An Integrated Approach for Video Captioning and Applications

by

Soheyla Amirian

(Under the Direction of Hamid R. Arabnia)

Abstract

Physical computing infrastructure, data gathering, and algorithms have recently had significant advances to extract information from images and videos. The growth has been especially outstanding in image captioning and video captioning. However, most of the advancements in video captioning still take place in short videos.

In this dissertation, we caption longer videos only by using the keyframes, which are a small subset of the total video frames. Instead of processing tens of thousands of frames, only a few frames are processed depending on the number of keyframes. There is a trade-off between the computation of many frames and the speed of the captioning process. The approach in this research is to allow the user to specify the trade-off between execution time and accuracy. We apply this system to make titles for videos. If we can generate reasonably meaningful titles for videos, the search engines could use them as metadata. For example, we search for videos that contain a woman with a red dress and sunglasses. Generating titles would significantly help search engines to search for videos. An additional novel application involves processing a video and directly generating captions describing a person's activities during a period without being constrained to time or location. The proposed model could be assistive technology to foster and facilitate physical activities. This framework could potentially help people manage their activities to reduce the health risks of an inactive lifestyle. Our work could be a healthcare application used by physicians or the public.

We demonstrate that these models and procedures and the interactions they enable are a path towards Artificial Intelligence. Our contribution lies in designing hybrid deep learning architectures to apply in long videos by captioning video keyframes. We consider the technology and the methodology that we have developed as steps toward the applications discussed in this dissertation. Consequently, the system we developed would open up doors for Generative Adversarial Networks to generate videos; such a thing does not presently exist. However, this is a step toward that goal.

Index words: Deep learning, Image captioning, Video captioning, Keyframe.

An Integrated Approach for Video Captioning and Applications

by

Soheyla Amirian

M.Sc. in Information Technology Engineering, Computer Networks,
Amirkabir University of Technology (Tehran Polytechnic), Iran, 2013

A Dissertation Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor of Philosophy

Athens, Georgia

2021

An Integrated Approach for Video Captioning and Applications

by

Soheyla Amirian

Major Professor: Hamid R. Arabnia

Committee: Khaled Rasheed
Thiab R. Taha

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
August 2021

Dedication

To My Dearest Professors, Family and Friends.

Acknowledgments

There are many angels that I must thank for contributing to these five glorious years of my journey as a Ph.D. student at the University of Georgia.

Foremost, I would like to express my sincere gratitude to my major professor, Dr. Hamid R. Arabnia, the actual owner of this dissertation. He shaped and formed me from an eager but mostly confused student to a fit researcher. Furthermore, I am very grateful to Prof. Arabnia for teaching me about some aspects of computer fields, the art of communicating ideas and teaching me how to think. I am thankful for his patience, motivation, passion, and immense knowledge. He always listened to me to understand what I had issues with and guided me to a direction in which I could comfortably excel. It was because he knew what my strengths and weaknesses are. Since the days I wanted to apply to this program, he has supported and helped me join the University of Georgia. I could not have imagined having a better advisor and mentor for my Ph.D. study. I am thankful for exploring my interests and enthusiasm, motivating me, guiding me, and helping me stay on track with his constant supervision, insightful comments, invaluable feedback, and inspiring me to do my best. Anytime I had a terrible paper draft, he patiently gave me advice many times until I got it. I appreciate his kindness in replying to my emails and phone calls instantly all the time. Even though he is one of the busiest men on earth, he was always available, always there, always listened, and always cared. When we were in conferences together, it allowed me to know him more. His sense of humor and efforts to make us happy will leave those trips as joyful times in my memory. My success is due to his support and mentorship. I appreciate him so much and value everything I have learned from him. His mentorship has been a precious gift over the past few years. I know I cannot thank him enough, but I can promise that I will be a similar mentor to my students and friends.

I have been very fortunate to learn from many remarkable inspiring people, for who I have developed a lot of respect and admiration. I want to extend my sincere thanks and appreciation to my doctoral advisory committee members - not only for their time and extreme patience but for their intellectual contributions to my development as a scientist. Prof. Khaled Rasheed helped to shed new light on many of my ideas. He was always eager to help. I am also very grateful to him for his scientific advice and knowledge, insightful discussions, and practical suggestions. He guided me through tough times and was with me to check on my research progress. He was always welcoming me when I was knocking on his office door without scheduling an appointment ahead. He provided me a machine to do my experiments and advised me to complete my dissertation. I am very thankful for his assistance and support at every stage of the research. I am greatful to Prof. Thiab R. Taha, who always had golden advice and for his prompt emails. He is one of the founders of the computer science department, and I appreciate all the hard work he has been doing

for all of us. I have been here for about five years, and every year I have seen a significant improvement in all aspects. And that is because of Prof. Taha's supervision. It is my honor to have Prof. Taha as advising and participating in my Ph.D. committee. I admire his encouragement, insightful comments, advice, suggestions, and support.

I have been very fortunate to learn from many other remarkably inspiring people that I have developed a lot of respect and admiration for and become my role models in many respects. My sincere thanks go to Dr. Keshtgari for enlightening me with her positive and encouraging energy. In addition, I sincerely appreciate the friends I have had the distinct pleasure of working with and learning from; Zengyan Wang, Mohammadhossein Toutiaee, Abolfazl Farahani, and my other friends. I am also very grateful to Dr. DiCosty for his fantastic sense of respect when helping me edit the thesis. A warm word for Evette Dunbar and Kimberly Buffington made me feel special and permitted me to record a video and use it for my research.

An exceptional word of thanks goes to my parents, who always believe in me ten times more than I do and give me strength and enthusiasm. My heartfelt gratitude goes to my siblings for their blessings, love, and inspiration. Special and profound thanks to my sisters Pooran and Sara and my brother Mansoor, who offered invaluable support and humor over these years and for always showing how proud they are of me. Last, I would like to thank God for his showers of blessings throughout my life, giving me the patience, strength, wisdom, support, and opportunity to accomplish my goal.

I am thankful for their unconditional support. I will remain forever grateful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Based on Wikipedia's definition, Artificial intelligence (AI) is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals, which involves consciousness and emotionality. One of the researcher's goals in Artificial Intelligence and Computer Vision is to enable computers to understand the visual world around us. In addition, scientists try to communicate with machines in natural human language so that machines can benefit people by doing various tedious tasks. Humans can accomplish different tasks that need visual understanding, including communication and interpretation in natural language by looking at a picture. A human can extract the information and describe an immense amount of details. For instance, by looking at Figure 1.1, we can immediately describe it as "It is a man standing on a rock next to a small waterfall. He has a blue-white outfit with a hat. He is holding something in his left hand.". In another example, Figure 1.2, we can watch a video and describe it as "It is a woman, wearing a yellow top, standing in front of a computer and talking on the phone.".

1.1.1 Challenges

It may seem effortless for humans to see scenes and describe them. However, this is a complicated task for computers. The process through which computers can gain high-level understanding from digital images or videos to understand and automate tasks that the human visual system can do is dense. In the digital world, each image or video frame represents an extensive array of numbers called pixels. For example, each pixel indicates the brightness at any position. A typical image includes a few million of these pixels, and a computer must transform these values into semantic concepts. These concepts are categorized into different classes —for instance, humans, animals, objects, actions, and many others. Moreover, a different type of object seen under different lighting conditions, with a different camera angle, or a different pose might depict another caption. Here, "a woman standing in front of a computer", however, the brightness values could change and result in a different caption. Furthermore, patterns with very similar low-level statistics (high-frequency patterns) might instead be part of many different objects (planes, cars, and many others)

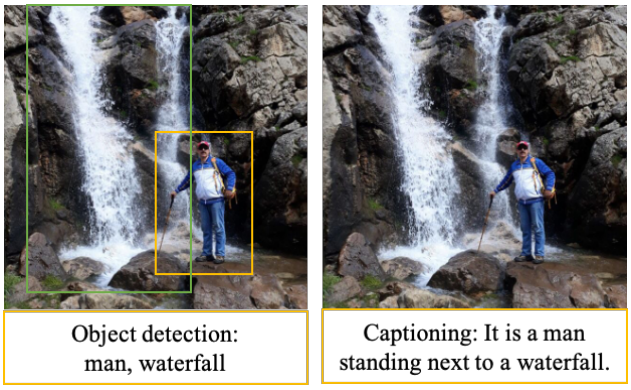


Figure 1.1: Left: detecting objects and information in an image. Right: detecting and describing the image

Figure 1.2: It describes the scenery and actions.

or animals (cats, dogs, bears, and many others). Thus, the challenges are no less severe on the language side. A natural language description such as "It's a woman standing in front of a computer and talking" will be represented in the computer as a sequence of integers indicating the index of each word in a vocabulary, for example [2, 1803, 409, and many others]. All told, the first step of accurately identifying and naming different parts of an image requires a complex pattern recognition process that needs much computation. Moreover, image captions often require detecting and describing complex high-level concepts that are not only visual but require problematic inferences such as actions. It requires detecting multiple objects and analyzing their poses, spatial arrangements, or even facial features for characters. For example, someone could be described as "delivering" something, "dancing," or "jumping". Moving on to videos, complexity increases as each video consists of many images (frames). In addition, time matters, so the sequence and order of frames would be important too. Consequently, there would be a massive computation process to extract all the information, and assign a meaningful caption to the video to describe it appropriately.

Recent Progress

There has been reasonably good progress in deep learning, visual recognition, image captioning, and video captioning systems. Notably, the state-of-the-art image recognition models based on deep convolutional neural networks [71] have become capable of distinguishing thousands of visual categories. Similarly, advances in related tasks such as segmentation and object detection have been dramatic [94]. Much impactful research has been done on image captioning [41], [65], [128], [134], and video captioning [69], [83]. In addition, many fundamental training datasets have become available. Together, these advances enabled many real-world applications. They include face detection, visual recognition, activity detection, photo search, automatic caption generation for images and videos for people who suffer from various degrees of visual impairment, general-purpose robot vision systems, and many others. Moreover, these advances can positively and significantly impact many other task-specific applications.

The overall approach in most of these applications is to model the visual recognition part. The first step is to extract the critical information to classify and detect objects, actions, and many others, from images. And then, the model attempts to caption images in real-time. For example, R-CNN significantly improves the quality of candidate bounding boxes and takes a deep architecture to extract high-level features [47]. Fast and Faster R-CNN focus on speeding up the R-CNN framework by sharing computation and using neural networks to propose regions. Nonetheless, while Fast and Faster R-CNN offer speed and accuracy improvements over R-CNN, both still fall short of real-time performance [92]. On the other hand, deep Residual Networks (ResNet) present a residual learning framework and gain accuracy from considerably increased depth [53]. Temporal Action Detection is a tool that focuses on localizing the temporal extent of each action for a detected object [20]. Object detection benchmarks use a different set of chosen categories (e.g., "car," "person," "plant," and many others). Different datasets for scene classification, action classification, or attribute classification have their own sets of categories. While visual classes constitute a convenient modeling assumption, this approach sometimes pales compared to the complexity of descriptions that humans can compose for images and videos (e.g., Figure 1.3).

Figure 1.3: Here are two examples of inaccurate captioning.

1.2 Related Work

Image processing has been and will continue to be playing an important role in science and industry. Applications include visual recognition [94], remote sensing [75] and scene understanding [27]. Before the advent of Deep Learning, most researchers used imaging methods that worked well on rigid objects in controlled environments with specialized hardware [13], [14], [31], [39], [59], [60], [81], [87], [115], [120].

Deep Learning

Deep Learning is a platform for solving impactful and challenging problems. Deep Learning allows computers to learn from experience and understand the world in terms of a hierarchy of concepts [48]. In recent years, deep learning-based convolutional neural networks have positively impacted image recognition and increased flexibility. Many different deep learning models are able to extract and process detailed information from images. In 2012, the deep convolutional neural network (DCNN) won the ImageNet classification competition. That model achieved the top-five test error rate of 15.3%, while the second-best model was 26.2% [71]. Since then, many researchers have advanced the deep learning model design, applications, and interpretation. The science behind deep learning goes back more than a half-century, but an increasing abundance of digital data and powerful GPUs have accelerated the development of deep learning research. Convenient development libraries such as TensorFlow and PyTorch, the open-source community, large labeled data sets (e.g., ImageNet, MNIST, PASCAL VOC, COCO, CIFAR, SVHN) [47], [61], [71], [92], and effective demonstrations have stimulated the growth of the deep learning field explosively.

Image and Video Captioning

Describing a short video in natural language is a trivial task for most people but challenging for machines. From the methodological perspective, integrating the models or algorithms is tricky because it is challeng-

ing to combine the contributions of the visual features and the adopted language model into the final description. Automatically generating natural language sentences describing a video clip generally has two components: extracting the visual information as Encoder and expressing it in a grammatically correct natural language sentence as Decoder. With a convolutional neural network, the objects and features are extracted from the video frames. Then, a neural network is used to generate a natural sentence based on the available information, using an image captioning method for captioning the frames [6].

In image captioning, Aneja et al. [12] developed a convolutional image captioning technique with existing Long Short Term Memory (LSTM) techniques and analyzed the differences between Recurrent Neural Networks (RNN) based learning and their method. Their techniques differ primarily in their intermediary components. Aneja et al. use masked convolution in a CNN-based approach, whereas RNN employs LSTM or GRU. The intermediary component of Aneja et al. is feed-forward without any recurrent function, and their CNN with attention (Attn) achieved comparable performance to the RNN-based approach. They also experimented with an attention mechanism and attention parameters using the conv-layer activations. The results of the CNN+Attn method were increased relative to the LSTM baseline. For better performance on the MS COCO, they used ResNet features, and the results show that ResNet boosts their performance on the MS COCO. The results on MS COCO with ResNet101 and ResNet152 were impressive.

In video captioning, Krishna et al. [69], presented Dense-captioning, which focuses on detecting multiple events that occur in a video by jointly localizing temporal proposals of interest and then describing each with natural language. This model introduced a new captioning module that uses contextual information from past and future events to describe all events jointly. They implemented the model on the ActivityNet Captions dataset. The captions that came out of ActivityNet shifted sentence descriptions from being object-centric in images to action-centric in videos. Ding et al. [34] proposed novel techniques for the application of long video segmentation, which can effectively shorten the retrieval time. Redundant video frame detection based on the Spatio-temporal interest points (STIPs) and a novel super-frame segmentation are combined to improve the effectiveness of video segmentation. Next, the super-frame segmentation of the filteblue long video is performed to find an action-containing clip. Then, keyframes from the most impactful segments are converted to video captioning using the saliency detection and LSTM variant network. Finally, the attention mechanism is used to select more crucial information from the traditional LSTM. Generative Adversarial Networks help to have more flexibility in these methods [5]. Therefore, Sung Park et al. [113] applied Adversarial Networks in their framework. They propose to use adversarial techniques during inference, designing a discriminator which encourages multi-sentence video description. They decouple a discriminator to evaluate visual relevance to the video, language diversity and fluency, and coherence across sentences on the ActivityNet Captions dataset.

Sequence models such as recurrent neural network (RNN) [26] have been widely utilized in speech recognition, natural language processing, and other areas. In addition, sequence models can address supervised learning problems such as machine translation [25], name entity recognition, DNA sequence analysis, video activity recognition, and sentiment classification. *LSTM*, as a particular RNN structure, has proven to be stable and robust for long-range modeling dependencies in various studies and can be

adopted as a building block for complex systems. The rugged unit in Long Short Term Memory is called a memory cell. Each memory cell is built around a central linear unit with a fixed self-connection [55]. LSTM is historically proven to be more powerful and more effective than a regular RNN since it has three gates (forget, update, and output). Therefore, Long Short Term Memory recurrent neural networks can be used to generate complex sequences with long-range structure [68], [129].

Text Summarization:

Automatic text summarization produces a concise and fluent summary while preserving key information content, and overall meaning [3]. Extractive and Abstractive are the two main categories of summarization algorithms. Extractive summarization systems form summaries by copying parts of the input. Extractive summarization is implemented by identifying the crucial sections of the text, processing them, and combining them to create a meaningful summary. Abstractive summarization systems generate new phrases, possibly rephrasing or using words that were not in the original text. Abstractive summaries are generated by interpreting the raw text and developing the same information in a different and concise form using complex neural network-based architectures such as RNNs and LSTMs. Paulus et al. [89] proposed a neural network model with a novel intra-attention that attended over the input while it continuously generating the output. Separately, a new training method combined standard supervised word prediction and reinforcement learning (RL). When standard word prediction was combined with RL's global sequence prediction training, the resulting summaries became more readable. Also, Roul et al. [91] combined the landscape of transfer learning techniques for Natural Language Processing (NLP) with a unified framework that converts every language problem into a text-to-text format. Text summarization can be further divided into two categories: single and multi-text summarization. In single text summarization [102], the text is summarized from one document. In contrast, Multi-document text summarization systems can generate reports that are rich in critical information and present varying views that span multiple documents.

Sources of Data:

The considerable majority of modeling approaches in this area fall under data-driven techniques, in which a model learns from human explanation data. It is therefore essential to highlight the available datasets. There are a few datasets that relate the visual domain with the field of natural language for image captioning models. This dissertation uses the Flickr8K [65], Flickr30K [65], [131], and Microsoft COCO [24], [131] datasets consisting of a set of images, each annotated with five descriptions written by humans on the Amazon Mechanical Turk. We use a few different datasets for video captioning and classify them into five domains based on the video contents: People, Open Subjects, Social Media, Cooking, and Movie (Automatic Image and Video Caption Generation with Deep Learning: Algorithmic Overlap chapter). We used videos from the Microsoft Video Description dataset (MSVD) and MSR Video to Text (MSR-VTT) Dataset for the experiments in the introduction and conclusion chapters. The Microsoft Video Description dataset (MSVD) [21] contains 1,970 YouTube clips with human-annotated sentences. The duration

of each video in the MSVD dataset is typically between 10 to 25 seconds. On average, 41 descriptions for each video are there. In total, this video dataset has approximately 80,000 description pairs and about 16,000 vocabulary words, which Microsoft Research provides. MSR Video to Text (MSR-VTT) [130] is created by collecting 257 popular queries from a commercial video search engine, with 118 videos for each query. MSR-VTT provides 41.2 hours of 10K web video clips with 200K clip-sentence pairs in total, covering a list of 20 categories. A caption could be a sentence that may explicitly mention non-obvious aspects of the scene (such as people, locations, or dates) that the model cannot derive from the image alone (i.e., it adds information). In the case of image-sentence datasets, the human annotators describe the image's content with a sentence. If the captions are inconsistently defined, it is impossible to achieve high accuracy. Therefore, these datasets contain numerous image descriptions since it is challenging for people unfamiliar with the specific context to caption the image appropriately.

1.3 Outline of Contributions

Each video consists of many frames. Some of these frames have the same contents, and some have essential information. Instead of giving all these frames to a video captioning model to assign a caption, we extract the video frames that contain necessary information called keyframes. Then we apply captioning techniques to those keyframes. Therefore, we caption videos only by using the keyframes, which are a small subset of the total video frames. Instead of processing tens of thousands of frames, a few frames are processed depending on the number of keyframes.

Figure 1.4 is an example from Wang et al. [123] for a video captioning model that generates a caption based on the video. Wang et al. proposed a reconstruction network (RecNet) with an encoder-decoder-reconstructor architecture, which leverages both the forward (video to sentence) and backward (sentence to video) flows for video captioning.

Video

Figure 1.4: SA-LSTM: a man is in the water; RecNet-local: a man is taking pictures on boat; RecNet-global: people are riding a boat [123]

Next, Figure 1.5 shows some selected keyframes from the sample video. Then, captions are generated by an image captioning model [80] model[1]. It concludes with exciting results as we can see more descriptions and more details about the video.

In an application, we apply this system to make titles for videos. If we can generate reasonably meaningful titles for videos, the search engines could increase efficiency by using the titles as metadata.

[1]https://github.com/ruotianluo/self-critical.pytorch

Selected Keyframes

Figure 1.5: A group of people standing next to each other. A man is holding his camera up to take a picture. A group of people riding a boat across a body of water. [80]
Ground Truth: bunch of people taking pictures from the boat and going towards ice.

An additional novel application involves taking in a video and directly generating captions describing a person's activities during a period without being constrained to time or location. The proposed model could be assistive technology to foster and facilitate physical activities. This framework could potentially help people trace their daily movements, and reduce the health risks of an inactive lifestyle by managing their activities. Our work could be a healthcare application used by physicians or the public. Figure 1.6 illustrates this model.

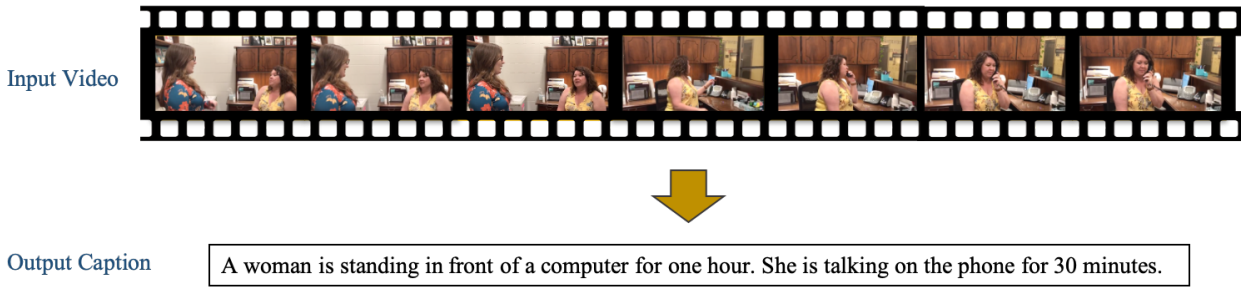


Figure 1.6: As an example of the proposed framework, the generated story describes the type of each physical activity and the corresponding duration of the video.

1.3.1 Motivations

The above contributions offer helpful and practical applications in the short term and will lead to enhanced machine understanding in the long term.

Long-term Motivations

The present research is a step towards a future where we can interact with computers, incredibly where these interactions can be appropriately grounded in physical environments. Therefore, working towards

artificially intelligent assistants will require us to make large amounts of information about how our world works available to computers. Concretely, there are extensive sources of knowledge for this motivation. The physical domain contains information about the world, scenes, objects, and interactions (vision, language, and visual sensors). The Internet's digital domain includes a vast amount of information not accessible from the physical domain (e.g., what happened in 1981). Because vision and language are the primary means to access the world's knowledge, we must develop techniques that can relate information across these two areas instead of processing each one independently. A solid short-term example might allow a computer to generate a caption based on the objects that it extracts from the images or videos. For instance, "it's a man that is standing in the snow","it's a wooden statue in a park".In the longer-term future, the computer could understand more by examining the procedure and events in the images or video and inform us about it if my kid "wants to touch a pan on a stove" or if they "want to jump over a fire". Furthermore, the system we developed would open up doors for Generative Adversarial Networks (GANs) to be used for generating videos; such a thing does not presently exist. However, this is a step toward that goal.

Short-term Motivations

The aspirations to select keyframes to generate captions and descriptions for a video can also be motivated by more concrete, short-term, and practical arguments. An Artificial Intelligence system reads a video; representative image frames are identified and selected. Indeed, selecting keyframes and information from all the video frames is essential and challenging. Then, it captions the image frames. We use image captioning models that accelerate the processing of a video. Natural Language Processing generates captions, and text summarization generates a video title. Figure 1.7 shows an example for selecting video keyframes and captioning them by image captioning model. Here, it captioned the video as "It's a car driving down a street, and a person riding a bike down a street." by selecting the keyframes. We demonstrate that our mod-

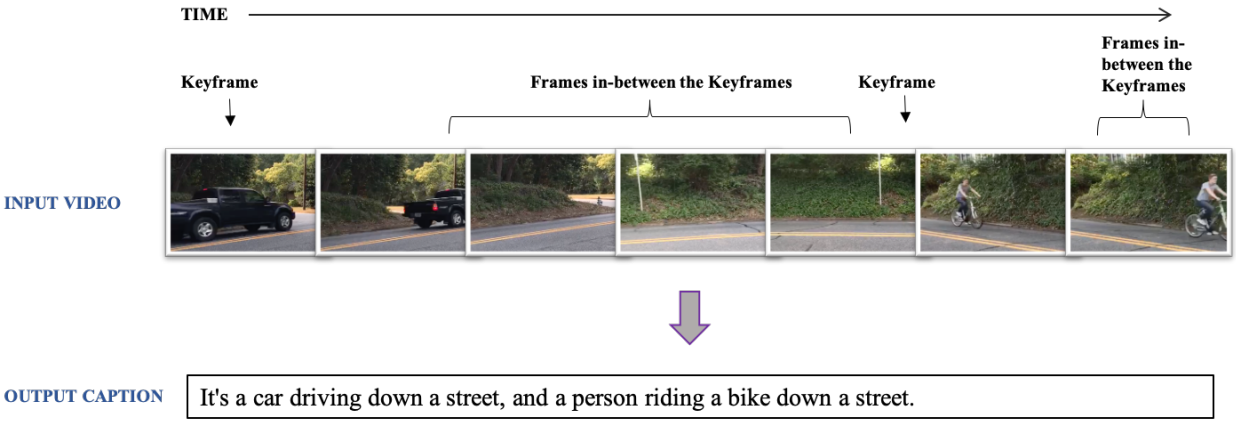


Figure 1.7: Notice the keyframes.

els, the procedures they take advantage of, and the interactions they enable, are a path towards Artificial Intelligence. In addition, we argue that linking images, videos, and natural language offers many practical

benefits and immediate valuable applications. From the modeling perspective, instead of designing and staging explicit algorithms to process videos and generate captions in complex processing pipelines, our contribution lies in designing hybrid deep learning architectures to apply in long videos by captioning video keyframes. In addition, we show how our model facilitates tedious human tasks by utilizing different deep learning models.

Challenges of this Approach

However, the Keyframes approach also poses some challenges. For example, which frames should we call keyframes? And how can we select the keyframes for captioning? One common criticism is that we may miss some vital information when we choose only some keyframes in a video. In addition, selecting the keyframes may generate descriptions that may not be accurate. Further, assessing the accuracy itself is complex. Presently, we have to evaluate accuracy to a human interpretation. In our work, we select each keyframe based on a transition that happens in the video. And we use the state-of-the-art image captioning model that offers the highest correlations with human judgments. We expect further improvement by evaluating the whole end caption with the integration model in the future.

Another challenge is that this approach integrates the keyframe selection (which images give us more information), visual recognition task (how to extract the information from keyframes), with the language modeling task (how to generate a meaningful caption based on the information we have). Therefore, it may feel safer to disintegrate these tasks, study them separately and then compose them to form the full model later. On the other hand, addressing these tasks combined will allow us to formulate a single model that automatically generates captions for a video. In Figure 1.8, we picked an example from Gao et al. [44]. First, we chose a keyframe from the video. Then we generated captions with an image captioning model [80]. The result is notably impressive.

Figure 1.8: Left: a-LSTMs: a woman is washing her hands [44]; Right: a woman standing in a kitchen next to a stove top oven [80].
GT: A woman is stirring some ingredients.

In another example, Figure 1.9, we see that one keyframe does not give us a clear caption. Therefore, we need to choose more keyframes from the video to have a better caption. Consequently, we added another keyframe. Figure 1.10 shows the generated caption.

As a result, common sense tells us that the more frames we process in a video, the more accurate a caption will be. However, more computations increase the expense. Therefore, there is a trade-off between

Figure 1.9: Left: a-LSTMs: a group of man are playing football [44]; Right: A blurry image of a baseball game in progress. A baseball game is being played on a green field. [80].
GT:People are playing football.

Figure 1.10: A blurry image of a baseball game in progress. A baseball game is being played on a green field. [80].

the computation of many frames and the speed of the captioning process. However, the approach in this research is to allow the user to specify the trade-off between execution time and accuracy.

1.4 Dissertation Outline

This dissertation proposes models by integrating deep learning, image and video captioning, and natural language processing models to accelerate the captioning process and offer new Artificial Intelligence applications. In these proposals, we use keyframes to caption long videos. These applications apply to videos for real-world purposes.

Chapter 2 provides relevant deep learning structures and models for Classification, Object Detection, and Segmentation [9]. Then we have a case study; Integrated Plant Growth and Disease Monitoring with IoT and Deep Learning Technology [42]. Hereabouts we discussed what the proposed system hopes to achieve and implications for future research, possibly setting the stage for a follow-up study in collaboration with an agricultural subject matter expert. Following another case study, Stereotype-Free Classification of Fictitious Faces. In this case study, we investigate different Generative Adversarial Networks. We present a novel approach through penalized regression to label stereotype-free GAN-generated synthetic unlabeled images [117].

Chapter 3 is a concise review of image captioning methodologies based on deep learning, strengths and limitations, the datasets, and the evaluation metrics used in automatic image captioning [6]. Then we explore Autoencoders as neural networks that learn data codings in an unsupervised manner [5].

Chapter 4 presents Automatic Image and Video Caption Generation with Deep Learning. This chapter does not mean only a comprehensive review of image captioning; instead, it is a concise review of image captioning and video captioning methodologies based on deep learning. This study treats both image and video captioning by emphasizing the algorithmic overlap between the two [8].

In **Chapter 5**, we specifically address the problem of generating novel descriptions by keyframes in long videos. This chapter proposes an architecture that utilizes image/video captioning methods and Natural Language Processing systems to create a title and a concise abstract for a video. Many application domains, including the cinema industry, video search engines, security surveillance, video databases/warehouses, data centers, and many others, can potentially utilize such a system [7].

Finally, in **Chapter 6**, we propose another application for video captioning. This study focuses on one application: video captioning for fostering and facilitating physical activities. In broad terms, we consider assistive technology [4].

Chapter 2

Dissection of deep learning with applications in image recognition

Deep Learning has dramatically improved the accuracy of image recognition. Image recognition is considered to be one of the most challenging problems in imaging science. This chapter briefly surveys the application of Deep Learning in classification, object detection, semantic segmentation, and instance segmentation of objects in digitized images. The chapter reviews widely used Deep Learning models in classifying, detecting, and segmenting objects based on convolutional neural networks (CNN). It provides a summary and presents some discussion about the performance of each model, including training time and accuracy[1]. For a more thorough and slower-paced introduction, we recommend the Deep Learning book from Goodfellow et al. [48].

Image processing has been and will continue to be playing an important role in science and industry. The applications spread in many areas, including, visual recognition [94], remote sensing [75] and scene understanding [27] to name a few. Before the advent of Deep Learning, most researchers used imaging methods that worked well on rigid objects in controlled environments with specialized hardware [13], [14], [31], [39], [59], [60], [81], [87], [115], [120]. In more recent years, deep learning-based convolutional neural networks have positively and significantly impacted image recognition, allowing a lot more flexibility. In this chapter, we attempt to highlight recent advances in image recognition in the context of deep learning.

In 2012, the deep convolutional neural network (DCNN) won the ImageNet classification competition. That model achieved the top-5 test error rate of 15.3%, while the second-best model was at 26.2% [71]. Since then, many researchers have advanced the deep learning model design, applications, and interpretation. The science behind deep learning goes back to more than a half-century, but an increasing abundance of digital data and powerful GPUs accelerate the development of deep learning research. Convenient development libraries such as TensorFlow and PyTorch, the open-source community, large labeled data sets

[1] Amirian, Soheyla, Zengyan Wang, Thiab R. Taha, and Hamid R. Arabnia, "Dissection of deep learning with applications in image recognition," in Computational Science and Computational Intelligence; "Artificial Intelligence" (CSCI-ISAI); 2018 International Conference on IEEE CPS (IEEE XPLORE, Scopus), 2018, ISBN-13: 978-1-7281-1360-9, pp. 1132–1138.

(e.g., ImageNet, MNIST, PASCAL VOC, COCO, CIFAR, SVHN) [47], [61], [71], [92], and splendid demonstrations simulate the growth of the deep learning field explosively.

In this chapter, we discuss the overall process of a deep learning model in Section 2.1. We present a brief introduction of the tasks in deep learning in Section 2.2. Then the components from the recent deep learning architectures are surveyed in Section 2.3. Next, Section 2.4 introduces some widely used models for each application category in image classification, object detection, semantic segmentation, and instance segmentation. Then, in Section 2.5, we discuss the Internet of Things (IoT) technology and image classification. Moreover, in Section 2.6, we focus on generative adversarial network (GAN) models by proposing an approach that helps to label data by minimizing a penalized version of the least-squares cost function between realistic pictures and target pictures.

2.1 Deep Learning

We need a sophisticated model with a significant learning capacity to learn about thousands of objects from a large number of images [71]. Deep learning presents a developing tool dealing with this data in the field of image recognition. It is a set of approaches that allows a machine to be fed with data and discover the descriptions needed for classification, detection, and segmentation. Deep learning methods are learning methods with multiple levels of representation. The levels of representation are obtained by composing simple parts that each transform the representation at one level into a representation at a higher (slightly more abstract) level with the composition of enough such conversions. Then, very complicated functions can be learned. For classification tasks, higher layers of representation, develop aspects of the input that are notable for inequality and suppress unrelated changes. An image, for example, comes in the form of a matrix of pixel values, and the learned features in the first layer of representation typically represent the presence or absence of edges at particular orientations and locations in the image. The second layer usually detects motifs by detecting specific arrangements of edges, regardless of small changes in the edge positions. The third layer may assemble motifs into larger combinations that correspond to parts of familiar objects, and subsequent layers would detect objects as combinations of these parts. The crucial point of deep learning is that these layers of features are learned from data using a general-purpose learning technique [73].

2.2 Image Based Applications

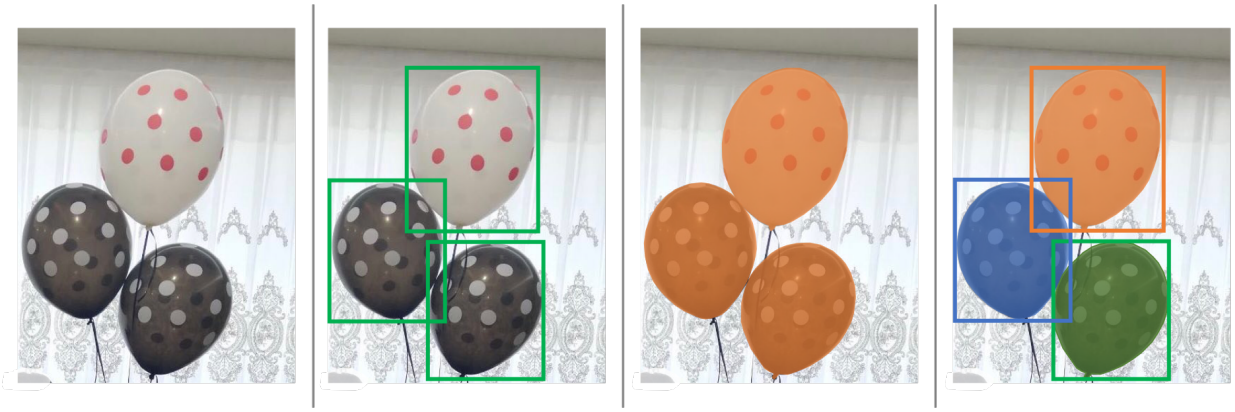
Deep learning based convolutional neural network significantly plays a crucial role in many applications such as image recognition. For example, image recognition is used to perform many visual tasks, such as understanding the content of images, self-driving cars, traffic sign detection, image caption, automatic colorization, and more. Here, we categorized deep learning tasks for image recognition into classification, object detection, semantic segmentation, and instance segmentation. Fig. 2.1 illustrates an example of the tasks introduced in this chapter.

1. **Classification**: A classification dilemma can be formally defined as the task of computing the label y of a K-dimensional input vector x, where $x \in X \subseteq R^K$ and $y \in Y = \{C_1, C_2, ..., C_Q\}$ [90] (Fig. 2.1. a).

2. **Object Detection**: This is a further step of classification, not only the label but also produces a bounding box around the objects within an image. Therefore, detection requires localizing objects within an image [47]. Deep learning upon convolutional networks is doing an impressive job in detecting objects in computer vision by localizing objects in images or video frames [46], [94] (Fig. 2.1. b).

3. **Semantic Segmentation** Each pixel is assigned a label to indicate its semantic class in an image, such as flower, person, road, car (Fig. 2.1. c).

4. **Instance Segmentation** Instance segmentation is the task of identifying individual instances of one semantic class in an image [22], [75], [100] or across video frames. Instance segmentation is the most challenging vision task than the last three tasks (Fig. 2.1. d).



a. Classification b. Object Detection c. Semantic Segmentation d. Instance Segmentation

Figure 2.1: a: There is a balloon in this image; b: There are 3 balloons in this image at these locations; c: Prediction of all pixels belong to balloons; d: Each balloon has a unique identification.

2.3 Deep Learning Structure

Here, we dissect some parts of the deep learning model based on a convolutional neural network. And, we introduce some operations and functions that are applied to optimize the task of the layers. Convolutional neural networks (CNNs) with several layers of convolution, pooling and non-linear units have shown considerable success in image recognition tasks [67] (Fig. 2.2 offers an overview of those tasks). The architecture of a typical convolutional neural network is structured as a series of blocks. Each block is composed of two types of layers: convolutional layers and pooling layers. CNNs are designed to process data that

come in the form of multiple matrices, for example, a color image composed of a three-dimensional matrix containing the pixels (units) of the three color channels [73] Units in a convolutional layer are organized in feature maps, within which each unit is connected to local patches in the feature maps of the previous layer through a set of weights called a filter bank. The result of this local weighted sum is then passed through a non-linearity such as a ReLU (Rectified Linear Unit) function. All units in a feature map share the same filter bank. Different feature maps in a layer use different filter banks. Several convolutions, non-linearity, and pooling blocks are stacked, followed by more convolutional and fully connected layers. By adding convolutional blocks, the system can extract high-level information from images. Finally, the backpropagation algorithm indicates how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer [73].

2.3.1 Convolutional Layer

The architecture of a typical convolutional neural network is typically structured as a series of convolutional blocks.

- **Convolution:** The convolution operation calculates sum of the element-wise multiplication between the input matrix and kernel matrix.
- **Transpose Convolution:** The network of the transposed convolution operation or Fractionally-Strided Convolution and Deconvolution learns how to upsample optimally. The transposed convolution operation backwardly performs the same connectivity as the general convolution.
- **Atrous Convolution:** There are two challenges in a deep convolutional neural network: 1) reducing feature resolution causes loss of information and 2) the objects exist in multi-scale images [22], [23]. Atrous convolution offers a possible solution without additional parameters and computational time. The kernel is inflated by inserting spaces between the kernel elements.
- **Encoder-Decoder:** It is also called a recognition or inference network. This architecture has a symmetric down-sampling and up-sampling structure. The down-sampling system (encoder) extracts high-level features through convolution, and the up-sampling structure (decoder) recovers the spatial resolution. In addition, there are skip connections between the encoder and decoder of the same size to help reconstruct information with features extracted previously [101], [106].
- **Multi-task:** Multi-task learning uses one deep model to achieve multiple related tasks [105].

2.3.2 Classification Layer

- **Fully Connected layer:** The most common CNNs have a few Conv-ReLU layers, following with Pool layers and repeats this pattern until the image is merged spatially to a small size. The Fully-Connected (FC) layers mostly follow a stack of convolutional layers (which has a different depth in different architectures) [53], [67], [92], [111]. And, using the dropout technique in FC layers might

significantly reduce overfitting [61]. However, some models would not have FC layer(s) in their architecture [101].

- **Softmax:** The last fully-connected layer or normalization layer, called Softmax, holds the output, such as the class scores [17], [111]. Softmax function squashes a K-dimensional vector of arbitrary real values to a K-dimensional vector of real values. Each value is in the range of 0 and 1, and all the values add up to 1. The softmax function is widely adopted by many CNNs [53], [71] due to its simplicity and probabilistic interpretation. Together with the cross-entropy loss, they form arguably one of the most commonly used components in CNN architectures [56], [78], [101].

2.3.3 Optimization Operation

- **Loss Function:** Choosing a loss function [56], [100], that accurately reflects the objective we want to achieve is a fundamental key for any model to be able to learn a given task, which is used to measure the inconsistency between predicted values and actual label.

    1. **Hinge Loss:** Current widely used data loss functions in CNNs include Euclidean loss [111], (square) hinge loss, information gain loss, contrastive loss, triplet loss, softmax loss, and many others [78].
    2. **Cross Entropy Loss:** Cross entropy loss and softmax are arguably one of the most commonly used supervision components in CNNs [78], [101].

- **Gradient Descent:** For optimization [56], normally the Stochastic Gradient Descent (SGD) will work well [17], [78]. Stochastic gradient-based optimization is of core practical importance in many fields of science and engineering. Adam, a method for efficient stochastic optimization that only requires first-order gradients with less memory requirement, computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients [67]. For problems with very noisy and sparse gradients, efficient stochastic optimization techniques are required [67].

2.3.4 Some Gears

- **Attention Model:** An attention-based model is for recognizing multiple objects in images. The proposed model by Lei Ba et al. is a deep recurrent neural network trained with reinforcement learning to attend to the most relevant regions of the input image [16].

- **Salient:** Salient Object Detection (SOD) has several applications such as segmentation, object proposal generation, and image resizing. SOD aims at highlighting salient object regions in images [126].

- **GRU:** Gated Recurrent Units (GRUs) is a gating mechanism in recurrent neural networks (RNNs), introduced in 2014 by Kyunghyun Cho et al. [25].

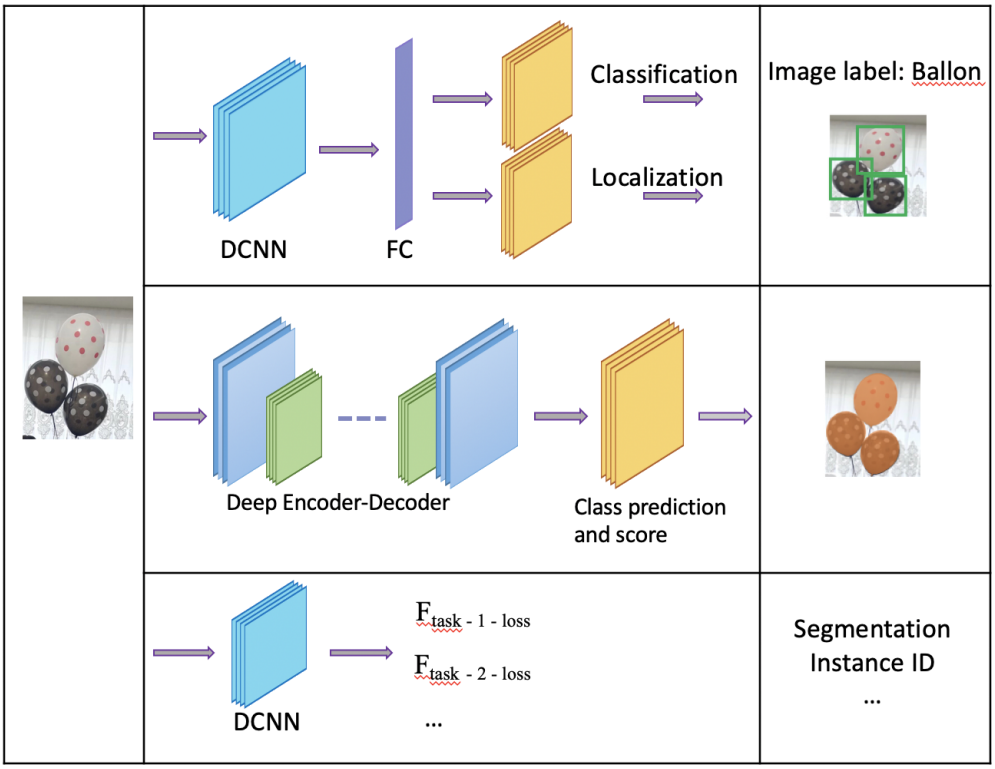


Figure 2.2: Overview of deep learning structure

2.4 Deep Learning Models

Deep learning presents computational models that are composed of multiple processing layers to learn representations of data in images [73]. Several well-known models in CNNs based on image classification, object detection, semantic segmentation, and instance segmentation are listed here. Some are summarized according to the architecture, task, and the number of parameters in Table 2.1.

2.4.1 Classification

There are several commonly used architectures in the field of Convolutional Networks by now that some are listed:

- **AlexNet:** The first effort that popularized Convolutional Networks in Computer Vision was the AlexNet, developed by Alex Krizhevsky et al. The neural network, using thoroughly supervised learning, contains eight learned layers (five convolutional layers, some of which are followed by max-pooling layers, and three FC layers followed by a softmax layer). To reduce overfitting in the FC layers, it also uses the dropout technique [71].

Table 2.1: Summary of models

Model	Architecture	Tasks	Parameters
AlexNet	CNN+FC	Classification	$62M$
VGG	CNN+FC	Classification	VGG16: $138M$ VGG19: $143M$
GoogleNet	CNN+FC	Classification	$4M$
ResNet	CNN+FC	Object detection	$1.7M$
DenseNet	CNN+FC	Object detection	$7M$

*M for million.

- **VGG:** The runner-up in ILSVRC 2014 was the network from Karen Simonyan and Andrew Zisserman, introduced as the VGGNet. Their work evaluated profound convolutional networks for large-scale image classification. The input to the ConvNets is a fixed-size 224×224 RGB image. The image passes through a stack of convolutional layers. The network contains 16 CONV/FC layers and features a highly homogeneous architecture that only performs 3×3 convolutions and 2×2 pooling from the beginning to the end. It also uses 1×1 convolution filters (a linear transformation of the input channels). The convolution stride is fixed to 1 pixel. Spatial pooling is carried out by five max-pooling layers, which follow some of the convolutional layers. Three FC layers follow a stack of convolutional layers (a different depth in different architectures). The final layer is the softmax layer [111].

- **GoogleNet:** The ILSVRC 2014 winner was a Convolutional Network from Szegedy et al. from Google, a 22 layer deep network. Its main contribution was the development of Inception architecture that spectacularly reduced the number of parameters in the network so that inferences can be run on individual devices, including even those with limited computational resources. Also, it uses average pooling instead of fully connected layers at the top of the ConvNet. A 1×1 convolution with 128 filters for dimension reduction and rectified linear activation is used. An FC layer with 1024 units and rectified linear activation, a dropout layer with a 70% ratio of dropped outputs, and a linear layer with softmax loss as the classifier has been used [114].

2.4.2 Object Detection

- **RCNN Series:** R-CNN significantly improves the quality of candidate bounding boxes and takes a deep architecture to extract high-level features. R-CNN obtained a mean average precision (mAP) of 53.3% with more than a 30% improvement over the previous works on PASCAL VOC 2012 [47]. Fast R-CNN employs several innovations to improve speed while also increasing accuracy. Fast R-CNN trains the very deep VGG16 network $9\times$ faster than the R-CNN, is $213\times$ faster at test-time, and achieves a higher mAP on PASCAL VOC 2012. A Fast R-CNN network takes as input an

entire image and a set of object proposals. The network first processes the whole image with several conv. and max-pooling layers to produce a conv. feature map. Then, for each object proposal, a region of interest (RoI) pooling layer extracts a fixed-length feature vector from the feature map. Each feature vector is fed into a sequence of FC layers that finally branch into two sibling output layers: softmax and another layer that outputs real-valued numbers for each of the object classes [46]. Fast and Faster R-CNN focus on speeding up the R-CNN framework by sharing computation and using neural networks to propose regions. While they offer speed and accuracy improvements over R-CNN, both still fall short of real-time performance [92].

- **YOLO:** You Only Look Once, or YOLO detection network, has 24 convolutional layers followed by two fully connected layers. Alternating 1×1 convolutional layer reduces the features space from preceding layers. It pretrains the convolutional layers on the ImageNet classification task at half the resolution (224×224 input image) and then doubles the resolution for detection. YOLO is a general-purpose detector that learns to detect various objects simultaneously; but struggles with small objects that appear in groups, such as flocks of birds [92]. YOLOv3 has significant benefits over other detection systems. Namely, it is faster and better. It processes images in real-time but still struggles with accuracy [94].

- **ResNet:** The runner-up in ILSVRC 2015 was a network from Kaiming He et al. that became known as the ResNet won first place on ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation. Deep Residual networks present a residual learning framework and can gain accuracy from considerably increased depth. On the ImageNet dataset, it evaluates residual nets with a depth of up to 152 layers, $8\times$ deeper than VGG nets (the more, the merrier!) [53].

- **DenseNet:** Gao Huang et al. introduced the Dense Convolutional Network (DenseNet), which connects each layer to every other layer in a feed-forward fashion. DenseNet has several compelling advantages: It alleviates the vanishing-gradient problem, confirms feature propagation, provokes feature reuse, and considerably reduces the number of parameters. For CIFAR-10, CIFAR-100, SVHN datasets, the DenseNet has three dense blocks, each with an equal number of layers. A convolution with 16 output channels is performed on the input images. For convolutional layers with kernel size 3×3, and zero-padded by one pixel. 1×1 convolution followed by 2×2 average pooling. At the end of the last dense block, an average pooling is performed, and then a softmax is attached. The feature-map sizes in the three dense blocks are 32×32, 16×16, and 8×8, respectively. In the experiments on ImageNet, a DenseNet-BC structure with four dense blocks on 224×224 input images is used. The initial convolution layer comprises 2k convolutions of size 7×7 with stride 2. DenseNet tends to yield consistent improvement in accuracy with the growing number of parameters, without any signs of performance degradation or overfitting. Under multiple settings, it achieved state-of-the-art results across several highly competitive datasets. Moreover, DenseNets require substantially fewer parameters and less computation to achieve state-of-the-art performances [56].

2.4.3 Semantic Segmentation

- **FCN:** Fully convolutional networks that address many pixelwise tasks are a rich class of models. FCNs by transferring pre-trained classifier weights, fusing different layer representations, and learning end-to-end on whole images dramatically improve accuracy for semantic segmentation. End-to-end, pixel-to-pixel operation simultaneously simplifies and speeds up learning and inference [79], [106].

- **U-Net:** Ronneberger et al., winner of the ISBI cell tracking challenge 2015, present a network and training strategy that relies on the strong use of data augmentation to use the available annotated samples more efficiently. The network architecture consists of a contracting path and an expansive path. The contracting path follows the typical architecture of a convolutional network. It consists of the repeated application of two 3×3 convolutions, each followed by a rectified linear unit and a 2×2 max pooling operation with stride 2 for downsampling. Every step in the expansion path consists of an upsampling of the feature map followed by a 2×2 convolution that halves the number of feature channels and two 3×3 convolutions followed by a ReLU. At the final layer, a 1×1 convolution is used to map each 64-component feature vector to the classes. In total, the network has 23 conv. layers. The architecture consists of a contracting path to capture context and a symmetric expanding approach that enables precise localization. Furthermore, the network is fast. The U-net architecture achieves excellent performance on very different biomedical segmentation applications [101].

- **SegNet:** Badrinarayanan et al. present a novel and practical deep fully CNN for semantic pixel-wise segmentation termed SegNet. This core trainable segmentation engine consists of an encoder network, a corresponding decoder network, followed by a pixel-wise classification layer. SegNet architecture has no fully connected layer, and hence, it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs a convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a softmax classifier for pixel-wise classification. This chapter reveals the practical tradeoffs involved in designing architectures for segmentation, particularly training time, memory versus accuracy. SegNet performs competitively, achieving high scores for road scene understanding [17].

2.4.4 Instance Segmentation

- **Mask R-CNN:** Kaiming He et al. presented Mask R-CNN, which extends Faster R-CNN. It detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. It also allows us to estimate human poses [52]. Instance segmentation, the automatic depiction of different objects appearing in an image, is a problem within computer vision that has attracted a fair amount of attention. Such interest is motivated by its potential applicability to various scenarios and the stimulating technical challenges it poses. Instance segmentation is more

challenging than other pixel-level learning problems, such as semantic segmentation, which deals with classifying each pixel of an image, given a set of classes. Each pixel can belong to a set of predefined groups (or categories), whereas in instance segmentation, the number of groups (instances) is unknown a priori. This difference exacerbates the problem: wherein semantic segmentation one can evaluate the prediction pixel-wise, instance segmentation requires the clustering of pixels to be evaluated with a loss function invariant to the permutation. That leads to further complexities in the learning of these models. Hence, instance segmentation has remained a more complex problem to solve [22], [100].

2.5 Case Study 1: Integrated Plant Growth and Disease Monitoring with IoT and Deep Learning Technology

This study proposes an Integrated Plant Growth and Disease Monitoring solution that combines Internet of Things (IoT) sensor data, high-resolution imagery, and manual intervention data in a synchronized time-series database environment. The overall system architecture integrates individual components, including sensors, drone imagery, image processing, database framework, and alerting mechanism. These components are brought together and synchronized in a time-series database. By synchronizing all the variables, this solution presents a comprehensive view, and better means for intervention [42].

Artificial Intelligence (AI) has permeated human life in all aspects, and employing it is beneficial in health, science, agriculture, economics, and finance. In agriculture, there are many applications for AI. For example, monitoring health and disease on plants plays a vital role in farm profitability, as AI can reduce costs and time-to-market. AI alone does not achieve this goal; instead, an integrated solution utilizing various technologies would best serve the market. In this research, *sensors* detect environmental variables such as air temperature, light level, soil moisture, and CO_2 levels. *Time-series databases* play a central role as we measure the condition of plants over time. Intervention data, sensor data, and image processing data all come together in this database and are synchronized by timestamp to show the effects of input and output across all variables. *Image classification* as a topic of image recognition, classifies the contextual information in images. In our proposed system, the classification task is formally defined as the process of labeling plant leaf images as Healthy or Sick, in order to detect diseases. The classification is an important component of our solution, one part of a whole system working in unison.

Care of plants in any large-scale agricultural setting is a heavily manual process that is time-consuming and dependent upon both the human eye and inaccurate estimates of soil conditions. The effect of water and nutrients is not known in real-time, and any adverse effects show up as latent indicators. At present, the time, labor, and inaccuracies in plant care make scalability a significant concern in large-scale agricultural operations.

2.5.1 Architecture and Methodology

The system includes sensors, drones, and a log of manual activities such as watering or pruning. The data are compiled in a cloud virtual machine (VM) and synchronized in a time-series database. A deep learning model processes the images to detect signs of disease, and an alerting protocol notifies the user via email or SMS text if the detection threshold is met.

Analytics Server. The analytics server is an Ubuntu virtual machine (VM) running on the Linode cloud service. Key components of the VM include the time-series database (InfluxDB), data collector (Telegraf), Python scripts, dashboard service (Grafana), and Apache webserver.

Activity Log. The activity log keeps manual interventions such as watering, feeding, pruning, relocating, and many others. We maintain such a log of manual interventions in the time-series database to complete a picture of cause and effect on the plants. For example, if soil moisture spiked at a particular time, we may expect to see a corresponding entry in the event log that indicates the plant was watered.

Sensors. The data collected from the individual plant is handled by a modified Particle Photon sensor [88]. The sensor collects ambient temperature, humidity, light level, soil moisture level, CO_2 level, and TVOC level. These readings are made and sent to the Particle Cloud and then delivered to the analytics server, powered by InfluxDB and its related Telegraf data collector plugin [57] every ten (10) seconds.

Image Intake and Analysis. Drones flying over the specified area take high-resolution images of individual plants at regular intervals. These images are inputs to the analytics server, where a Python script stores the file in block storage and pushes the file location string to the time-series database. A second Python script, known as the classifier, analyzes the image utilizing a deep learning model trained on photos of healthy and diseased plants. The classifier script triggers an alert via email or SMS message. The Keras framework powers the classifier script itself. The images taken by drones feed into the classifier. The Convolutional Neural Network system extracts the information from the image, and based on the trained process, the system classifies it as Healthy or Sick. Figure 2.3 shows the classifier process.

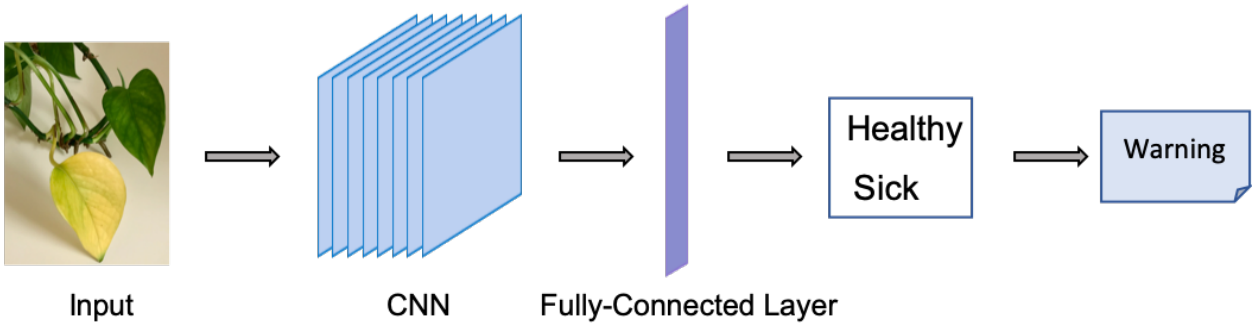


Figure 2.3: Overview of the deep learning methodology for classifying images as Healthy or Sick.

Alerting System. When an image is identified as a sick plant, designated recipients receive an alert. This alert includes the image, the likelihood of disease score, and the time stamp. The intention is for the recipient(s) to query the database for activities and sensor readings from a specified time window, searching for variations in the sensor readings and manual activity logs to understand what conditions

led to the sick plant. Next, the recipients decide whether or not to manually intervene to identify and change the conditions that led to the sick plant. We intend to catalog those manual interventions to build a model for various responses to the conditions that led to the sick plant. The ultimate goal is to have an automated response with minimal manual intervention.

2.5.2 Discussion

The concept presented in this research is an ongoing investigative research project. Here we discussed what we hope to with the proposed system and implications for future research, possibly setting the stage for a follow-up study in collaboration with an agricultural subject matter expert. An essential next step is developing the alerting system to transition to a semi-automated or fully automated response system based on the grower's preferences. A fully automated response system would utilize the machine learning models produced to apply the correct interventions. A semi-automated response system would leverage the same process but allow the grower to choose the intervention from a list recommended by the model.

2.6 Case Study 2: Stereotype-Free Classification of Fictitious Faces

Studies are developing Equal Opportunity and Fairness in artificial intelligence. However, Stereotyping as another source of discrimination has not been unstudied in literature. If human perception classifies GAN-made faces, it would be exposed to such discrimination. It is possible to eliminate the human impact on fictitious faces classification tasks using statistical approaches. This study proposes a novel approach through penalized regression to label stereotype-free GAN-generated synthetic unlabeled images to help to label new data (fictitious output images) by minimizing a penalized version of the least-squares cost function between realistic pictures and target pictures [117].

Despite the appealing application and success in Machine Learning tasks, a significant field within Artificial Intelligence that began slower but has expanded enormously in recent years is fairness. Discrimination refers to the effect of bias against people's lives due to their membership in different population subgroups. These subgroups are differentiated by the sensitive (protected) attributes recognized by national and international legislation. Many applications of machine learning, including decision-making, can, perhaps unintentionally, result in an unfortunate lack of fairness. As an example, their outcomes can asymmetrically deprive (or enrich) specific subgroups of people with one or more common protected attributes such as race, gender, caste, and religion. [62] enumerated a few examples of applications in policing, hiring, and lending where the systematic decision process discrimination might be inevitable. Thus, this realization motivates a new era of research in machine learning in the knowledge of fairness actions.

Scientists have extensively practiced around structured data. In Fairness Through Unawareness work (FTU) [51] proposed a definition of a fair algorithm provided that sensitive attributes A are not explicitly trained in the model. Experts in individual fairness study (IF) [38] presented that individuals i and j are similar under a pre-defined metric function if their predictions are similar.

Many researchers have targeted discrimination-aware decision-making approaches, each of which proposes a new quantification of discrimination. Thus, fairness definition in predictive models is still controversial with the absence of consensus among researchers. A new school of thought in fairness function is often published via research papers to dampen the discrimination effect. The variety of approaches leads to difficulty evaluating the progress in the field and can assess no strengths and weaknesses for further recommendations accordingly.

This research focuses on the Generative Adversarial Network (GAN) in this study for two reasons; first, it can produce fictitious outputs (imaginary images). Second, it has inspired a legion of scientists to evolve GAN under the impression of producing more realistic-looking data [66]. Many domains widely applied GANs due to their impressive performance, especially on the image generation paradigm. In literature, GANs have only been used to help mitigate bias in data.

We observe that the super-high-quality fictitious images of humans generated by a state-of-the-art GAN, such as Style-GAN [66], might be prejudiced by gender or race to stereotype. In that case, a GAN-made picture of a person with long hair should not be realized necessarily as a woman, and a dark-complexion person would not be an African-American individual. According to the literature, the protected attributes of a person induce more sensitivity around the subject, and Stereotyping is another source of discrimination. The evolution of GANs output images (or videos) and Stereotyping issue is a motivation of this study. This work proposes an interpretable and practical approach to classify synthetic faces of Style-GAN [66] without symptoms of discrimination. The proposed method is applied after an image is generated and considers two binary sensitive attributes (race and gender).

A property of a GAN architecture is the lack of ground truth in the outcomes. In other words, faces that appear in outputs should not be labeled based on stereotypes such as White women or Asian men since these attributes are self-reported. In this case, the fictitious images cannot be self-reported by an individual who does not even exist. In this research, the novel approach alleviates the prejudiced view of sensitive attributes (such as gender or race) by minimizing the distance of a given output image from all other realistic input images that produce that target output.

Generative Adversarial Networks [49] are machine learning models that can imagine new samples. Many variants on GAN are extended by machine learning enthusiasts, while some are very prominent. We chose Style-GAN as the main architecture to bear imaginary faces, then we classify faces (bring them to life). **Style-GAN** [66]. We utilize Style-GAN in this study, since it produces high quality of images.

2.6.1 Proposed Approach

Linear regression and unbiasedness in our work are very similar. Both try to minimize the average distances of data points to a particular line, the regression line. The position of this line classifies unreal images fairly. In this study, Ridge regression would help locate this fair line. The Ridge regression model (Figure 2.5) would determine the attributes of GAN-made faces by finding the relation between real and imaginary faces. Coefficients estimation, a process of minimizing the cost function, would ultimately aid in performing the task.

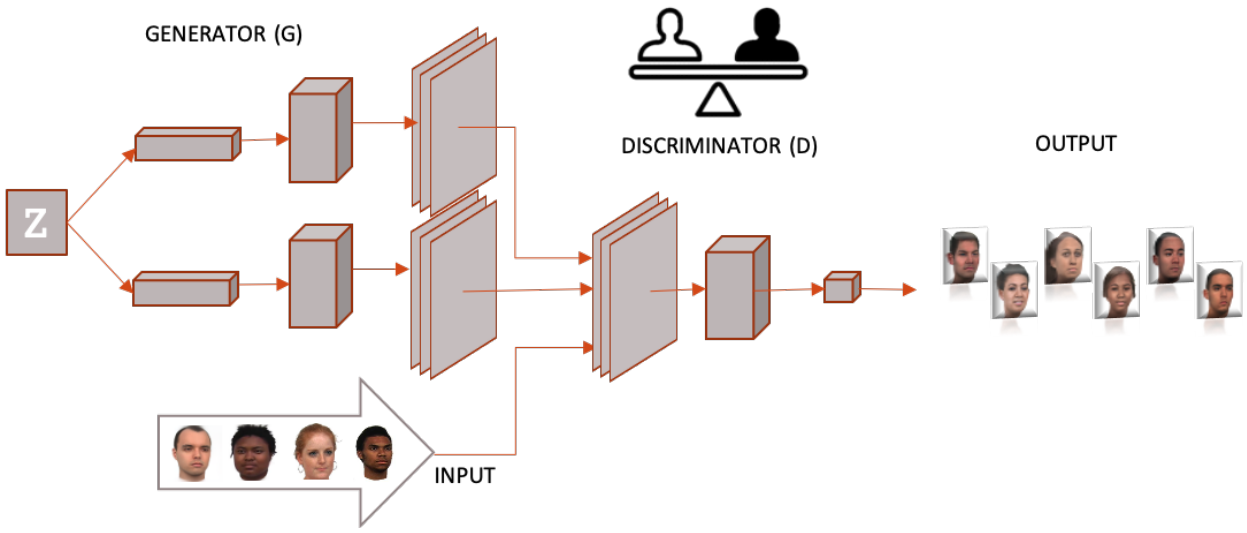


Figure 2.4: An adversarial training architecture for generating imaginary images.

Stereotype-Free Classification Method: Each coefficient X_j represents j-th GAN-made image that is supposed to be labeled. Thus, a dataset containing several observations n_i with several features X_j (imaginary images) and a response variable y_i (a protective attribute), which is given as the ground truth per observation form a classic dataset which can apply to all the statistical assessments and principles. The Earth Mover Distance (EMD), which reflects the similarity between content-based images, can be computed using various practical algorithms in the image retrieval domain.

Similarity Measure: We evaluate image similarity between X_j and n_i by *Earth Mover's Distance* (EMD), which has been studied in computer vision and image retrieval for a long time [104]. Discrete Kantorovich formulation (i.e., EMD), which arises from the idea of optimal transport, provides a better distinction between the images approximated by the histograms instead of other conventional measures such as Euclidean distance.

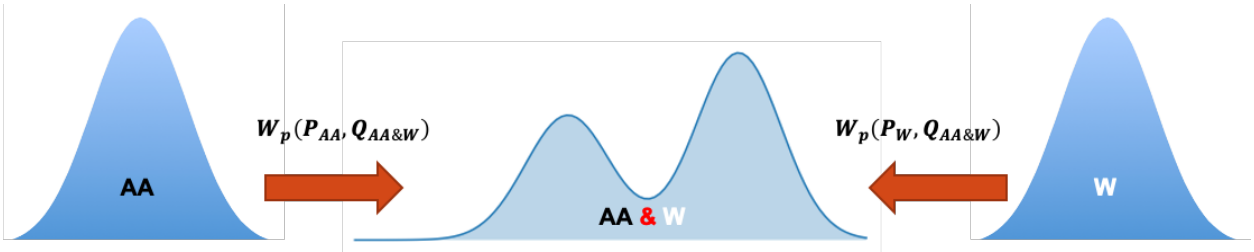


Figure 2.5: Earth Mover Distance (or 1D-Wasserstein) as similarity metric between images. "W" and "AA" are representing White and African-American, respectively.

Table 2.2: [Top] Gender: Male (M) Vs. Female (F). [Bottom] Race: African-American (AA) Vs. White (W) Results obtained from several evaluated methods for Gender (left) and Race (right). The first and second number per method indicate the number of instances that method has detected. The third number (P-value) shows the probability of the null hypothesis being true under the statistical threshold (0.05) to test whether the corresponding method prefers one particular attribute over another. If the related P-value exceeds the threshold (0.05), then one concludes that the impartial preference is rejected and the corresponding method is biased against the minority. As both tables confirm, Ridge method tends to propagate labels to all images (64 out of 64) without any sign of discrimination (p-value $>$ 0.05).

Method(M)(F)(P-value)	**#detection**	**Biased?**
Human-Perception(31)(33)(0.9)	64	No
Face Classification(31)(33)(0.9)	64	No
Ridge(34)(30)(0.7)	64	No

Method(AA)(W)(P-value)	**#detection**	**Biased?**
Human-Perception(18)(46)(0.0006)	64	Yes
Face Classification(22)(36)(0.043)	58	Yes
Ridge(36)(28)(0.38)	64	No

Implementation Process

One positive side-effect of the similarity distance between inputs and outputs is a tendency of our approach to be less prone to classify unfairly. The earth mover distance (EMD) is chosen as a similarity metric since it matches better with the human perception of differences compared to other distances such as Euclidean distance or χ^2 divergence [103]. The EMD level obtained per output image encodes the relation between the unlabeled imaginary image and previously labeled images fed into the Adversarial training network (Figure 2.4).

As the number of unlabeled images increases, the dataset may suffer from the curse of dimensionality, and all the principles affected in the high-dimensional dataset are enforced. We primarily adopt the Ridge Logistic Regression model with an extra regularization term (penalty) to handle such high-dimensional data. One positive property of Ridge Logistic Regression is that it preserves the features in the model, which ensures the true attributes tag all the imaginary faces.

Fairness Evaluation: Race preference study of GANs is another contribution of this work as it measures which race is preferred over another. In comparison with the baseline and face classification methods, we construct the null hypothesis H_0 to test statistically whether the protected attributes are equally preferred vs. the alternative (H_a) that they are not equally preferred across all the methods (Table 2.2). Then a sign test, with an α-level of 0.05, of the null hypothesis (H_0) is evaluated statistically by the two-tailed P-values in each approach, calculated by Binomial distribution.

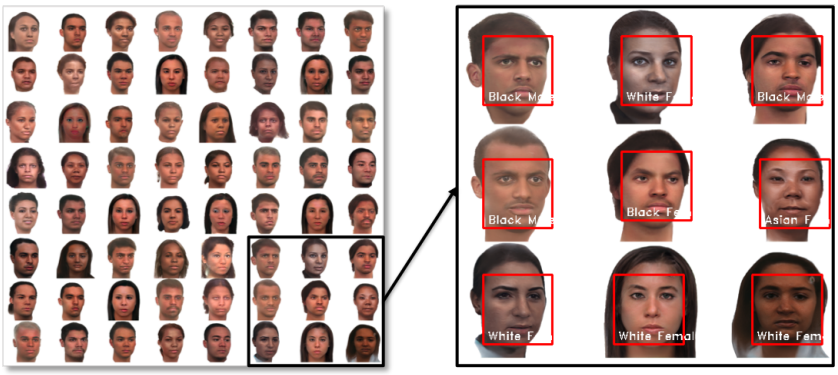


Figure 2.6: Left: Imaginary faces generated from GAN. Right: Attribute tagging by face classification.

2.6.2 Experimental Results and Conclusion

In our experiments, Ridge Logistic Regression tends to yield stereotyping-free labeling in generated unlabeled images in both tasks, without any signs of performance degradation. These evidences are provided in Table 2.2.

There is no ground truth for outputs of GAN-made images of fictitious people since GANs generate imaginary data. Images generated from GANs are mostly classified by human perception, and human perception suffers from stereotyping paradigm. We discussed that stereotype is another source of discrimination, which leads to behavioral bias against sub-groups of people. We presented a stereotype-free labeling approach to eliminate such discrimination as a result of human perception. This new angel of view stimulates a viewer's attention by looking at the supernatural imaginary faces. The proposed method is proper when there is no ground truth for faces, and it does not apply to images with true labels. The results revealed that Ridge Logistic Regression labeled all the imaginary images fairly due to its shrinkage property. At the same time, human perception and the trained face classification are biased in favor of one sub-group. Based on our experiment, it is becoming clear that human perception is not a reliable source for judging synthetic faces, and stereotypes govern it.

2.7 Summary

Deep learning is a pervasive technique in image recognition. Image recognition presents a challenging problem for scientists over recent decades and has received a significant deal of attention because of its many applications in science and industry. This chapter briefly surveyed the deep learning tasks in classification, object detection, semantic, and instance segmentation of objects in images. Then, some of the well-known deep learning models in classifying, detecting, and segmenting objects in the images based on convolutional neural networks are summarized. In the end, each model's performance parameters, such as training time and accuracy, are reviewed.

This chapter attempted to give an idea of the state of the art of image recognition technology. For example, YOLOv3 improved the real-time object detection, U-Net architecture achieved excellent performance on speed for semantic segmentation applications, but scientists still have challenges with instance segmentation.

Finally, we described two real examples of deep learning. First, we discussed the Internet of Things (IoT) technology and image classification in the subsequent case study as an example of combining technology and deep learning models to apply in AI. Moreover, we focused on generative adversarial network (GAN) models by proposing an approach that aids labeling data in the second case study.

Chapter 3

A Short Review on Image Caption Generation with Deep Learning

Methodologies that utilize Deep Learning offer sophisticated potential for applications that automatically attempt to generate captions or descriptions about images. Image captioning is considered to be one of the intellectually challenging problems in imaging science. The application domains include automatic caption (or description) generation for images for people who suffer from various degrees of visual impairment; the automatic creation of metadata for images (indexing) for use by search engines; general-purpose robot vision systems; and many others. Each of these application domains can positively and significantly impact many other task-specific applications. This chapter is a concise review of image captioning methodologies based on deep learning, strengths and limitations, the datasets, and the evaluation metrics used in automatic image captioning. Then, a quick discussion about the software and hardware requirements for implementing an image captioning method is presented. Finally, we discuss a study about Image Captioning with Generative Adversarial Network[1].

In recent years, deep learning based convolutional neural networks has positively and significantly impacted image captioning, allowing a lot more flexibility. In this chapter, we attempt to highlight recent advances in image captioning in the context of deep learning. Since 2012, many researchers have advanced the deep learning model design [9], applications and interpretation [71]. The science and methodology behind deep learning have been in existence for decades, but an increasing abundance of digital data and the involvement of powerful GPUs has accelerated the development of deep learning research in recent years. Convenient development libraries such as TensorFlow and PyTorch, the open-source community, large labeled datasets (e.g., MSCOCO, Flicker, and more) [95], [98], and splendid demonstrations simulate the explosive growth of the deep learning field.

Describing a scene in an image is a highly demanding task for humans. To create machines with this capability, computer scientists have explored methods to connect the science of understanding human lan-

[1]Amirian, Soheyla, Khaled Rasheed, Thiab R. Taha, and Hamid R. Arabnia, "A Short Review on Image Caption Generation with Deep Learning", The 23rd International Conference on Image Processing, Computer Vision and Pattern Recognition (IPCV'19), World Congress in Computer Science, Computer Engineering and Applied Computing (CSCE'19), IEEE, 2019, ISBN: 1-60132-506-1, pp. 10-18.

guage with the science of automatic extraction and analysis of visual information. Image captioning needs more effort than image recognition because of the additional challenge of recognizing the objects and actions in the image and creating a concise, meaningful sentence based on the contents found. Nevertheless, the advancement of this process opens up enormous opportunities in many application domains in real life, such as aid to people who suffer from various degrees of visual impairment, self-driving vehicles, sign language translation, human-robot interaction, and more. This chapter surveys the state-of-the-art approaches with a focus on deep learning models for image captioning. The models and generated captions are evaluated using BLEU, METEOR, CIDEr [24], [121], [131], and other metrics.

This chapter is a concise review of image captioning methodologies based on deep learning. This review begins by introducing the Image Captioning in Section 3.1. Then, a few recent methods of Image Captioning, the Datasets and Metrics are discussed in Section 3.2. Afterward, Required Software and Hardware Platforms for implementing a model are mentioned in Section 3.3. Finally, we discuss about Image Captioning with Generative Adversarial Network in Section 3.4.

3.1 Image Captioning

Image captioning is the process of generating a concise description of an input picture/ image (See Figure 3.1). Typically, such functions are done manually. Automating this process would be a significant contribution. Much research has been done on image captioning [41], [65], [128], [134] that are pretty impactful. A system that automatically generates image captions can be utilized in many applications. Examples include: enhancing the accuracy of search engines, recognition, and vision applications; enriching and creating new image datasets; enhancing the functionality of systems similar to Google Photos; and enhancing the optical system analysis of self-driving vehicles.

There are some challenges in extracting visual information from the input and transforming the visual information into a proper and meaningful language for image captioning. Captioning research started with the classical retrieval, and template-based [41], [74] approaches in which Subject, Verb, and Object are detected separately and then joined using a sentence template. However, the advent of Deep Learning and the tremendous advancements in Natural Language Processing has equally affected the area of captioning. Hence, the latest approaches follow deep learning based architectures that encode the visual features with Convolutional Neural Networks and decode with a language-based model, which translates the features and objects given by an image-based model to a meaningful sentence.

3.2 Image Captioning Methodologies

Automatically generating natural language sentences describing an image generally has two components: extracting the visual information and expressing it in a grammatically correct natural language sentence. Figure 3.2 depicts a simple Encoder-Decoder deep learning based captioning framework for image captioning. By convolutional Neural Network the objects and features are extracted from the image, then we need a network to generate a natural sentence based on the information we have.

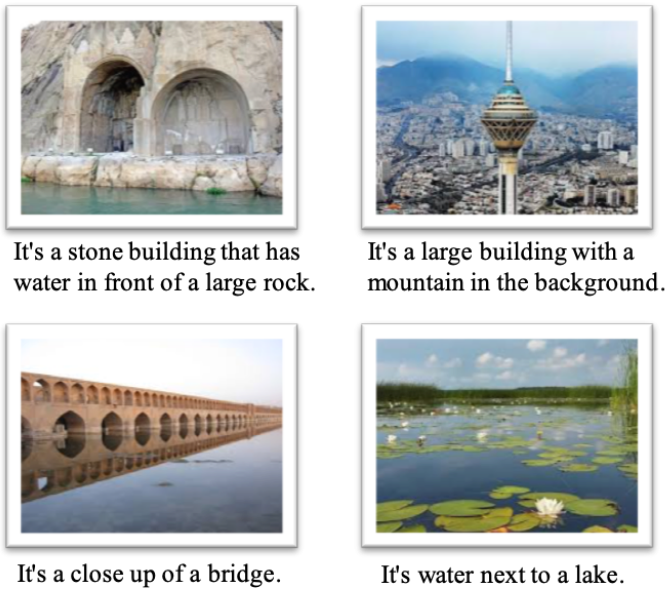


Figure 3.1: These are a few examples of captions that has been generated for images.

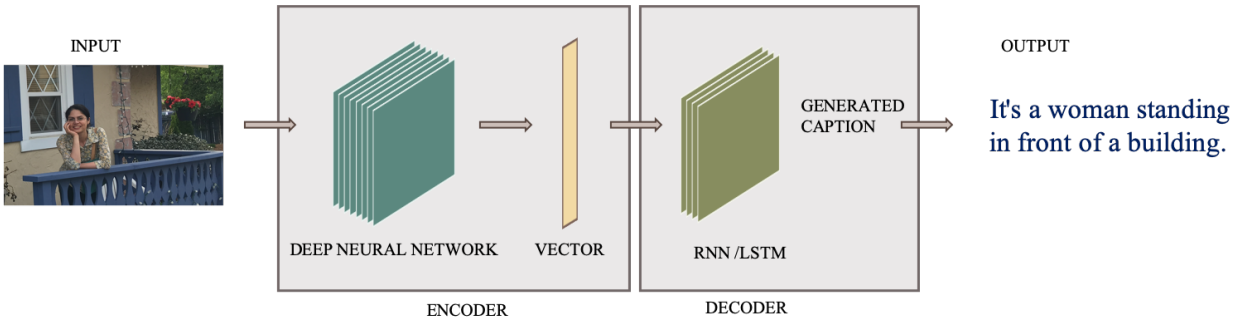


Figure 3.2: This is an overall encoder-decoder structure for image captioning models. A deep learning model encodes the image into a feature vector. The language model takes the input vector to generate a sentence that describes the image.

- Convolutional Neural Network: There is a need for a model with a large learning capacity to learn about thousands of objects from a large number of images [71]. Deep learning presents computational models that are composed of multiple processing layers to learn representations of data in images [9], [73]. Deep learning based Convolutional Neural Networks play a key role in many applications, one of which is image recognition. Image recognition is used to perform many visual tasks, such as understanding the content of images. There are several well-known models [9] in the field of CNNs based on object detection [46], [47], [94] and segmentation [100].

- Recurrent Neural Networks: Sequence models like recurrent neural network (RNN) [26] have widely been utilized in speech recognition, natural language processing, and other areas. Sequence models can address supervised learning problems like machine translation [25], name entity recognition, DNA sequence analysis, and sentiment classification.

- Gated Recurrent Unit: Gated recurrent unit (GRU) is a gating mechanism in RNN, introduced in 2014 by Cho et al. [25]. The basic RNN algorithm runs into a vanishing gradient problem (a difficulty in training artificial neural networks). The gated recurrent units are an effective solution for addressing the vanishing gradient problem. They allow neural networks to capture a much longer range of dependencies [26]. The advantage of the GRU is that it is a simple model, and so it is easy to build an extensive network. Also, it only has two gates, so it computes quickly.

- Long Short Term Memory: LSTM, as a particular RNN structure, has proven to be stable and robust for long-range modeling dependencies in various studies. LSTM can be adopted as a building block for complex structures. The rugged unit in Long Short Term Memory is called a memory cell. Each memory cell is built around a central linear unit with a fixed self-connection [55]. LSTM is historically proven more powerful and more effective than a regular RNN since it has three gates (forget, update, and output). Long Short Term Memory recurrent neural networks can be used to generate complex sequences with long-range structure [68], [129].

3.2.1 Recent Deep Learning based Models:

There are many methods for image captioning. Earlier methods, prior to deep neural networks (DNNs), were retrieved-based [41] or template-based [74] models. Current methods are based on deep neural networks. Generating an automatic caption for describing an image has two stages. First, the information needs to be extracted from the image and put it in a feature vector. This stage focuses on visual recognition by deep learning models. Then the feature vector is fed into the second stage. The second stage is caption generation, which describes what is extracted in a grammatically correct natural language sentence. Figure 3.2 depicts an overall encoder-decoder structure for image captioning methods. So, we classified DNN-based methods based on the main framework into subcategories that they respectively use. Here, a review of recent deep learning based methods for automatic image captioning is discussed. All are summarized in Table 3.1.

A breakthrough in image captioning occurred in 2014 through the application of encoder-decoder models. Kiros et al. introduced an encoder-decoder pipeline model in which an encoder network takes the image as an input and extracts a fixed-size feature vector that a decoder network maps to a sequence of words. They set new best results when using the 19-layer Oxford convolutional network. They were also developing an attention-based model that jointly learns to align parts of captions to images. The generated descriptions are arguably the nicest ones to date [68]. The attention model is one of the models used in deep learning from one of the most curious facets of the human visual system. The attention-based model learns to focus on different parts of the image. This is important when there is a lot of clutter in an image. However, this may cause losing information which could be helpful for richer and more descriptive captions [131].

Xu et al. proposed the attention-based approach that gives the state of the art performance on three benchmark datasets using the BLEU and METEOR metric (See section 3.2.3). They also showed how the learned attention could be exploited to give more interpretability to the model generation process and demonstrate that the learned alignments correspond very well to human intuition. Finally, their model encourages future work in using visual attention [131].

You et al. also proposed a model of semantic attention. The algorithm combines top-down and bottom-up strategies to extract richer information from the image. It fuses them with an RNN that can selectively attend to rich semantic attributes detected from the image. They performed their method on the Microsoft COCO and the Flickr30K, and the captioning system was implemented based on the LSTM network. The image feature vector is extracted from the last 1024 dimensional convolutional layer of the GoogleNet [9] CNN model. Furthermore, their framework employs attention at both input and output layers to the RNN module. Their effort exploited abundant fine-grain visual semantic aspects and fused global and local information for generating a better caption. The experimental results show that the algorithm significantly outperforms the state-of-the-art approaches consistently across different evaluation metrics [134].

Fu et al. proposed the image caption system that exploits the parallel structures between images and sentences. One contribution of this system is that it aligns generating captions and the attention shifting among the visual regions. Another is that it introduces the scene-specific contexts to LSTM that adapt language models for word generation to specific scene types. The architecture is that an image is first analyzed and represented with multiple visual regions from which visual features are extracted. The visual feature vectors are then fed into an LSTM network, which predicts both the sequence of focusing on different areas and generating words based on the transition of visual attention. A scene vector also governs the neural network model, a global visual context extracted from the whole image. Finally, intuitively, it selects a scene-specific language model for generating text. This has been tested on several popular datasets, including MSCOCO, Flickr8K, and Flickr30K. They evaluated captions in BLEU-n, METEOR, ROUGE-L and CIDEr-D metrics (See Table 3.1). Either region-based attention or scene-specific contexts alone improve performance, but combining the two provides a further improvement [43].

In 2018, Aneja et al. developed a convolutional image captioning technique with existing LSTM techniques and analyzed the differences between RNN based learning and their method. This technique

Table 3.1: The summary of a few recent works for Image Caption (All the results have been converted to percentages).

Model	Architecture	Evaluation
(2014, Kiros et al.)	CNN+LSTM encoder–decoder Attention-based	Image Annotation result R@1 R@5 R@10 Med r 23.0 50.7 62.9 5 (OxfordNet) on Flickr30K. R@K is Recall@K (high is good). Med r is the median rank (low is good).
(2015, Xu et al.)	CNN+RNN Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR 71.8 50.4 35.7 25.0 23.04 Hard attention on MSCOCO
(2016, You et al.)	CNN+RNN Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR 70.9 53.7 40.2 30.4 24.3 Using the ground-truth visual attributes on MSCOCO
(2017, Fu et al.)	Region-based attention and scene-specific contexts VGG/Alex/ResNet + LSTM Attention-based	ENSEMBLE result BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE-L CIDEr-D 72.4 55.5 41.8 31.3 24.8 53.2 95.5 on MSCOCO
(2018, Aneja et al.)	CNN+LSTM ResNet152	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE-L CIDEr 72.5 55.5 41 29.9 25.1 53.2 97.2 on MSCOCO
(2018, Anderson et al.)	CNN+LSTM Faster R-CNN;ResNet101 Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE-L CIDEr SPICE 80.2 64.1 49.1 36.9 27.6 57.1 117.9 21.5 on MSCOCO

contains three main components. The first and the last components are input and output word embedding, respectively. However, while the middle component contains LSTM or GRU units in the RNN case, masked convolutions are employed in their CNN-based approach. This component is feed-forward without any recurrent function. Their CNN with attention (Attn) achieved comparable performance. They also experimented with an attention mechanism by attention parameters using the conv-layer activations. The results of the CNN+Attn method were increased rather than the LSTM baseline. For better performance on the MSCOCO, they used ResNet features, and the results show ResNet boosts their performance on the MS COCO. The results on the MS COCO with Resnet101 and Resnet152 were comparable to previous works. Table 3.1 shows that the METEOR and CIDEr results are outstanding [12].

Anderson et al. proposed a combined bottom-up and top-down attention mechanism that enables attention to be calculated at the level of objects and other salient image regions. The bottom-up attention uses Faster R-CNN with ResNet-101 [9], which represents a natural expression of a bottom-up attention mechanism. The top-down mechanism uses task-specific context to predict an attention distribution over the image regions. The attended feature vector is then computed as a weighted average of image features over all regions. Their results on the MS COCO dataset present a new state-of-the-art for the task, achieving CIDEr, BLEU-4 scores of 117.9 and 36.9, respectively. The broad applicability of the method,

applying the same approach to Visual Question Answering, was obtained first in the 2017 VQA Challenge [11].

3.2.2 Image Captioning Datasets:

There are a few datasets that are widely used to evaluate and compare image captioning methods: Flickr8K [65], Flickr9K [131], Flickr30k [65], [131] and Microsoft COCO [24], [131].

- Flickr: The Flickr8K, 9k, and 30k datasets contain more than 8000, 9000, and 30000 images, respectively. Each image is annotated using Amazon Mechanical Turk with five independent sentences. The Flickr8K dataset mainly contains human and animal images, while the Flickr30k dataset contains humans involved in everyday activities and events. For each image, five sentences are provided [65], [131].
- COCO: Microsoft Common Objects in COntext (MS-COCO) is large-scale object detection, segmentation, and captioning dataset that contains 91 object categories, 328K images, and five assigned captions to each image [24], [77], [131].

3.2.3 Image Captioning Evaluation Metrics:

Captions are evaluated using the BLEU, METEOR, CIDEr, and other metrics [24], [121], [131]. These metrics are common for comparing the different image captioning models and have varying degrees of similarity with human judgment [110].

- **BLEU:** BiLingual Evaluation Understudy is a method of automatic machine translation evaluation that is a precision-based metric, correlates highly with human evaluation, and has a little marginal cost per run [86], [131]. BLEU has different n-grams based versions for candidate sentences concerning the reference sentences.
- METEOR: Metric for Evaluation of Translation with Explicit ORdering is an automatic metric that evaluates translation hypotheses. It is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations [18], [24], [32], [131].
- CIDEr: Consensus-based Image Description Evaluation enables objective comparison of machine generation approaches based on their human-likeness, without having to make arbitrary calls on weighing content, grammar, saliency, and many others with respect to each other [121]. CIDEr was first developed specifically for evaluating image captioning tasks, but it is also used in video captioning methods.
- ROUGE: Recall-Oriented Understudy for Gisting Evaluation determines the quality of a summary by comparing it to other summaries created by humans. ROUGE, similar to BLEU, has different n-grams based versions [76].

- SPICE: Anderson et al. introduced Semantic Propositional Image Captioning Evaluation, a novel semantic evaluation metric that measures how effectively image captions recover objects, attributes, and the relations between them. It correlates more with the human judgment of semantic quality than previously reported metrics [10].
- WMD: Word Mover's Distance measures the dissimilarity between two text documents. Therefore, the sensitivity of this metric when compared to BLUE, ROUGE, and CIDEr is low about word order or synonym swapping, but, like CIDEr and METEOR, provides high correlation with human judgments [72].

3.2.4 Discussion:

In this section, we briefly reviewed a few methods according to the standard approaches that they have used. For a fair comparison of the models, Table 3.1 shows the results of attention-based methods on the MS COCO dataset, the common dataset that they have utilized. We could state that Anderson et al. performed better on the MS COCO dataset. This method outperformed previous works. It uses the attention mechanism, which focuses only on relevant objects of the image. Also, We found that the performance of a technique can vary across different metrics, parameters, and datasets. Here, we tried to compare them based on the common performance criteria. However, image captioning still has a long way to improve the accuracy of captioning the events in images (See Figure 3.3).

3.3 The Required Platform for Implementation:

Deep Learning has dramatically improved the accuracy of image recognition. However, image recognition is considered to be one of the most challenging problems in image science. In recent years, deep learning based convolutional neural networks have positively and significantly impacted image recognition allowing a lot of flexibility. Deep Learning is responsible for many of the recent breakthroughs in image science, such as image captioning. Despite Deep Learning's popularity, it is difficult to accurately predict the time it takes to train a deep learning network to solve a given problem. The training time can be seen as the product of the training time per epoch and the number of periods that need to be performed to reach the desired level of accuracy. We define the features which could influence the prediction of execution time while performing the training. We categorize these features into layer, implementation, and hardware features. Each of these categories can contain almost an endless list of features. Layer (Algorithm or model) Features include Activation Function (ReLU, Softmax, Tanh, and more), Optimizer (Gradient Descent, Momentum, Adam, and more), Batch Size (the number of training samples which are processed together as part of the same batch), number of inputs to the layer, the neurons within the layer, Matrix, Kernel, Stride, and Padding size. Hardware Features include CPU, GPU, or TPU technology (memory, clock, speed, bandwidth,...) [64].

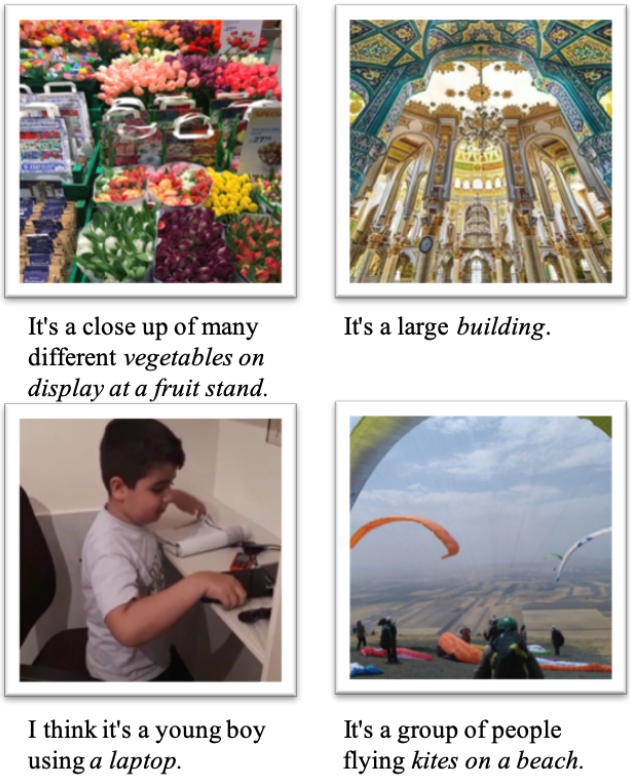


Figure 3.3: Image captioning still have a big room in improving the accuracy of describing the events and objects in images.

3.3.1 Software Requirement:

- Tensorflow: TensorFlow is an end-to-end open-source platform for machine learning. TensorFlow is developed by Google and has integrated the most common units in deep learning frameworks. It supports many up-to-date networks such as CNN and RNN with different settings. TensorFlow is designed for remarkable flexibility, portability, and high efficiency of equipped hardware [109].

- PyTorch: PyTorch is a Python-based scientific computing package that serves two purposes: as a replacement for NumPy to use the power of GPUs, and as a deep learning research platform that provides maximum flexibility and speed[2] [12].

- Keras: Keras is a high-level neural network API, written in Python and capable of running on TensorFlow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation. Going from idea to result with the least possible delay is the key to doing good research. Keras allows for easy and fast prototyping (through user-friendliness, modularity, and extensibility). Keras supports both convolutional networks and recurrent networks, as well as a combination of both. Keras runs seamlessly on CPU and GPU[3].

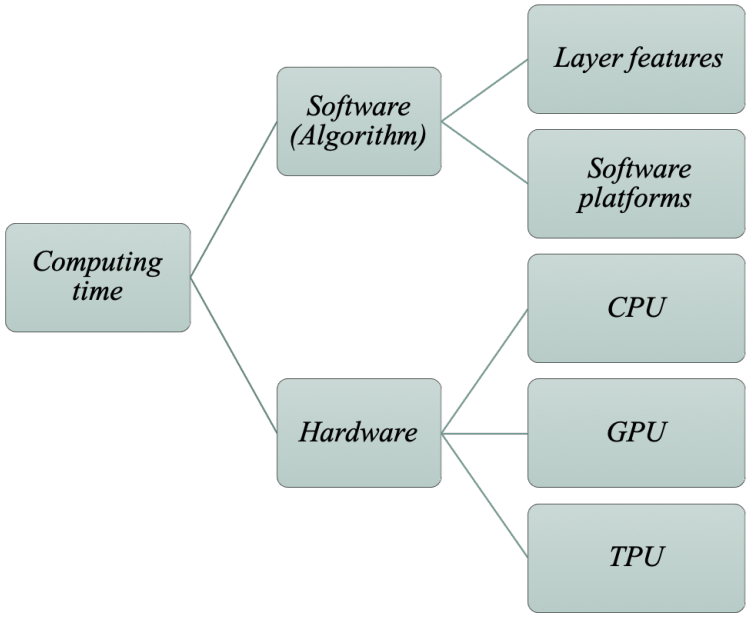


Figure 3.4: Computing time dependencies

[2]https://pytorch.org/tutorials/beginner/blitz/tensor-tutorial.html
[3]https://keras.io/

3.3.2 Hardware Requirement:

The science and methodology behind deep learning have been in existence for decades. In recent years, however, there has been a significant acceleration in the utilization of deep learning due to an increasing abundance of digital data and the involvement of powerful hardware.

- GPU: Compared to CPU, the performance of matrix multiplication on Graphics Processing Unit is significantly better. With GPU computing resources, all the deep learning tools mentioned achieve much higher speedup when compared to their CPU-only versions [109]. GPUs have become the platform of choice for training large, complex Neural Network-based systems because of their ability to accelerate the systems. For example, it used to take a few days to train AlexNet (the work of Krizhevsky et al. [71] which outperformed all other image recognition approaches at the time [9]) on the ImagetNet dataset with an NVIDIA K40 machine. Now with DGX-2, the NVIDIA group can train AlexNet in a few minutes[4]. Shi et al. worked to evaluate the running time performance of a set of modern deep learning software tools and see how they perform on different types of neural networks and different hardware platforms. They showed that all tested tools could use GPUs to achieve significant speedup over their CPU counterparts. However, there is no single software tool that can consistently outperform others, which implies that there are some opportunities to optimize further the performance [109].

- TPU: Tensor Processing Unit (Domain-Specific Architecture) is a custom chip that has been deployed in Google data centers since 2015. Tensors dominate dNNs, so the architects created instructions that operate on tensors of data rather than one data element per instruction [63]. To reduce deployment time, TPU was designed to be a coprocessor on the PCI Express (PCIe) I/O bus rather than be tightly integrated with a CPU, allowing it to plug into existing servers just as a GPU does. The goal was to run whole inference models in the TPU to reduce I/O between the TPU and the host CPU. Minimalism is a virtue of domain-specific processors. Jouppi et al. show in their paper that the TPU leverages its advantages to run 15 times as fast as the K80 GPU, resulting in a performance/ Watt advantage of 29 times. While future CPUs and GPUs will surely run inference faster, a redesigned TPU using circa-2015 GPU memory would go three times faster and boost the performance/ Watt advantage to nearly 70 over the K80, and 200 over Haswell CPU [63], [129].

3.4 Case Study: Image Captioning with Generative Adversarial Network

In recent years, investigators have been using Deep Learning to caption images with some success. However, the reported results suffer from several deficiencies, namely, accuracy, lack of diversity, and emotions in resultant captions. We propose using Generative Adversarial models to produce new and combinatorial samples to address some of these deficiencies. More specifically, we offer to explore various autoencoders to

[4]https://devblogs.nvidia.com/tensor-core-ai-performance-milestones/

generate more accurate and meaningful captions for images. Autoencoders are neural networks that learn data codings in an unsupervised manner. The research outlined in this study is an ongoing investigative research project[5].

The methods based on the only combination of CNN and LSTM lack diversity and naturalness in the generated captions. Researchers are investigating to improve the image captioning models by adding sentiment and diversity to have more human-like descriptions in recent past years. The Generative Adversarial Network model that Ian Goodfellow invented and his colleagues in 2014 [49] comes to help in this improvement. In this work, we propose a framework with GAN architecture for captioning an image. More specifically, we offer to explore various autoencoders to generate more accurate and meaningful captions for images. Autoencoders are neural networks that learn data codings in an unsupervised manner. The research outlined in this study is an ongoing investigative research project.

3.4.1 Background

Automating generating a natural description for an image is called Image Captioning, which can be utilized in many applications. For example, enhancing the accuracy of search engines, also recognition, and vision applications. How to extract visual information from the input and transform the visual information into a proper and meaningful language was the challenge. Captioning research started with the classical retrieval and template-based [41], [74] approaches. While latest approaches follow deep learning based on the Encoder-and-Decoder paradigm that adopts Maximum Likelihood Estimation (MLE) as their learning method [6]. Therefore, based on the encoder-and-decoder, many variants are proposed, where the attention mechanism appears to be the most effective add-on. We could also see many improvements in generating more natural and diverse descriptions in which, each time the system gets an image, we could describe it differently. We could observe these improvements in using one of the recent deep learning architectures called Adversarial Neural Network.

Adversarial Network Architecture:

In 2014, for the first time, Goodfellow et al. [49] proposed a new framework for estimating generative models via an adversarial process. Generative Adversarial Networks are machine learning models that can imagine new samples; in which they simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G (See Figure 3.5). GAN is drawing a sample from the probability distribution over all hypothetical images matching that description. Caption discriminator distinguishes between captions generated by caption generator and accurate captions generated by humans.

Most generative models are trained by adjusting parameters to maximize the probability that the generator net will generate the training data set. The discriminator is just a regular neural net classifier.

[5]Amirian, Soheyla, Khaled Rasheed, Thiab R. Taha, and Hamid R. Arabnia, "Image Captioning with Generative Adversarial Network," in Computational Science and Computational Intelligence; "Artificial Intelligence" (CSCI-ISAI); 2019 International Conference on IEEE CPS (IEEE XPLORE, Scopus), 2019, ISBN -13: 978-1-7281-5584-5, pp. 272-275.

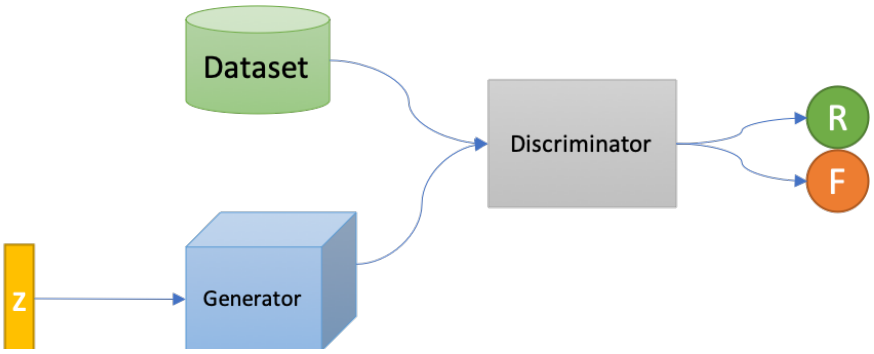


Figure 3.5: Generative Adversarial Network Architecture.

The generator takes random noise values z from a prior distribution P_z and maps them to output values x via function $G(z)$. Figure 3.5 illustrates the explanation. The goal of caption generator G is to generate a caption to achieve a maximum reward value from the caption discriminator. The objective function of D and G are respectively defined as:

$$\max_{D} \quad \mathbb{E}_{x\sim P_{data}}[logD(x)] + \mathbb{E}_{z\sim P_z}[log(1 - D(G(z)))],$$

$$\min_{G} \quad \mathbb{E}_{z\sim P_z}[log(1 - D(G(z)))].$$

Thus, the GAN is formulated as a $\min_G \max_D V(G, D)$, namely as:

$$V(G, D) = \mathbb{E}_{x\sim P_{data}}[logD(x)] + \mathbb{E}_{z\sim P_z}[log(1 - D(G(z)))].$$

3.4.2 Literature Review

This section provides relevant background on previous models and investigations that Generative Adversarial Networks have done.

In 2017, Dai et al. presented a new framework based on Conditional Generative Adversarial Networks (CGAN), which jointly learns a generator to produce descriptions conditioned on images and an evaluator to assess how well a description fits the visual content. They explore an approach to creating sentences with three properties: Fidelity in semantics, Naturalness, and Diversity. This work proposed a different task for the GAN method. Applying GANs to text generation is nontrivial. It comes with two significant challenges because of the unique nature of linguistic representation. First, challenging to apply back-propagation directly that devised via Policy Gradient, originating from reinforcement learning. Second, in the conventional GAN setting, vanishing gradients and error propagation in the training of the generator because of feedback from the evaluator, which tackled by getting early feedback by an approximated expected future reward through Monte Carlo rollouts. This framework not only results in a generator

that can produce natural and diverse expressions but also yields a description evaluator called G-GAN, which is substantially more consistent with human evaluation. This method is the first in applying the GAN method for image captioning [29].

In 2019, Nezami et al. propose ATTEND-GAN model. Their contribution is to generate human-like stylistic captions in a two-stage architecture by ATTEND-GAN using both the designed attention-based caption generator and the adversarial training mechanism on the SentiCap dataset. The architecture of the ATTEND-GAN model is spatial-visual features that are generated by ResNet-152 network and the caption discriminator is inspired by the Wasserstein GAN (WGAN) [84].

Arjovsky et al. claim that Wasserstein Generative Adversarial Network minimizes a reasonable and efficient approximation of the Earth Mover distance. Also, Wasserstein GAN can learn the distribution without mode collapse. Using the WGAN algorithm has significant practical benefits: a meaningful loss metric that correlates with the generator's convergence and sample quality; improved stability of the optimization process [15].

Makhzani et al. proposed Adversarial Autoencoder (AAE), a probabilistic autoencoder that uses generative adversarial networks to perform variational inference by matching the aggregated posterior of the hidden code vector of the autoencoder with an arbitrary prior distribution, which results in meaningful samples [82].

3.4.3 Our Approach

One of the important directions of this work would be generating more natural and diverse descriptions rather than the predefined templates. We are adding a sentiment to the captions to achieve naturalness. As discussed in [29], meaningful captions shall have three properties, fidelity, naturalness, and diversity, where the last two are essential properties of human language. However, most of the existing image captioning works mainly focus on the fidelity of the generated descriptions. Most recently, some works try to achieve natural and diverse captions by means of contrastive learning [30], conditional GAN [29] and variational auto-encoder [125].

In this research, we plan to utilize Variational Autoencoder. **Variational Autoencoder (VAE)** helps to map the input to a distribution. The usual bottleneck after the encoder is two vectors, mean vector and standard deviation vector, that represent the mean distribution and standard deviation distribution, respectively. Then we get the sample distribution of these vectors as an input vector to the decoder. In Disentangled Variational Autoencoder, [54] proposed hyperparameter β that directly affects the degree of learnt disentanglement. where

$$F(\theta, \phi, \beta; x, z) \geq (\theta, \phi; x, z, \beta) = E_{q\phi(x|z)}[logp_\theta(x \mid z)] - \beta D_{KL}$$

Varying β changes the degree of applied learning pressure during training, thus encouraging different learned representations.

Part of the implementation is from ATTEND-GAN [84] which has two core components: first, an attention-based caption generator to correlate different parts of an image with a caption; and second, an

adversarial training mechanism (WGAN) to help the caption generator to add diverse stylistic components to the generated captions. By combining VAE, WGAN, ATTEND-GAN, and sentiment to MSCOCO dataset [77] and SentiCap Dataset in our experiment, we are hoping to get a better and more meaningful description for each image.

3.4.4 Discussion

Image captioning is the process of automatically generating captions that describe the content of the image. A system that automatically creates image captions can be utilized in many applications. Examples include: enhancing the accuracy of search engines; recognition and vision applications. In recent years, researchers have been exploring Deep Learning to caption images with some success. However, the reported results suffer from several deficiencies, namely, accuracy, lack of diversity, and emotions in resultant captions. We propose using Generative Adversarial models to produce new and combinatorial samples to address some of these deficiencies. More specifically, we offer to explore various autoencoders to generate more accurate and meaningful captions for images. The research outlined in this study is an ongoing investigative research project.

3.5 Summary

Many models have already been presented to generate meaningful captions for images. These models are pretty good but have some constraints. Image captioning still has a long way to go in improving the accuracy of captioning the events in images (See Figure 3.3). We reviewed some of the recent deep learning-based works. It is hard to compare different works due to the different combinations of structures, using different parameters and implying various datasets. We also noticed that there is a lot of room for improving accuracy. By improving image captioning models, we can further aid people with hearing or sight impairments and improve search engines.

Chapter 4

Automatic Image and Video Caption Generation with Deep Learning: Algorithmic Overlap

Methodologies that utilize Deep Learning offer sophisticated potential for applications that automatically attempt to generate captions or descriptions about images and video frames. Image and video captioning are considered to be intellectually challenging problems in imaging science. The application domains include automatic caption (or description) generation for images and videos for people who suffer from various degrees of visual impairment; the automatic creation of metadata for images and videos (indexing) for use by search engines; general-purpose robot vision systems; and many others. Each of these application domains can positively and significantly impact many other task-specific applications. This chapter is not meant to be a comprehensive review of image captioning; rather, it is a concise review of both image captioning and video captioning methodologies based on deep learning. This study treats both image and video captioning by emphasizing the algorithmic overlap between the two[1].

Describing a scene in an image or a video clip is a highly demanding task for humans. To create machines with this capability, computer scientists have been exploring methods to connect the science of understanding human language with the science of automatic extraction and analysis of visual information. Image captioning and video captioning need more effort than image recognition, because of the additional challenge of recognizing the objects and actions in the image and creating a succinct meaningful sentence based on the contents found. The advancement of this process opens up enormous opportunities in many application domains in real life, such as aid to people who suffer from various degrees of visual impairment, self-driving vehicles, sign language translation, human-robot interaction, automatic video subtitling, video surveillance, and more. This chapter surveys the state-of-the-art approaches, focusing on deep learning models for image and video captioning. The models and the generated captions are evaluated by using BLEU, METEOR, CIDEr [24], [121], [131], and other evaluation metrics.

[1]Amirian, Soheyla, Khaled Rasheed, Thiab R. Taha, and Hamid R. Arabnia, "Automatic Image and Video Caption Generation with Deep Learning: A Concise Review and Algorithmic Overlap," in IEEE Access, vol. 8, pp. 218386-218400, 2020.

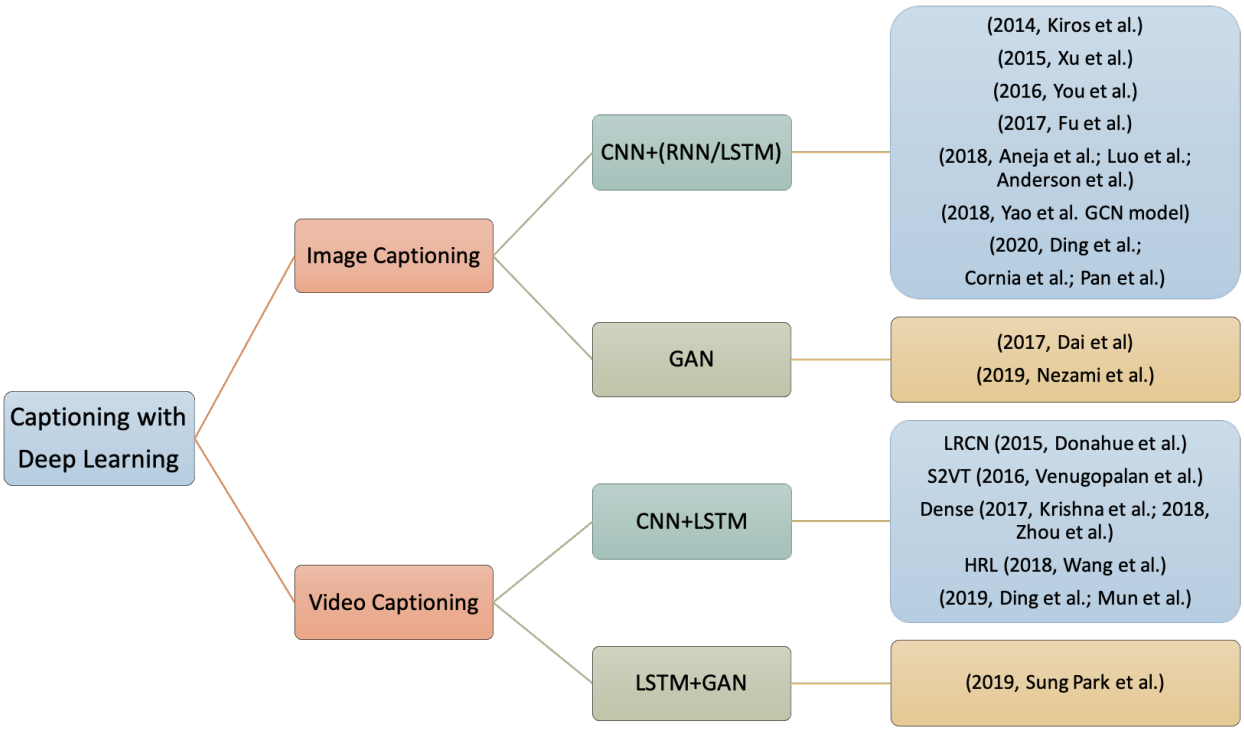


Figure 4.1: The Taxonomy of the reviewed works in this research.

This chapter summarizes both image and video captioning methodologies based on deep learning, focusing on the algorithmic overlap between the two. This review begins by introducing the Image and Video Captioning in Section 4.1. Then, a few recent methods of Image and Video Captioning, their Datasets, and evaluation metrics are discussed in Section 4.2. To facilitate the discussions about image and video captioning, we use the taxonomy shown in Figure 1. Figure 1 shows the conventional methods currently utilized in image and video captioning as well as the corresponding and relevant publications.

4.1 Image and Video Captioning

Many impressive studies have been done about image captioning [41], [65], [128], [134]. Image captioning is often regarded to be the process of generating a concise description of objects and/or information about the scenes in an image. Some examples (images and their corresponding captions) are shown in Figure 4.2. Often, captions of images are generated manually. Automating this process would be a significant contribution. A system that automatically generates image captions can be utilized in many applications. Examples include: enhancing the accuracy of search engines; recognition and vision applications; enriching and creating new image datasets; enhancing the functionality of systems similar to Google Photos; and enhancing the optical system analysis of self-driving vehicles. In image captioning, the main challenges include extracting visual information from the picture and transforming this visual information into a proper and meaningful language. Captioning research started with the classical retrieval [41] and template-based [74] approaches in which Subject, Verb, and Object are detected separately and then joined using a sentence template. However, the advent of Deep Learning and the tremendous advancements in Natural Language Processing have equally and positively affected the field of captioning. Hence, the latest approaches follow deep learning-based architectures that encode the visual features with Convolutional Neural Networks and decode with a language-based model, which translates the features and objects given with an image-based model to a meaningful sentence. We dissect the image captioning process and models in Section 4.2.

Video description is the automatic generation of meaningful sentences that describes the events in a video. Many researchers present different models on video captioning [36], [69], [116], [122], [127], mostly with limited success and many constraints. Video captioning can also be achieved by applying image captioning methods to the video frames as images. The advancement of video description opens up opportunities in a wide range of applications like human-robot interaction, automatic video subtitling, and video surveillance. Section 4.2 provides a detailed discussion of the video captioning process and recent models.

4.2 Captioning Methodology

Automatically generating natural language sentences describing an image or a video clip generally has two components: Encoder and Decoder. Here we specifically explain the architecture of each part. The Encoder utilizes a Convolutional Neural Network, which extracts the objects and features from an image

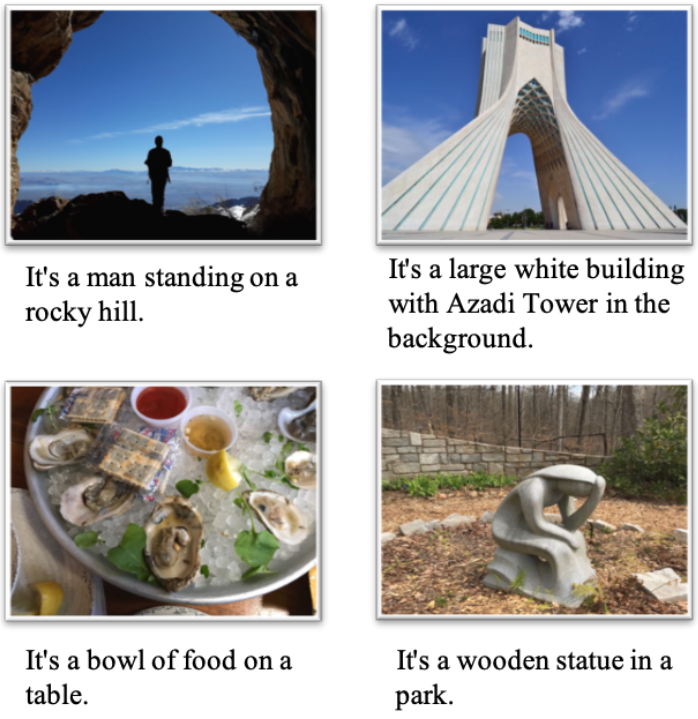


Figure 4.2: Some examples of image captioning. Each caption describes the image above it. These captions are generated with the model presented in [96] and the images are taken by the authors.

or video frame. A neural network is needed for the Decoder to generate a natural sentence based on the available information.

Convolutional Neural Network: A model with a large learning capacity to learn about thousands of objects from a large number of images [71] is needed. Deep learning presents computational models that are composed of multiple processing layers to learn representations of data in images [9], [73]. Deep learning-based Convolutional Neural Networks play a key role in many applications, one of which is image recognition (See Figure 4.3). Image recognition is used to perform many visual tasks, such as understanding the content of images. Several well-known models [9] in the field of CNNs based on object detection [46], [47], [94] and segmentation [100] exist that are heavily used in image captioning and video captioning architecture to extract the visual information.

Recurrent Neural Networks: Sequence models like recurrent neural network (RNN) [26] have widely been utilized in speech recognition, natural language processing, and other areas. In addition, sequence models can address supervised learning problems like machine translation [25], name entity recognition, DNA sequence analysis, video activity recognition, and sentiment classification. Gated recurrent unit (GRU) is a gating mechanism in RNN, introduced by Cho et al. [25] in 2014. The basic RNN algorithm runs into a vanishing gradient problem (a difficulty in training artificial neural networks).

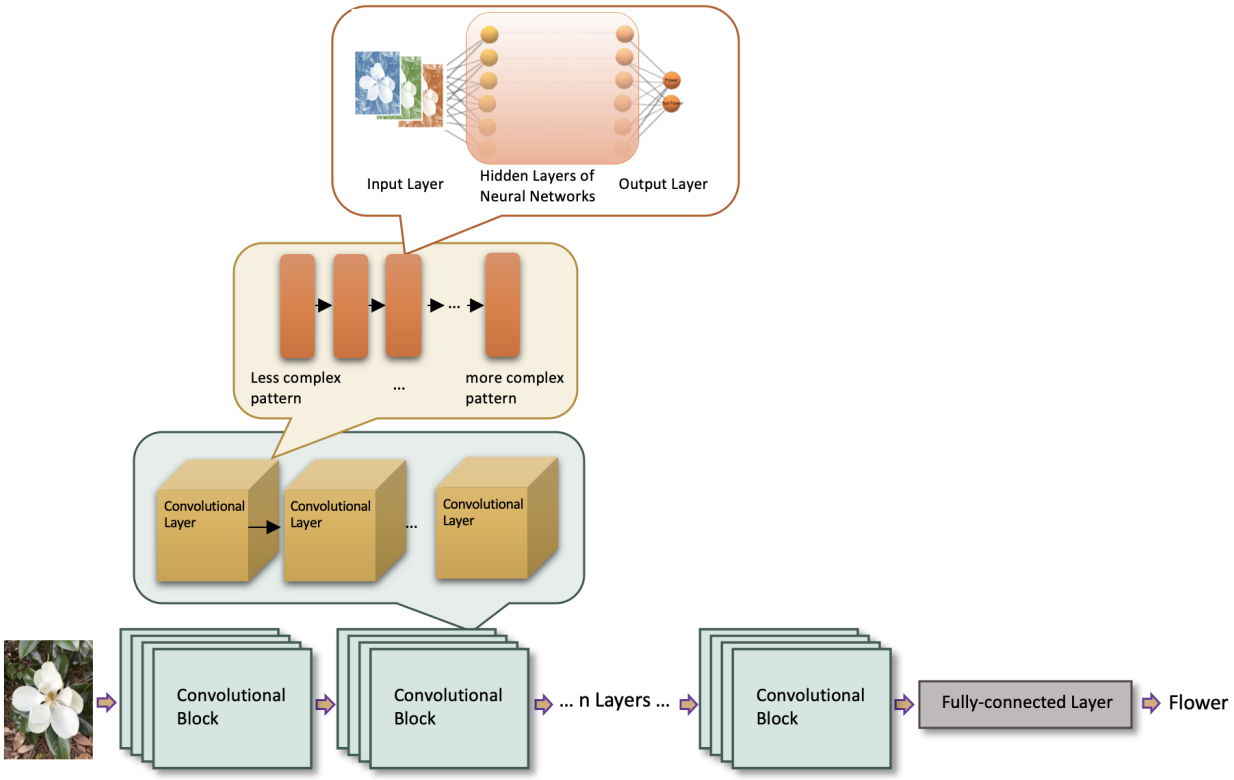


Figure 4.3: Overall architecture of Convolutional Neural Network that shows each Convolutional Block consists of n Convolutional layers and each of these Convolutional layers is built up of convolutions with filters.

The gated recurrent units are an effective solution for addressing the vanishing gradient problem. Furthermore, they allow neural networks to capture much longer range dependencies [26]. The advantage of the GRU is that it is a simple model. Therefore it is easy to build a big network with GRU. Also, it only has two gates, as a result, it computes quickly.

Long Short Term Memory: LSTM, as a special RNN structure, has proven to be stable and powerful for modeling long-range dependencies in various studies. LSTM can be adopted as a building block for complex structures. The complex unit in Long Short Term Memory is called a memory cell. Each memory cell is built around a central linear unit with a fixed self-connection [55]. LSTM is historically proven more powerful and more effective than a regular RNN since it has three gates (forget, update, and output). Long Short Term Memory can be used to generate complex sequences with long-range structure [68], [129].

4.2.1 Image Captioning Methodologies:

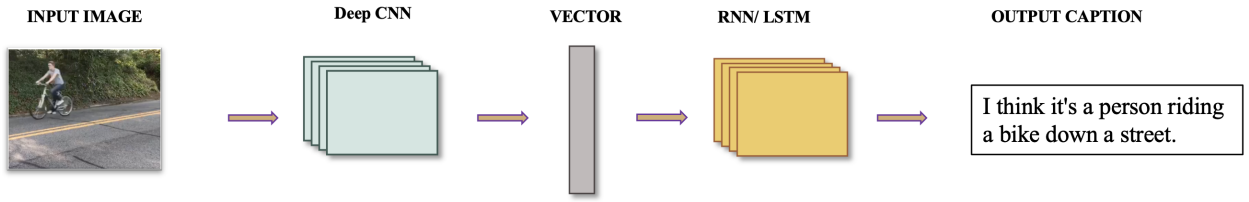


Figure 4.4: The early attempts of image captioning as an active research area exploit the encoder-decoder architecture. A deep learning model encodes the image into a feature vector. The language model takes the input vector to generate a sentence that describes the image, leading to promising results for this task.

Many methods for image captioning are there. Earlier methods, prior to deep neural networks (DNNs), were retrieved-based [41] or template-based [74] models. Recent methods are based on deep neural networks. Generating an automatic caption for describing an image has two stages. First, the information needs to be extracted from the image and put it in a feature vector. This stage focuses on visual recognition through deep learning models. Then the feature vector is fed into the second stage. The second stage is caption generation which is describing what is extracted in a grammatically correct natural language sentence (See Figure 4.4). So, we classified DNN-based methods based on the main framework into subcategories that they respectively use. Here, a review of recent deep learning-based works for automatic image captioning is discussed. All are summarized with more details about the evaluation results in Table 4.1.

A breakthrough in image and video captioning occurred in 2014 through the application of encoder-decoder models. Kiros et al. [68] introduced an encoder-decoder pipeline model in which an encoder network takes the image or video as an input and extracts a fixed-size feature vector that a decoder network maps to a sequence of words. They set new best results when using the 19-layer Oxford convolutional network. Then, a series of innovations such as attention mechanism have been introduced to boost image captioning by encouraging more interactions between the two different modalities. They were

Table 4.1: The summary of a few recent works for Image Captioning.

Model	Architecture	Evaluation (on MSCOCO)	Comment
(2014, Kiros et al.)	CNN+LSTM encoder–decoder Attention-based	Image Annotation result R@1 R@5 R@10 Med r 23.0 50.7 62.9 5 (OxfordNet) on Flickr30K. R@K is Recall@K (high is good). Med r is the median rank (low is good).	The generated descriptions are arguably the nicest ones to date.
(2015, Xu et al.)	CNN+RNN Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR 71.8 50.4 35.7 25.0 23.04	They encourage future work in using visual attention.
(2016, You et al.)	CNN+RNN Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR 0.709 0.537 0.402 0.304 0.243 Using the ground-truth visual attributes	They need to experiment with phrase-based visual attributes with their distributed representations.
(2017, Fu et al.)	VGG/Alex/ResNet + LSTM Attention-based	ENSEMBLE result BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE-L CIDEr-D 72.4 55.5 41.8 31.3 24.8 53.2 95.5	Either region-based attention or scene-specific contexts improve performance. Combining these two modeling ingredients provides a further improvement.
(2017, Dai et al)	GAN VGG/G-MLE/G-GAN	BLEU-3 BLEU-4 METEOR ROUGE L CIDEr SPICE E-NGAN E-GAN G-MLE: 0.393 0.299 0.248 0.527 0.1020 0.199 0.464 0.427 G-GAN: 0.305 0.207 0.224 0.475 0.795 0.182 0.528 0.602	E-NGAN regard G-GAN as the best generator. This framework also provides an evaluator that is more consistent with human's evaluation.
(2018, Aneja et al.)	CNN+LSTM (ResNet152)	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE-L CIDEr 0.725 0.555 0.41 0.299 0.251 0.532 0.972	ResNet is capable of encoding better feature vectors for images. The Meteor and CIDEr results are comparable with previous works.
(2018, Luo et al.)	ResNet101 + LSTM ATTN+CIDER+DISC	BLEU-4. ROUGE. METEOR CIDEr SPICE 0.3274 0.2574 **0.5457** 1.0231 0.1939	Incorporating a discriminability loss, in training image caption generators improves the quality of resulting captions. Their model needs more sophisticated visual semantic embedding model.
(2018, Anderson et al.)	CNN+LSTM Faster R-CNN; ResNet101 Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE-L CIDEr SPICE c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 80.2 95.2 64.1 88.8 49.1 79.4 36.9 68.5 27.6 36.7 57.1 72.4 117.9 120.5 21.5 71.5	They obtained first place in the 2017 VQA Challenge.
(2018, Yao et al.)	GCN+LSTM ResNet101 Attention-based	BLEU-1 BLEU-4 METEOR ROUGE-L CIDEr-D SPICE 80.9 38.3 28.6 58.5 128.7 **22.1**	They build graphs over the detected objects in an image based on their spatial and semantic connections.
(2019, Nezami et al.)	ATTEND-GAN	BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-L METEOR CIDEr SPICE 56.55 33.85 20.80 13.05 44.45 18.35 62.85 16.05	It also adds sentiment and naturalness to the sentences.
(2020, Ding et al.)	CNN+LSTM (VGG-19)	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE CIDEr 0.748 0.525 0.365 0.235 0.235 0.505 1.041	They introduced the theory of attention in psychology to image caption generation.
(2020, Cornia et al.)	CNN+LSTM Faster R-CNN; ResNet101 Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE CIDEr c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 81.6 **96.0** 66.4 **90.8** 51.8 **82.7** 39.7 **72.8** 29.4 39.0 59.2 **74.8** 129.3 132.1	Novelty in using a stack of memory-augmented encoding layers and a stack of decoder layers.
(2020, Pan et al.)	CNN+LSTM SENet-154 Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE CIDEr c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 **81.9** 95.7 **66.9** 90.5 **52.4** 82.5 **40.3** 72.4 29.6 39.2 **59.5 75** **131.1 133.5**	They used a novel unified X-Linear attention block for image captioning. Using SENet-154 makes it proceed other models.

developing an attention-based model that jointly learns to align parts of captions to images. The generated descriptions are arguably the nicest ones to date [68]. The attention model is one of the models used in deep learning that got from one of the most curious facets of the human visual system. The attention-based model learns to focus on different parts of the image. This is crucial when much clutter is in an image. However, this may cause losing information which could be helpful for richer and more descriptive captions. Xu et al. [131] proposed the attention-based approach that gives the state of the art performance on three benchmark datasets using the BLEU and METEOR metric (See section 4.2.5). They showed how the learned attention can be exploited to give more interpretability to the model generation process and demonstrate that the learned alignments correspond well to human intuition. Their model encourages future work in using visual attention. Next, You et al. [134] proposed a model of semantic attention that learns to focus on the semantic attributes in the image selectively. The algorithm combines top-down and bottom-up strategies to extract richer information from the image. It fuses them with an RNN that can selectively attend to rich semantic attributes detected from the image. They performed their method on different datasets, and the captioning system was implemented based on the LSTM network. The image feature vector is extracted from the last 1024 dimensional convolutional layer of the GoogleNet [9] CNN model. Furthermore, their framework employs attention at both input and output layers to the RNN module. Their effort was to exploit abundant fine-grain visual semantic aspects and fuse global and local information to generate a better caption. The results show that the algorithm significantly outperforms the state-of-the-art approaches consistently across different evaluation metrics. We see in the next research, Fu et al. [43] proposed the image caption system that exploits the parallel structures between images and sentences. One contribution of this system is that it aligns the process of generating captions and the attention shifting among the visual regions. Another is that it introduces the scene-specific contexts to LSTM that adapt language models for word generation to specific scene types. An image is first analyzed and represented with multiple visual regions from which visual features are extracted in that system. The visual feature vectors are then fed into an LSTM network that predicts the sequence of focusing on different regions and the sequence of generating words based on the transition of visual attention. A scene vector also governs the neural network model, a global visual context extracted from the whole image. Intuitively, it selects a scene-specific language model for generating text. They evaluated captions in BLEU-n, METEOR, ROUGE-L and CIDEr-D metrics by testing on several popular datasets, including the MSCOCO, Flickr8K, and Flickr30K (See Table 4.1). Either region-based attention or scene-specific contexts alone improve performance but combining the two provides a further improvement.

Researching more with CNN and LSTM models, in 2018, Aneja et al. [12] developed a convolutional image captioning technique with existing LSTM techniques and analyzing the differences between RNN based learning and their method. This technique contains three main components. The first and the last components are input/output word embeddings, respectively. However, while the middle component contains LSTM or GRU units in the RNN case, masked convolutions are employed in their CNN-based approach. This component is feed-forward without any recurrent function. Their CNN with attention (Attn) achieved comparable performance. They also experimented with an attention mechanism with attention parameters using the conv-layer activations. The results on CNN+Attn method were improved

relative to the LSTM baseline. For better performance on the MSCOCO they used ResNet features and the results show that ResNet boosts the performance. The results on the MSCOCO with Resnet101 and Resnet152 were comparable to previous works. Table 4.1 shows that the METEOR and CIDEr results are outstanding, therefore better captions. Then, we see Ding et al. [35] introduced the same architecture of CNN by VGG-19 and LSTM on the MSCOCO dataset, but with the theory of attention [131] in psychology to image caption generation with two types of attention mechanisms: The stimulus-driven for monitoring salient information by Color stimulus-driven, Dimension stimulus-driven and location perception stimulus-driven for attention detection. And the concept-driven is a classical question-guided attention mechanism. This approach enhances the encoder framework to suit complex scenes.

At the same year, Luo et al. [80] presented a method that has been trained with the COCO dataset for the Encoder part. For the image encoder in retrieval and FC captioning model, Resnet-101 is used. The spatial features are extracted from the output of a Faster R-CNN with ResNet-101, trained with an object and attribute annotations from Visual Genome [70]. The retrieval model uses GRU-RNN to encode text. The captions generated with this model describe valuable information about the images. However, richer and more diverse sources of training signals may further improve the training of caption generators. For experimenting with the output, we implemented their method, and some results are shown in Figure 4.5.

Figure 4.5: These captions are generated with the model presented in [80] and the images are scenes from the ActivityNet dataset.

Also, Anderson et al. [11] proposed a combined bottom-up and top-down attention mechanism that enables attention to be calculated at the level of objects and other salient image regions. The bottom-up attention uses Faster R-CNN with ResNet-101 [9], which represents a natural expression of a bottom-up attention mechanism. The top-down mechanism uses a task-specific context to predict an attention distribution over the image regions. The attended feature vector is then computed as a weighted average of image features over all regions. Their results on the MSCOCO dataset present a new state-of-the-art for the task, achieving CIDEr, BLEU-4 scores of 117.9, and 36.9, respectively. Demonstrating the broad applicability of the method, they applied the same approach to Visual Question Answering and obtained first place in the 2017 VQA Challenge. In a novel architecture, Yao et al. [133] proposed Graph Convolutional Networks plus Long Short-Term Memory (GCN-LSTM) model. This model integrates both semantic and spatial object relationships into the image encoder, which more remarkably increases CIDEr-D performance on the COCO testing set.

We also have Cornia et al. [28] which proposed a Transformer-based architecture. The architecture is composed of a stack of memory-augmented encoding layers and a stack of decoder layers. Image regions and their relationships are encoded in a multi-level fashion, in which low-level and high-level relations are taken into account. The model can learn and encode a priori knowledge by using persistent memory vectors. The generation of the sentence, done with a multi-layer architecture, exploits low-level and high-level visual relationships instead of just a single input from the visual modality. This goal is achieved through a learned gating mechanism, which weights multi-level contributions at each stage. They name this model Meshed-Memory Transformer as this creates a mesh connectivity schema between encoder and decoder layers. Based on their results, this approach achieves a new state of the art on COCO, ranking first in the on-line leaderboard.

The same year, Pan et al. [85] presented a unified attention block or X-Linear attention block that employs bilinear pooling to capitalize on visual information or perform multimodal reasoning selectively. In addition, they offered X-Linear Attention Networks that novelly integrates X-Linear attention block(s) to leverage higher-order intra- and inter-modal interactions. The experiments on the COCO benchmark show that their X-LAN obtains the best-published CIDEr performance of 132.0% on the COCO Karpathy test split so far. Moreover, by endowing Transformer with X-Linear attention blocks, CIDEr is boosted up to 132.8%.

The models mentioned above are all heavily utilized in image caption generation. Many other deep learning models have the potential to be used for applications such as image caption generation; one such model is Generative Adversarial Network. In 2014, Goodfellow et al. proposed a new framework for estimating generative models via an adversarial process for the first time. They simultaneously train two models. A generative model G captures the data distribution. A discriminative model D estimates the probability that a sample came from the training data rather than G. GAN has been successfully used in image generation. They can produce natural images almost indistinguishable from real photos [5], [49], [58]. Dai et al. [29] presented a new framework based on Conditional Generative Adversarial Networks (CGAN), which jointly learns a generator to produce descriptions conditioned on images and an evaluator to assess how well a description fits the visual content. This work proposed a different task

for the GAN method. They have a strategy stemming from Reinforcement Learning, which allows the generator to receive early feedback along the way. In their method, they implemented G-MLE: a generator trained based on MLE that is used to produce the descriptions, and G-GAN, the same generator, which is based on the conditional GAN formulations. For both G-MLE and G-GAN, VGG16 is used as the image encoder. They considered multiple evaluation metrics, including six conventional metrics BLEU-n, METEOR, ROUGE L, CIDEr, SPICE, and two additional metrics relevant to their formulation: E-NGAN and E-GAN particularly using their framework. This method was the first to apply GAN. We believe GAN has significant potential in image captioning. In 2019, Nezami et al. [84] proposed the ATTEND-GAN model. Their contribution is to generate human-like stylistic captions in a two-stage architecture, with ATTEND-GAN using both the designed attention-based caption generator and the adversarial training mechanism on the SentiCap dataset. The ATTEND-GAN model's architecture uses spatial-visual features generated with the ResNet-152 network, and the caption discriminator is inspired by the Wasserstein GAN (WGAN).

So far, we briefly reviewed a few methods according to the standard approaches that they have used. For a fair comparison of the models, Table 4.1 shows the results of attention-based methods on the MS COCO dataset, the common dataset that they have utilized. With this comparison, we could state that Anderson et al. performed well on the MS COCO dataset. Furthermore, their method outperformed previous works. It uses the attention mechanism, which focuses only on relevant objects of the image. Also, We found that the performance of a technique can vary across different metrics, parameters, and datasets. Here, we tried to analyze them based on the various methods they have used. However, image captioning remains active research, and it has a long way to go in improving the accuracy of captioning the information in images (See Figure 4.6).

4.2.2 A.1 Image Captioning Datasets:

A few datasets are widely used to evaluate and compare image captioning methods: Flickr8K [65], Flickr9K [131], Flickr30k [65], [131] and Microsoft COCO [24], [131].

- **Flickr:** The Flickr8K, 9k, and 30k datasets contain more than 8000, 9000, and 30000 images, respectively. Each image is annotated using Amazon Mechanical Turk with five independent sentences. The Flickr8K dataset mainly contains human and animal images, while the Flickr30k dataset contains humans involved in everyday activities and events. For each image, five sentences are provided [65], [131].

- **COCO:** Lin et al. [77] presented a new dataset for detecting and segmenting objects found in everyday life in their natural environments. Microsoft Common Objects in COntext (MS COCO) dataset contains a total of 2.5 million labeled instances in 328k images, 91 object categories with 82 of them having more than 5,000 labeled instances, and five assigned captions to each image [24], [131].

Figure 4.6: Examples of poor image captioning generated by state-of-the-art systems. These captions are generated with the model presented in [37] and the images are taken by the authors.

Figure 4.7: Video: Keyframes and Frames in-between the Keyframes (Keyframe is a frame used to indicate the beginning or end of a change made to a parameter).

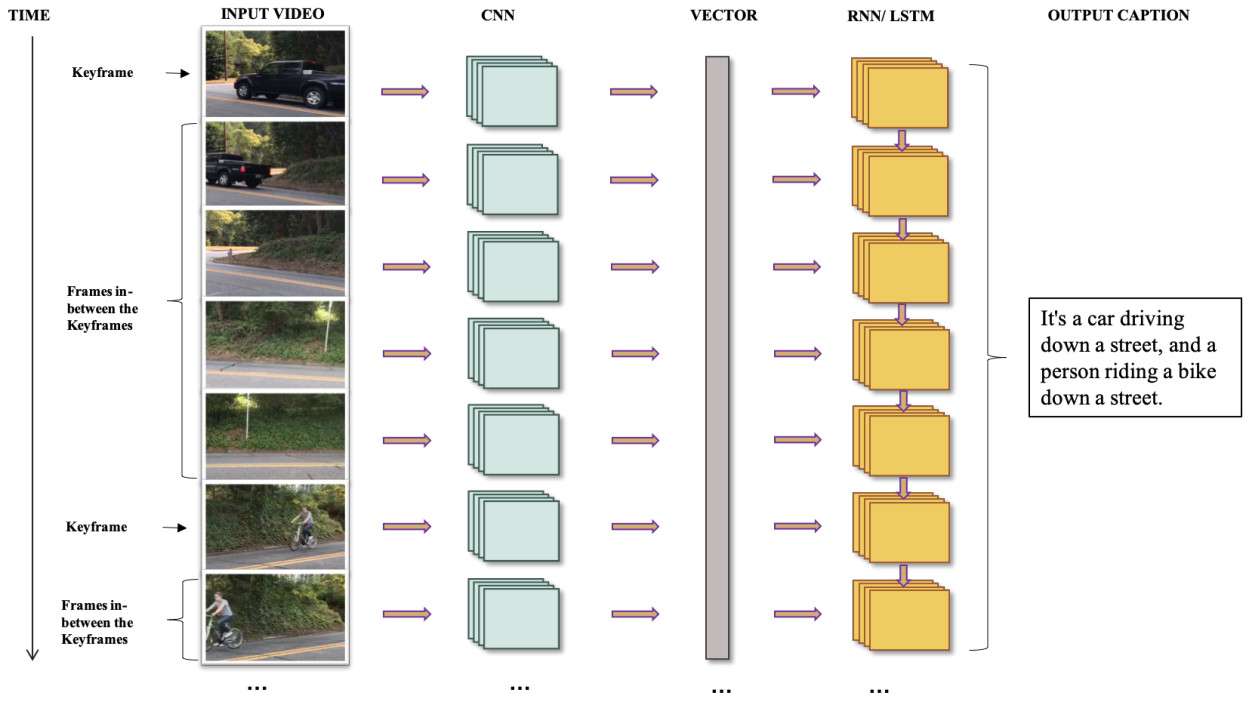


Figure 4.8: This is a basic structure for video captioning models. Each DCNN takes a frame of the video as an image, then encodes the frame into a common feature vector between all the other video frames. The language model takes the vector to generate a sentence or a paragraph that describes the video.

4.2.3 Video Captioning Methodologies:

Describing a video in natural language is a trivial task for most people but challenging for machines. Furthermore, from the methodological perspective, categorizing the models or algorithms is challenging because it is difficult to assert the contributions of the visual features and the adopted language model to the final description.

Video captioning can be achieved by applying image captioning (as discussed in Section 4.2.1) to the video keyframes and a small sample of the frames in-between the keyframes (See Figure 4.7). The encoder-decoder framework discussed for image captioning can also be extended to video captioning (compare Figure 4.4 and 4.8). Overall, generating natural language sentences describing the video content automatically has two stages. The first stage is understanding the objects. This stage focuses on visual recognition with deep learning models and extracts the performer, action, and the object of the action (e.g. human and activity detection) from the video clip. Next, the video clip is fed as a series of frames that are considered images. So, we have a series of frames in each clip that are input images. Then the extracted information from the clip is put in a common feature vector. Finally, this vector is fed into the second stage. The second stage is caption generation, which describes what is extracted in a grammatically correct natural language sentence, thus mapping the objects identified in the first stage. Here, we bring a combination of deep learning architectures for the encoding and decoding stages (See Table 4.2).

One of the most common architectures in deep learning that is used for video captioning is a combination of CNN and RNN models. Donahue et al. proposed Long-term Recurrent Convolutional Networks (LRCNs). LRCNs are a model for visual recognition and description. It combines convolutional layers and long-range temporal recursion and is end-to-end trainable. They considered three vision problems: activity recognition, image description, and video description. LRCN processes the variable-length visual input with a CNN, whose outputs are fed into a stack of recurrent sequence models, which finally produce a variable-length prediction. They evaluated the image architecture on the COCO and Flickr30k datasets, using BLEU to measure the descriptions' similarity. Finally, they evaluated the video description approach on the TACoS multilevel dataset, using the BLEU-4 metric for scoring the results. The advantage of using LSTM here is that it allows them to model the video as a variable-length input stream. Although the LSTM outperformed the statistical model-based approaches, it was still not trainable in the end-to-end fashion [36]. We see that Venugopalan et al. [122] used the S2VT method (a sequence to sequence approach for a video to text), which is a combination of CNN and LSTM models. The S2VT architecture encodes a sequence of frames and decodes them into a sentence. They compared their model on the YouTube dataset, MPII-MD, and M-VAD (See Table 4.3). They evaluated the performance using METEOR and BLEU to compare the machine-generated descriptions to human ones. The results show significant improvements in human evaluations of grammar. Language alone is considerable; hence, it is essential to focus on both language and visual aspects to generate better descriptions. Later methods have adopted a similar framework, including attention mechanisms [116].

Deep learning has achieved much better results than previous models, and most methods are aimed at producing one sentence from a video clip containing only one bold event. Krishna et al. [69], however, presented Dense-captioning, which focuses on detecting multiple events that occur in a video by jointly

Table 4.2: The summary of a few recent works for Video Captioning.

Model	Architecture	Evaluation	Comment
LRCN (2015, Donahue et al.)	Long-term recurrent convolutional networks.	BLEU 28.8 on TACoS	It was still not trainable in an end-to-end fashion.
S2VT (2016, Venugopalan et al.)	A sequence to sequence approach. CNN+LSTM.	BLEU-4 METEOR **42.1** **31.4** on Youtube, MPII-MD and M-VAD	The contribution of language alone is considerable.
Dense (2017, Krishna et al.)	Attention mechanism. Dense-captioning, multiple events.	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR CIDEr 26.45 13.48 7.12 3.98 9.46 24.56 on ActivityNet	No single-sentence generation scenario.
(2018, Zhou et al.)	CNN+LSTM Attention based, Dense-captioning.	BLEU-3 BLEU-4 METEOR 4.76 2.23 10.12 End-to-end Masked Transformer, on ActivityNet	This model is able to produce proposal and description simultaneously.
HRL (2018, Wang et al.)	Hierarchical Reinforcement Learning. attention module.	BLEU-4 METEOR ROUGE-L CIDEr 41.3 28.7 **61.7** **48.0** on MSR-VTT	Outperformed all the other algorithms. Still needs a boost.
(2019, Ding et al.)	CNN+LSTM	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE-L **76.4** **56.3** **43.6** 31.7 26.5 53.5 on MSCOCO	The result that has been provided is for the image caption part.
(2019, Mun et al.)	A streamlined approach. C3D+GRU+RNN	BLEU-1 BLEU-2 BLEU-3 BLEU-4 CIDEr METEOR 17.92 7.99 2.94 0.93 30.68 8.82 on ActivityNet	Algorithm generates captions for events sequentially conditioned on the prior ones by detecting highly correlated events in a video.
(2019, Sung Park et al.)	LSTM+GAN	BLEU-4 METEOR CIDEr-D 10.02 16.69 21.07 GAN results on ActivityNet	A hybrid discriminator consists of three individual experts: language, one for relating the sentence to the video, and one pairwise, across sentences.

localizing temporal proposals of interest and then describing each with natural language. This model introduced a new captioning module that uses contextual information from past and future events to describe all events jointly. They implemented the model on the ActivityNet Captions dataset (See Table 4.3 and Section 4.2.4). The captions that came out of ActivityNet shift sentence descriptions from being object-centric in images to action-centric in videos. It does not aim to solve the single-sentence generation scenario, though.

The most similar work to Krishna et al. in using dense video captioning model is Zhou et al.[135] model. However, this model proposed an end-to-end transformer model for dense video captioning and is composed of an encoder and two decoders. The captioning decoder employs a masking network to restrict its attention to the proposal event over the encoding feature, which converts the event proposal to a differentiable mask to ensure the consistency between the proposal and captioning during training. Furthermore, this model employs a self-attention mechanism.

Another line of work is deep reinforcement networks, a relatively new research area for video description. Wang et al. [127] presented the Hierarchical Reinforcement Learning method that aims to generate one or more sentences for a sequence of one or more continuous actions. In this model, both the encoder and decoder are equipped with an attention module. The novel HRL method outperformed all the other algorithms on all metrics. Hence, the HRL agent needs more exploration in terms of attention space and utilizing features from multiple modalities.

In 2019, Ding et al. [34] proposed novel techniques for the application of long video segmentation, which can effectively shorten the retrieval time. First, redundant video frame detection based on the Spatio-temporal interest points (STIPs) and a novel super-frame segmentation are combined to improve the effectiveness of video segmentation. After that, the super-frame segmentation of the filteblue long video is performed to find the exciting clip of a long video. Then, keyframes from the most impactful segments are converted to video captioning using the saliency detection and LSTM variant network. Finally, the attention mechanism is used to select more crucial information to the traditional LSTM. This method is benchmarked on the VideoSet dataset and evaluated with the BLEU, Meteor, and Rouge on the image captioning part. However, the language model still has a significant performance gap from humans in small object recognition or object recognition at lower resolutions. Similar to Krishna et al. [69] work, Mun et al. [83] proposed a dense video captioning framework that models temporal dependency across events in a video explicitly and leverages visual and linguistic context from previous events for coherent storytelling. They have used the Single-Stream Temporal Action model to get some proposals at a single scan. By implying PtrNet, the highly correlated events that make an episode fed into a sequential captioning network produce a caption by RNN systems. The proposed technique achieves outstanding performances on the ActivityNet Captions dataset in terms of METEOR. By injecting GAN to DL, Sung Park et al. [113] applied Adversarial Networks in their framework by designing a discriminator to evaluate visual relevance to the video, language diversity, fluency, and coherence across sentences. Thus, GAN helps to generate more accurate, diverse, and coherent multi-sentence video descriptions. The task of the discriminator (D) is to score the captions generated with the generator (G) for a given video. They

propose to compose D out of three separate discriminators, each focusing on one of the above tasks. They denote this design as a hybrid discriminator.

In this section, we reviewed a few methods ordered chronologically according to the recent techniques of CNN, LSTM, and attention-based that they have used. Table 4.2 shows the performance of these methods. We do not intend to compare them because they are using different approaches, techniques, and datasets. Nevertheless, the performance and accuracy are getting better each year due to the methods, extensive datasets and captions assigned, and the advancements in hardware.

4.2.4 Video Captioning Datasets:

Many datasets are used to evaluate video captioning methods. Here, we mention just a few of them and classify them into five domains based on the video contents: People, Open Subjects, Social Media, Cooking, and Movie (See Table 4.3).

- **People:** The Charades dataset [110] is built up by combining 40 objects and 30 actions in 15 scenes. Sigurdsson et al. proposed Charades, which contains 9,848 videos (7,985 for training and 1,863 videos for test purposes) with an average length of 30 seconds of people's daily activities. The dataset comprises of 66,500 annotations describing 157 actions. It also provides 27,847 descriptions covering all the videos.

Table 4.3: Some of the Video Caption Datasets.

Domain	Dataset	Total duration
People	Charades [110]	82h
Open	MSVD [21], ActivityNet Captions [69], MSR-VTT [130]	5.3h, 849h, 41.2
Social Media	VideoStory [45]	396h
Cooking	MPII [99], TACoS [95], YouCook2 [135]	490m, 15.9h,176h
Movie	LSMDC [98], MPII-MD [97], M-VAD [116]	158h, 73.6h, 84.6h

- **Open Subject:** The Microsoft Video Description dataset (Chen and Dolan, 2011) contains 1,970 YouTube clips (1,200 videos for training, 100 videos for validation, and 670 videos for testing) with human-annotated sentences. The duration of each video in the MSVD dataset is typically between 10 to 25 seconds. On average, 41 descriptions for each video [21] are there. Krishna et al. [69] presented the ActivityNet Captions dataset, a large-scale benchmark for dense-captioning events, which contains 20k videos amounting to 849 hours with 100k total descriptions. Xu et al. [130] presented the MSR-VTT dataset (standing for MSR-Video to Text). This is created by collecting 257 popular queries from a commercial video search engine, with 118 videos for each query. MSR-VTT provides 41.2 hours of 10K web video clips with 200K clip-sentence pairs in total, covering a list of 20 categories.

- **Social media:** VideoStory [45] is a dataset for telling the stories of social media videos. It contains 20k videos amounting to 396 hours of video with 123k sentences.

- **Cooking:** Max Plank Institute for Informatics (MPII) Cooking dataset [99] presents 65 fine-grained cooking activities. The dataset is comprised of 44 videos with an average length of 600 seconds per clip. Regneri et al. [95] presented Textually Annotated Cooking Scenes (TACoS), which provides coherent textual descriptions for high-quality videos, and contains 26 fine-grained cooking activities in 127 videos. In 2018, Zhou et al. [135] collected a large-scale procedure segmentation dataset with procedure segments temporally localized and described; they used cooking videos and named the dataset YouCook2. It contains 176 hours of runtime comprised of 2000 videos nearly equally distributed over 89 recipes from Africa, America, Asia, and Europe.

- **Movie:** The Large Scale Movie Description Challenge (LSMDC, Rohrbach et al., 2017) dataset [98], which provides transcribed and aligned Audio Description and script data sentences, is based on 200 movies and has 128,118 sentences with aligned clips (around 150 hours of video in total). LSMDC is based on the MPII-MD dataset and the M-VAD dataset, which were initially collected independently but are presented jointly in this work. The MPII Movie Description (MPII-MD) [97] dataset contains a parallel corpus of over 68K sentences and video snippets from 94 HD movies. The Montreal Video Annotation dataset (M-VAD) [116] includes over 84.6 hours of paired video and sentences from 92 DVDs.

4.2.5 Image and Video Captioning Evaluation Metrics:

Captions are evaluated using the BLEU, METEOR, CIDEr, and other metrics [24], [121], [131]. These metrics are common for comparing the different image, and video captioning models and have varying degrees of similarity with human judgment [110].

- **BLEU:** BiLingual Evaluation Understudy is a method of automatic machine translation evaluation that is a precision-based metric, correlates highly with human evaluation, and has a little marginal cost per run [86], [131]. BLEU has different n-grams based versions for candidate sentences concerning the reference sentences.

- **METEOR:** Metric for Evaluation of Translation with Explicit ORdering is an automatic metric that evaluates translation hypotheses. It is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations [18], [24], [32], [131].

- **CIDEr:** Consensus-based Image Description Evaluation [121] enables an objective comparison of machine generation approaches based on their human-likeness, without having to make arbitrary calls on weighing content, grammar, saliency, and many others concerning each other. CIDEr was first developed specifically for evaluating image captioning tasks, but it is also used in video captioning methods.

- **ROUGE:** Recall-Oriented Understudy for Gisting Evaluation [76] determines the quality of a summary by comparing it to other summaries created by humans. ROUGE, similar to BLEU, has different n-grams based versions.
- **SPICE:** Anderson et al. [10] introduced Semantic Propositional Image Captioning Evaluation, a novel semantic evaluation metric that measures how effectively image captions recover objects, attributes, and the relations between them. It correlates more with the human judgment of semantic quality as compared to previously reported metrics.
- **WMD:** Word Mover's Distance [72] measures the dissimilarity between two text documents. Therefore, the sensitivity of this metric when compared to BLUE, ROUGE, and CIDEr, is low about word order or synonym swapping, but, like CIDEr and METEOR, it provides a high correlation with human judgments.

4.3 Conclusion

Many models have been proposed and presented to generate captions for images and short videos in recent years. Although these models are helping to advance the technology, they suffer from inaccuracies due to fundamental constraints, resulting in limited use in practical situations. Many of the earlier models proposed treat image captioning and video captioning differently using different algorithms and methodologies. This chapter focuses on methods that perform video captioning by using image captioning methods as building blocks. Thus, the video captioning process is considered to be a compilation of the summarization of image captions. For the above reason, in this chapter, we only focused on the algorithmic overlap between image and video captioning. Therefore, this chapter is not meant to be a comprehensive review of image and video captioning; instead, a concise review of the algorithm overlaps between the two. Furthermore, this chapter only considered those algorithms that used deep learning.

In general, comparing different deep learning models used for image and video captioning is difficult because researchers use different image datasets, different parameters, different classification methods, additional preprocessing, different combinations of structures, and others. Therefore, despite the vast differences in this study, we focused on the general overlap between these methods.

A reliable, accurate, and real-time video and image captioning method can be used in many applications. Researchers attempt to give sight to the machines. First, machines learn to see. Then, they help us to see better. We will not only use the machines because of their intelligence, but we will also collaborate with them in ways that we cannot even imagine. Image and video captioning systems can be used as an essential part of Assistive Technologies that would help people with hearing or sight impairments. The captions can be used as meta-data for search engines, taking the search engine's functionality to a new dimension. Captions can be used as part of recommendation systems in many applications.

Future Research Direction and Broader Impact: As mentioned earlier, the current technologies used for image and video captioning often generate captions that are not very accurate. There is much room for improvement and enhancement. The fusion and processing of image, video and audio would

provide more accurate captions. Audio-to-Word converters are available, and they are pretty reliable. Integrating an Audio-to-Word converter with a video and combining the captions/words generated via audio and video would generate more accurate and meaningful captions even though a detailed text/sentence summarization would have to be performed.

Another challenge with video captioning is the very compute-intensive nature of the problem. With the current technology, only concise videos can be captioned (only a few seconds long). However, using the next generation of GPUs and with explicit algorithm parallelization (targeted at the GPU machine architectures), we can get closer to real-time performance for longer videos. A sophisticated opportunity in video captioning is to design and develop a strategy that would permit users to request video captions at varying levels of detail. However, we believe that the most fundamental and challenging research problem with video captioning is that different captions based on different interpretations can be generated for the same video - in the same way as two individuals can come up with two different views/description by watching the same video. We believe that this fundamental problem can be addressed by studying relevant concepts and making the process more interactive.

Chapter 5

Automatic Generation of Descriptive Titles for Video Clips Using Deep Learning

Over the last decade, Deep Learning in many applications produced results comparable to and, in some cases surpassing human expert performance. The application domains include diagnosing diseases, finance, agriculture, search engines, robot vision, and many others. This chapter proposes an architecture that utilizes image/video captioning methods and Natural Language Processing systems to generate a title and a concise abstract for a video. Such a system can potentially be utilized in many application domains, including the cinema industry, video search engines, security surveillance, video databases/warehouses, data centers, and many others The proposed system functions and operates as followed: it reads a video; representative image frames are identified and selected; the image frames are captioned; NLP is applied to all generated captions together with text summarization; and finally, a title and an abstract are generated for the video. All functions are performed automatically. Preliminary results are provided in this chapter using publicly available datasets[1].

The use of extensive neural networks as Deep Learning methods that are inspired by the human brain system has recently dominated most of the researchers' work in several domains to help in improving the results and make it more desirable for people. Machine Translation, Self-driving cars, Robotics [112], Digital Marketing, Customer Services, and Better Recommendations are some applications for deep learning. In more recent years, deep learning [9] have positively and significantly impacted the field of image recognition specifically, allowing much more flexibility. In this research, we attempt to utilize image/video captioning [80], [113] methods and Natural Language Processing systems to generate a sentence as a title for a long video that could be useful in many ways. Using an automated system instead of watching many

[1]Amirian, Soheyla, Khaled Rasheed, Thiab R. Taha, and Hamid R. Arabnia, "Automatic Generation of Descriptive Titles for Video Clips Using Deep Learning," in Springer Nature - Research Book Series: Advances in Artificial Intelligence and Applied Cognitive Computing, Transactions on Computational Science and Computational Intelligence; Springer ID: 89066307 (Book ID: 495585_1_En), ISBN #: 978-3-030-70295-3, 2020, pp. 17–28

videos to get titles could be time-saving. It can also be used in the cinema industry, search engines, and supervision cameras, to name a few. We present an example of the overall process in Figure 5.1.

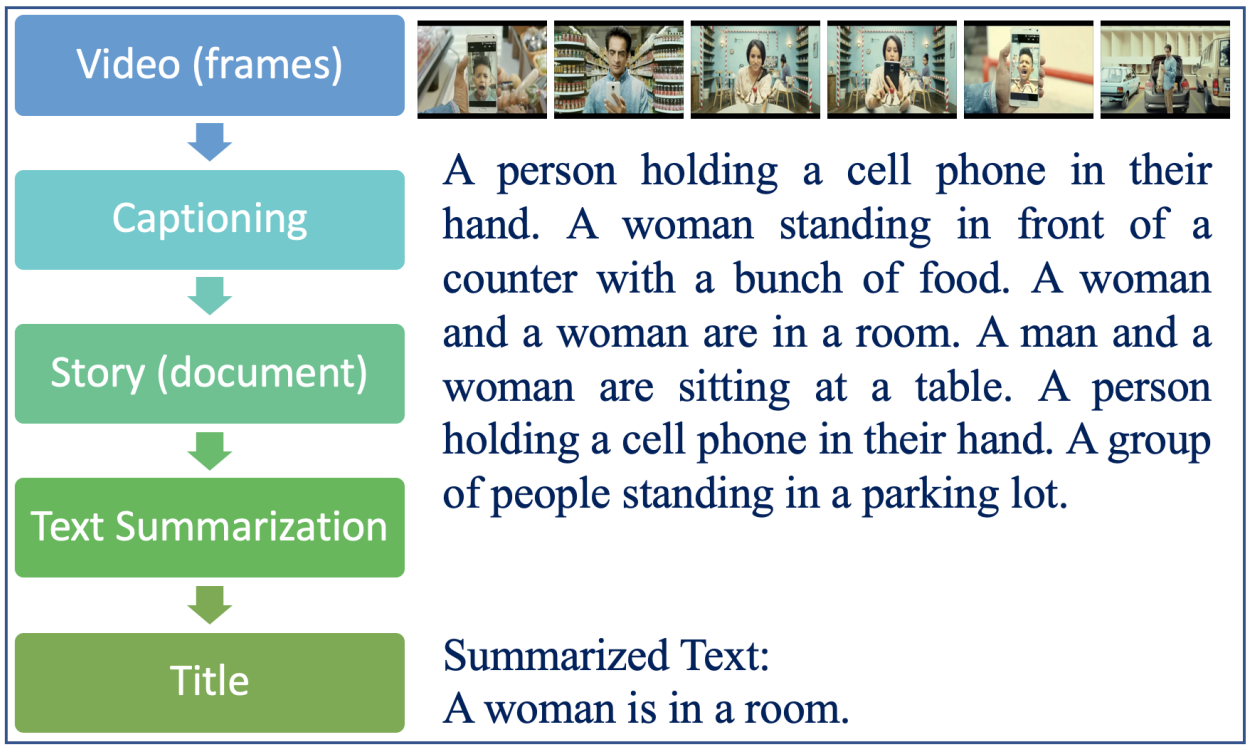


Figure 5.1: This is an overall example of the proposed system. The key-frames of the video are selected and captioned. The resulting document is processed with the text summarization method and the output is a possible title for the corresponding video.

Image and video captioning with deep learning are used for the difficult task of recognizing the objects and actions in an image or video and creating a concise meaningful sentence based on the contents found. Text summarization [3] is the task of generating a concise and fluent summary for a document(s) while preserving key information content. This chapter proposes an architecture by utilizing the image/video captioning system and text summarization methods to make a title and an abstract for a long video. For constructing a story about a video, we extract the keyframes of the video, which give more information, and then we feed those keyframes to the captioning system to make a caption or a document for them. Different methods such as encoder-decoder or generative adversarial networks exist to propose different object detection methods for the captioning system. Also, for the text summarization, we use both Extractive and abstractive methods [89] to generate the title and the abstract, respectively. We provide more details in the next sections.

The main contribution of this research is to explore the possibility of making a title and a concise abstract for a long video by utilizing deep learning technologies to save time through automation in many application domains. We described the different parts of our proposed architecture: image/video captioning and text summarization methods earlier. Here, we explain the methodology of the proposed

architecture and how it works. We present a proof of concept through experiments using publicly available datasets. The chapter is concluded with a discussion of the results and our future work.

5.1 Methodology

The proposed architecture consists of two different, complementary processes: Video Captioning and Text Summarization. In the first process for video captioning, the system gets a video as an input and then generates a video story. The generated story will feed to the second process as a document, and it summarizes the document into a sentence and an abstract. Figure 5.2 shows the complete process of the suggested architecture. Further, we explain the details of each part.

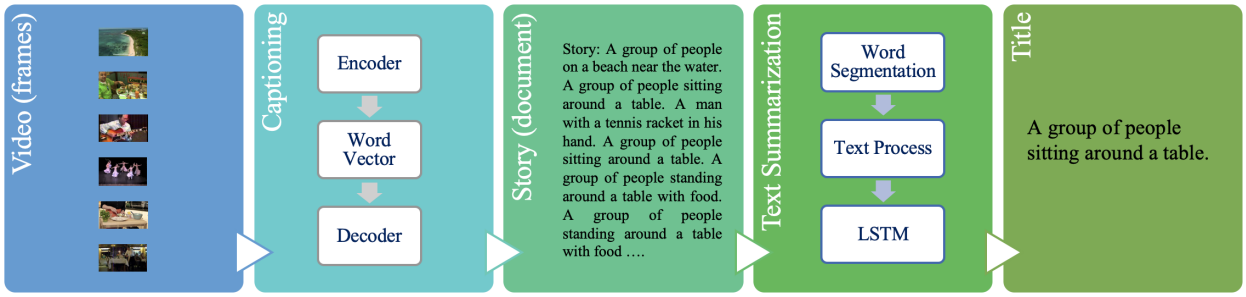


Figure 5.2: This is the overall architecture of the proposed method that parts to two separate process for the video captioning and text summarization.

5.1.1 Video to Document Process:

Image/video description is the automatic generation of meaningful sentences that describes the events in an image/video (frames). A video consists of many frames, each representing an image. Some of the images/frames give much information, and some are just basically repeating a scene. Therefore, we select some keyframes that include more information. The in-between frames are just repeating with subtle changes. A sequence of keyframes defines which movement the viewer will see. Therefore, the order of the keyframes on the video or animation determines the timing of the action.

One of our contributions in this research is doing some experiments by selecting different keyframes to have a story for long videos to see if we can have the same extracted information. So, one task is to get the keyframes and process them to be captioned instead of using all the video frames to save time and resources for getting the same result. See Figure 5.3 to illustrate the frames, keyframes, and in-between frames.

The captioning part consists of two phases: Encoder and Decoder. The Encoder part extracts the image information using convolutional neural networks like object detection methods to extract the objects and actions and then put them in a vector. ResNet, DenseNet, RCNN series, Yolo and more [9] can be used as object detection methods. Then, the vector enters the decoder phase. The Decoder

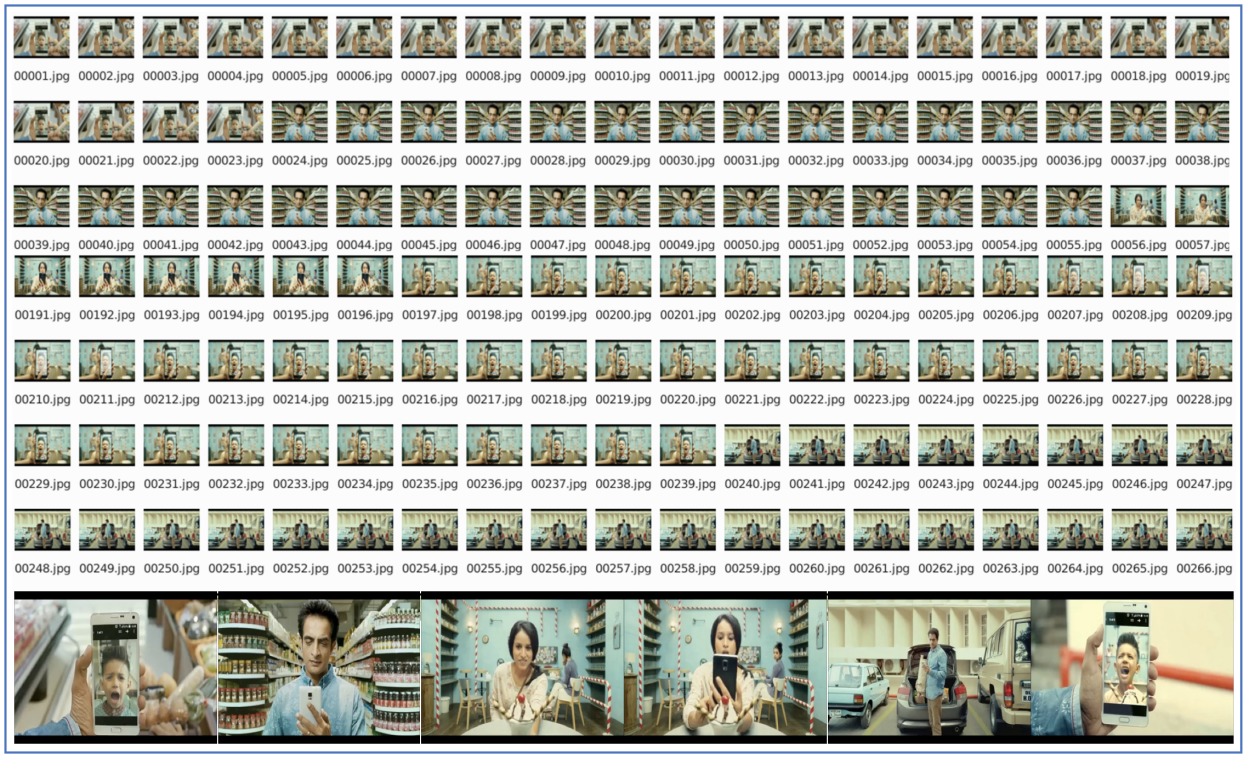


Figure 5.3: Video frames: in-between frames and the keyframes. We observe that many frames are repeating.

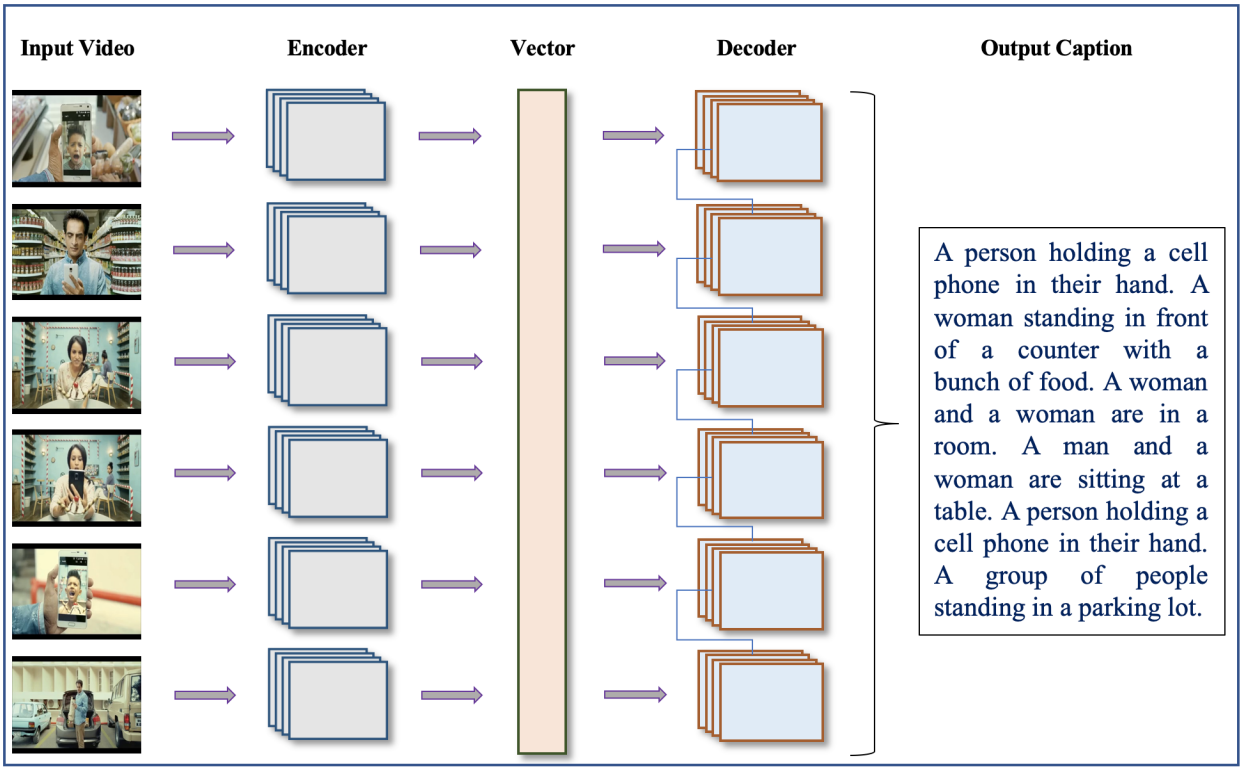


Figure 5.4: The task of video captioning can be divided logically into two modules: one module is based on an image-based model which extracts important features and nuances from video frames; another module is based on a language-based model, which translates the features and objects produced with the image-based model to meaningful sentences.

gets the vector and then, with RNN methods, generates a meaningful caption for the image. These two phases could work simultaneously. Figure 5.4 illustrates the captioning process. Captions are evaluated using the BLEU, METEOR, CIDEr, and other metrics [24], [121], [131]. These metrics are common for comparing the different image, and video captioning models and have varying degrees of similarity with human judgment [110].

5.1.2 Document to Title Process:

For generating and assigning a title to the video clip, we use an extractive text summarization technique. To keep it simple, we are using an unsupervised learning approach to find the sentence similarity and rank them [37]. The process is that we give the produced document as input. It splits the whole document into sentences. It removes stop words, builds a similarity matrix, generates rank based on the matrix, and at the end, it picks the top N sentences for a descriptive title. Figure 5.5 shows an example of the extractive text summarization system. Also, for having an abstract, we implemented and used the abstractive text summarization method for the video [50]. Abstractive summarization methods interpret and examine the text using advanced natural language techniques to generate a new shorter text that conveys the most critical information from the original text [3].

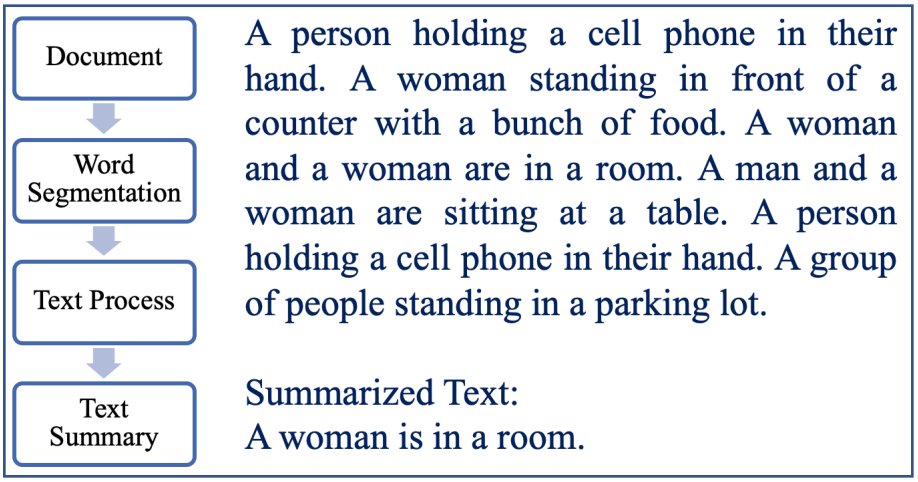


Figure 5.5: The output of the captioning system is a document(s) that is an input to the extractive text summarization method. Then the text would be processed to weight the words, and it shows the most likely descriptive title that the text could have.

5.2 Experiments

The main goal of our experiments is to evaluate the utility of the proposed architecture as a proof of concept. First, we need to get a story as a document from the image/video captioning model for implementing our idea. So, we explore the frames for a given video and then feed the selected keyframes to the system

to get the description. Implementing this part, captions have been generated by Luo et al. [80] method. The encoder has been trained with the COCO dataset. We utilized an image captioning system for this part. Some of the videos have been selected from the YouTube-8M dataset composed of almost 8 million videos totaling 500K hours of video [2], and some from the COCO dataset [77]. The captioning method has been trained and evaluated on the COCO dataset, which includes 113,287 images for training, 5,000 images for validation, and another 5,000 held out for testing. Each image is associated with five human captions.

For the image encoder in the retrieval and FC captioning model, Resnet-101 is used. For each image, the final convolutional layer output's global average pooling results in a vector of dimension 2048. The spatial features are extracted from the output of a Faster R-CNN with ResNet-101 [9], trained with the object and attribute annotations from Visual Genome [70]. Both the FC features and Spatial features are pre-extracted, and no fine-tuning is applied to image encoders. For captioning models, the dimension of LSTM hidden state, image feature embedding, and word embedding are all set to 512. The retrieval model uses GRU-RNN to encode text. The word embedding has 300 dimensions, and the GRU hidden state size and joint embedding size are 1024 [80]. The captions generated with this model describe valuable information about the frames. However, richer and more diverse sources of training signal may further improve the training of caption generators.

The Text Summarization method that has been used in the first experiment is extractive and single summarization. First, we read the generated document from the previous process. Then, we generate a Similarly Matrix across the sentences. We then rank the sentences in the similarity matrix. And at the end, we sort the rank and pick the top sentence. Figure 5.6 shows some experiments that have been done. The videos are selected from the YouTube-8M dataset [2], and some from the COCO dataset [77]. The reader can find all the guidance and code here[2] for replicating the experiments for each part of the process.

In another experiment, we implement the abstractive text summarization method [91] to generate an abstract for the video clips instead of assigning a title. Figure 5.7 shows the results by using Simple abstractive text summarization with pre-trained T5 (Text-To-Text Transfer Transformer) code [50].

5.3 Summary

The purpose of this research is to propose an architecture that could generate an appropriate title and a concise abstract for a video by utilizing image/video caption systems and text summarization methods to help in several domains such as search engines, supervision cameras, and the cinema industry. We utilized deep learning systems as captioning methods to generate a document describing a video. We then use extractive text summarization methods to assign a title and abstractive text summarization methods to create a concise abstract to the video. We explained the components of the proposed framework and conducted experiments using videos from different datasets. The results prove that the concept is valid. However, the results could become better by applying improved image/video captioning and text summarization methods.

[2] https://github.com/sohamirian/VideoTitle

A group of people on a beach near the water. A group of people sitting around a table. A man with a tennis racket in his hand. A group of people sitting around a table. A group of people standing around a table with food. A group of people standing around a table with food …

Possible title:
A group of people sitting around a table.

A church with a clock tower in the middle. A group of dogs standing next to each other. A man and a woman standing in a kitchen. A black and white photo of a person sitting on a bed. A group of people standing around a fire hydrant ...

Possible title:
A man and a woman standing in a kitchen.

A close up of a tree in a forest. A couple of people that are sitting on a bench. A man is riding on the back of a truck. A group of people riding bikes down a street ...

Possible title:
A man is riding on the back of a truck.

A man is holding a skateboard in his hand. A group of men playing a game of soccer. A woman holding a tennis racquet on a tennis court. A man and a woman sitting on a couch. A man walking down a street with an umbrella. A man is standing in front of a building.
A street scene with focus on a street. A man riding a skateboard down the side of a ramp. A group of people walking down a street. A city street with a stop sign on the corner. A living room with a couch and a window …

Possible title:
A man is holding a skateboard in his hand.

Figure 5.6: These are different videos of YouTube-8M and ActivityNet dataset that we captioned with [80] and made a document. Then, the possible title is generated with an extractive text summarization algorithm for each document.

A person holding a cell phone in their hand. A woman standing in front of a counter with a bunch of food. A woman and a woman are in a room. A man and a woman are sitting at a table. A person holding a cell phone in their hand. A group of people standing in a parking lot … Abstract: A person is holding cell phone in their hand. he stands in front of counter with food and eats food from the counter - and is sitting at table in the parking lot.
A group of people on a beach near the water. A group of people sitting around a table. A man with a tennis racket in his hand. A group of people sitting around a table. A group of people standing around a table with food. A group of people standing around a table with food … Abstract: A group of people sitting around table sits around the table with food. man with tennis racket in his hand. he is able to play tennis and swam in the water with his tv show on Thursday night.
A church with a clock tower in the middle. A group of dogs standing next to each other. A man and a woman standing in a kitchen. A black and white photo of a person sitting on a bed. A group of people standing around a fire hydrant ... Abstract: Church with a clock tower in the middle. dogs standing next to each other. man and woman standing in kitchen. black and white photo of person sitting on bed.
A close up of a tree in a forest. A couple of people that are sitting on a bench. A man is riding on the back of a truck. A group of people riding bikes down a street ... Abstract: A man is riding on the back of the truck. the group of people riding bikes down the street... and biking down street.
A man is holding a skateboard in his hand. A group of men playing a game of soccer. A woman holding a tennis racquet on a tennis court. A man and a woman sitting on a couch. A man walking down a street with an umbrella. A man is standing in front of a building. A street scene with focus on a street. A man riding a skateboard down the side of a ramp. A group of people walking down a street. A city street with a stop sign on the corner. A living room with a couch and a window … Abstract: A man is holding skateboard in his hand. he is standing in front of an umbrella and walking down the side of the ramp.

Figure 5.7: Here are the results with the abstractive text summarization method. Which generates a summary for each document.

We plan to explore more recent image/video captioning systems to generate a more natural story to describe the video clips in our future work. Therefore, the text summarization system could generate a better title using the extractive text summarization algorithms and a better abstract using the abstractive text summarization algorithms.

Chapter 6

A Novel Application: The Use of Video Captioning for Fostering Physical Activity

Video Captioning is considered to be one of the most challenging problems in the field of computer vision. Video Captioning involves combining different deep learning models to perform object detection, action detection, and localization by processing a sequence of image frames. Therefore, it is crucial to consider the sequence of actions in a video to generate a meaningful description of the overall action event. A reliable, accurate, and real-time video captioning method can be used in many applications. However, this study focuses on one application: video captioning for fostering and facilitating physical activities. Thus, in broad terms, the work can be considered to be assistive technology. Lack of physical activity appears to be increasingly widespread in many nations due to many factors, the most important being the convenience that technology has provided in workplaces. In addition, the adopted sedentary lifestyle is becoming a significant public health issue. Therefore, it is essential to incorporate more physical movements into our daily lives. Tracking one's daily physical activities would offer a base for comparison with actions performed in subsequent days. With the above in mind, this study proposes a video captioning framework that aims to describe the activities in a video and estimate a person's daily physical activity level. This framework could potentially help people trace their daily movements to reduce an inactive lifestyle's health risks. The work presented in this study is still in its infancy. The initial steps of the application are outlined in this study. Based on our preliminary research, this project has excellent merit[1].

With the recent availability of powerful machines (GPUs, CPU clusters), together with large amounts of training data, deep learning has made a comeback providing breakthroughs on image recognition and object detection [8], [9], [53], [107], [108], [118] with many applications. Object detection models [53], [93] are used to extract object information and localization from images and videos. The prosperity of

[1]1. Amirian, Soheyla, Abolfazl Farahani, Khaled Rasheed, Hamid R. Arabnia, and Thiab R. Taha, "The Use of Video Captioning for Fostering Physical Activity," in Computational Science and Computational Intelligence; 2020 International Conference on IEEE CPS (IEEE XPLORE, Scopus), ISBN-13: 978-1-7281-7624-6, pp. 611-614, 2020.

object detection models has made them suitable for other deep learning tasks, including image and video captioning, automation of making a title for videos [7], [8] and more.

Image captioning is a comparatively more trivial task than video captioning, as we have to deal with more objects in a video. Video captioning requires understanding the video contents accurately to detect the objects, their corresponding actions, and relations. In this task, detecting the objects' correlation plays an important role in generating a meaningful and consistent description. Besides, events may overlap as the length of events varies across videos, making object detection a difficult task. Some existing models address these concerns [1], [8], [69], [83]. A reliable, accurate, and real-time video captioning method can be used in many applications. Video captioning techniques are utilized in medical and healthcare applications, a guide to interacting with people with visual impairments, human-robot interaction, automatic video subtitling, video surveillance, automatic title/summary generation [7], self-driving vehicles, sign language translation, and many others [8]. Technology is moving faster than ever, and we could soon interact with robots in the same manner as we do with humans. Taking advantage of human-robot interaction, medical and healthcare applications, we introduce a novel video captioning framework that offers an activity summary. Using video captioning for fostering and facilitating physical activities, the proposed model can be used as assistive technology.

Lack of physical activity is increasingly widespread in many nations due to many reasons. The adopted sedentary lifestyle is becoming a significant public health issue. Therefore, it is essential to incorporate more physical movements into our daily lives. Tracking one's daily physical activities would offer a base for comparison with activities performed in subsequent days. With the above in mind, this study proposes a video captioning framework that aims to describe the activities in a video and estimate a person's daily physical activity level. This framework could potentially help people trace their daily movements to reduce the health risks of an inactive lifestyle. First, we feed a video of a person's daily activity into the model. Then, the model processes the video by extracting actions and generating captions. Finally, it summarizes an activity history. Figure 6.1 illustrate more details. This framework could potentially help people trace their daily movements to reduce the health risks of an inactive lifestyle by managing their activities [40].

The main contribution of this research is proposing a novel video captioning framework. This framework utilizes the Spatio-Temporal information in a video to generate accurate and coherent captions for filmed physical activities. The captions comprise temporal dynamics of discovered actions that could be used to follow a person's physical activity in daily life. The work presented in this study is still in its infancy. However, the initial steps of the application are outlined in this study. Based on our preliminary research, this project could be a healthcare application used by physicians or the public.

6.1 Examination of the Documents

6.1.1 Video Captioning

Video captioning is an automatic process that aims to generate natural language sentences describing a given video's contents. Thus, video captioning can be considered a system describing human activities

with natural language [8]. Recent video captioning techniques often consist of two main parts; an encoder and a decoder. In Dense-captioning proposed by Krishna et al. [69], the encoder part focuses on detecting multiple events recorded in a video by jointly localizing temporal proposals of interest, and the decoder describes the events in natural language sentences. This model introduces a new captioning module that applies the past and future contextual information to describe all the events jointly. Mun et al. [83] extend the Dense-captioning model by proposing Streamlined Dense Video Captioning. This technique utilizes multiple steps to generate coherent, consistent, and unique video descriptions by incorporating temporal dependencies across events. To achieve this, the model first introduces the Event Proposal Network (EPN) that adopts Single-Stream Temporal action proposals (SST)[19]. EPN tends to find a series of candidate event proposals that are semantically and temporally meaningful. In the next step, the Event Sequence Generation Network (ESGN) creates an episode for a video by selecting a highly correlated set of event proposals from the candidate proposals. Finally, SCN generates coherent captions for the selected event proposals using a sequential captioning network (SCN). SCN consists of an episode RNN and an event RNN. The episode RNN takes the selected proposals one at a time to model an episode's state. The event RNN generates a caption for each event proposal in a sequential manner, where the words in a caption are conditioned on the implicit representation of the episode.

6.1.2 Temporal Activity Proposals

Action recognition is an essential part of video captioning. This task generally provides fundamental tools and applications for action detection. For example, Temporal Action Detection is a tool that focuses on localizing the temporal extent of each action for a detected object [20]. Heilbron et al. [20] proposed Fast temporal activity proposals to detect human actions in untrimmed videos efficiently. This method is an end-to-end action detection pipeline that generates high-quality proposals in terms of localization and ranking. However, the proposed method's efficiency needs to be improved by interleaving or combining the feature extraction and proposal representation step. SST [19] can run continuously in a single stream over long input video sequences to find semantically meaningful temporal regions via a single scan of videos. SST produces high temporal overlapping proposals with grand truth action intervals by considering and evaluating many action proposals over densely sampled time scales and locations. In this technique, *Input* is a video consist of N frames, and *Visual Encoding* applies 3D Convolutional (C3D) network [119] to compute feature representation. *Sequence Encoding* progressively accumulates the evidence in a video sequence and simultaneously disregards the irrelevant background. This process continues until it is confident that an action is taking place in the video.

Finally, *Output* produces confidence scores of multiple proposals at each time step. 3D-CNNs are powerful tools to learn Spatio-temporal features by encapsulating information related to objects, scenes, and actions. They are popular networks, and many recent video captioning frameworks [1], [83], [124] benefit from them. Yang et al. [132] proposed the first end-to-end progressive optimization framework for video action detection known as Spatio-Temporal Progressive Learning (STEP). STEP recognizes the action of interset recorded in a video by localizing them in both space and time. Unlike previous methods that directly perform action detection in one run, STEP involves a multi-step optimization

process that progressively refines the initial proposals towards the final solution with spatial refinement and temporal extension. Multiple steps run in a sequential order, where the outputs of one stage are used as the proposals for the next step. Spatial refinement aims to improve the classification and localization of the action's regions. It starts with a small number of coarse-scale proposals and updates them iteratively. Temporal extension, on the other hand, focuses on improving classification accuracy by incorporating longer-range temporal information. Moreover, STEP achieves superior performance by using only a handful of proposals and preventing the need to generate and process large proposals.

6.1.3 Object Detection

Object detection is a computer vision task that collects the related functions to identify objects in an image or video frame by localizing and classifying them accurately. Deep learning models [9] have helped object detection by extracting high-quality representations that result in identifying more objects with precise localization. Many video captioning techniques benefit from object detection models to improve captioning performance. Redmon et al. [93] proposed YOLO (You Only Look Once), a real-time object detection system that can detect over 9000 object categories. Aafaq et al. [1] utilized YOLO in their video captioning framework. The proposed model processes the objects' locations and the corresponding multiplicity information extracted from YOLO to encode the scenes' spatial dynamics.

6.1.4 Decoder

Hierarchical recurrent neural networks are employed to generate coherent captions based on the detected events. Mun et al. [83] proposed framework includes a captioning network termed as Sequential Captioning Network (SCN) that generates the descriptions for the event proposals provided by other parts of the framework. SCN consists of two parts; episode RNN and event RNN. Episode RNN adopts a single-layer Long Short Term Memory (LSTM) with a 512-dimensional hidden state while event RNN utilizes captioning network with temporal dynamic attention and context gating (TDA-CG) [124]. Reinforcement Learning is applied to learn RNNs in SCN using event and episode-level rewards. Event-level reward allows the system to accurately capture specific content in each event, while the episode-level reward enforces the network to produce coherent descriptions from all generated captions. Dense-Captioning [69] uses a captioning LSTM network to describe the video. It develops an analogous model that groups the events to capture the temporal context. The captioning module of the framework then incorporates the context from neighboring events to capture the correlations between events. Instead of using a complex language model, Aafaq et al. [1] utilized multiple layers of Gated Recurrent Units (GRUs) [25] in the captioning part of their model. Gated Recurrent Units are simple sequential models and known to be robust to vanishing gradient problems. Furthermore, GRUs generate semantically rich captions from the visual representation discovered by the earlier part of the model.

6.2 Proposed Framework

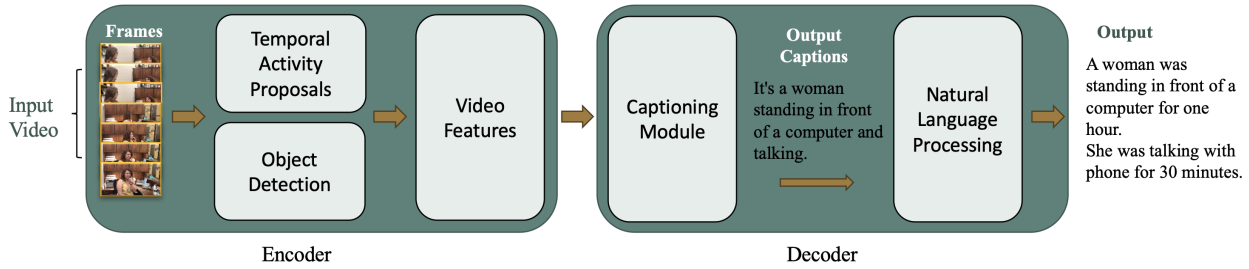


Figure 6.1: This is an overview of the proposed framework. It consists of an encoder and a decoder. The encoder part focuses on detecting objects, multiple events, and actions recorded in a video by jointly localizing temporal proposals of interest. The decoder first generates captions for each event proposal and then finds the correlation between them to summarize an activity history. The produced story describes the type of each physical activity and the corresponding duration. We can use this information to categorize the video into different physical activity levels.

In this section, we elaborate on the proposed idea for a video captioning framework. The model intends to generate a meaningful, coherent, and accurate video description for the overall action events by incorporating the Spatio-Temporal information of each event recorded in the video. The proposed technique involves two main components; an Encoder and a Decoder. The encoder utilizes multiple deep learning models to discover high-quality representations from the input video frames. These representations are then embedded into a high-dimensional feature space to create an input for the decoder. The decoder comprises a captioning module and a Natural Language Processing (NLP) module. The captioning module generates a description for each input video frame. In contrast, the NLP module aims to create a coherent story for all input videos by considering the generated captions and finding the correlation between them. In the following, we describe our framework in detail.

6.2.1 Encoder

In the encoding section, we propose a visual encoding technique for video action detection. This technique aims to compute representations enriched by utilizing Spatio-TEmporal Progressive (STEP) action detector [132] to recognize the actions of interest presented in a video while considering their spatial and temporal properties. STEP consists of spatial refinement and temporal extension, where the former tends to update the accurate localizations for the proposals iteratively, and the latter gradually incorporate relevant temporal context by increasing sequence length. STEP starts with a small number of initial proposals and updates them to classify better and localize action regions in the spatial refinement. In this section, multiple steps are carried out in a sequential order, where the outputs of one stage are utilized as the proposals for the next step. Temporal extension, on the other hand, improves classification accuracy by including longer-range temporal information. Generally, we aim to embed object labels, their frequencies of occurrence, and the evolution of their spatial locations in the encoding process. In this pipeline, the

temporal information is kept to be further utilized in the framework's final output. The decoding part of the framework uses this information to keep track of each action's duration. Furthermore, we encode spatial dynamics by processing objects' locations and their multiplicity information extracted from an Object Detector (i.e., YOLO [93], [94]). The output layers in both Object Detector and STEP are then fused to extract the object's semantics and actions of the video.

6.2.2 Decoder

The decoding section processes the representations extracted by the encoder to generate the captions describing each action event recorded in the video. Captioning can be obtained by naively treating each action individually. However, events are usually highly correlated in a video, and ignoring the correlations between actions could generate inconsistent or redundant descriptions. To address this challenge, the model needs to include the temporal dependency across events. Hence, our framework's decoder adopts the sequence modeling components of [1], which employs Gated Recurrent Units (GRU). We specifically employ two layers of GRUs in our model to use the temporal information produced by the encoder. Moreover, we utilize Natural Language Processing system [33] to find the relevant actions and calculate the duration for each action by analyzing the descriptions generated by GRUs. Thus, it helps the model to automate the process of generating the final report. The overall framework of the proposed method is illustrated in Figure 6.1.

6.3 Discussion and Future Work

Deep learning and video captioning frameworks can be used as technologies to help people to control their health. The captions can be used as meta-data for search engines, taking the search engine's functionality to a new dimension. Captions can also be used as part of recommendation systems in many applications. Video captioning is a very challenging task as it is naturally very compute-intensive. Moreover, the currently proposed captioning models can only deal with short videos that are a few seconds long. We could achieve real-time performance for longer videos with the next generation of GPUs and explicit parallel algorithms targeted at the GPU machine architectures. A sophisticated opportunity in video captioning is to design and develop a strategy that would permit users to request video captions with various levels of detail. In this research, we investigated different frameworks of video captioning and action detection models. First, we proposed a framework that takes a video from users containing their daily activity. Then, it analyzes and captions the video to inform the user about their physical activity in everyday life. We are currently implementing this framework that could potentially be used as a health recommendation application. Moreover, we are not concerned about the efficiency of the framework at the execution time. However, we hope to be able to address execution efficiency issues in our subsequent publications.

Chapter 7

Conclusion

We have observed active and fast advances in the field of Computer Vision (deep learning, image, and video captioning) over the last few years. Progress in this area has unlocked a wide variety of ambitious problems that once defied our efforts. In particular, in this dissertation, we developed different frameworks and techniques that push the frontier of video captioning. By selecting the keyframes in a video and proposing frameworks for various applications, we step towards Artificial Intelligence agents that can perceive the visual world and interact with us naturally. We argued that by selecting keyframes from videos, we could assign captions to the long videos and use them in different applications, accelerating the video captioning process. We evaluated this captioning qualitatively in this work; quantitative assessment could be the subject of future studies.

Concretely, in **Chapter 2**, we dissected relevant deep learning structures and models. Then we had two case studies; Integrated Plant Growth and Disease Monitoring with IoT and Deep Learning Technology. Hereabouts we discussed what the proposed system hopes to achieve and implications for future research, possibly setting the stage for a follow-up study in collaboration with an agricultural subject matter expert. In the second case study, we investigated the Stereotype-Free Classification of Fictitious Faces using Generative Adversarial Networks. And, we presented a novel approach through penalized regression to label stereotype-free GAN-generated synthetic unlabeled images. **Chapter 3** and **Chapter 4** were a comprehensive review of image captioning and video captioning methodologies based on deep learning. This study treated both image and video captioning by emphasizing the algorithmic overlap between the two. In **Chapter 5**, we developed a model that can select and extract the keyframes in a video. First, we utilized deep learning systems as captioning methods to generate a document describing a video. Then, we used extractive text summarization methods. This step assigned a concise abstract to the video and then created a title by text summarization methods. Finally, we explained the components of the proposed framework and conducted experiments using videos from different datasets. The results proved that the concept is valid. However, the results could become better by applying improved image/video captioning and text summarization methods.

Finally, in **Chapter 6**, we proposed another application by image/video captioning. We proposed a framework that takes a video from users containing their daily activity. Then, it analyzed and captioned

the video to inform the user about their physical activity in everyday life. This study proposed a video captioning framework that describes the activities in a video and estimates a person's daily physical activity level. This framework could potentially help people trace their daily movements to reduce an inactive lifestyle's health risks. The work presented in this study is still in its infancy.

Despite recent rapid progress in image/video captioning, it is clear that many challenges remain in the vision of machines that can sense the visual world and interact with us naturally to utilize them in many applications. We can employ the captioning process in search engines, the cinema industry, and computer vision tasks. Notably, we pick some of the keyframes in the video. If the keyframes do not give us a good caption, we can add more. There is a trade-off between execution time for video captioning and the accuracy of the generated captions. In other words, our goal is to reduce the computation process and speed up the captioning task for videos. Furthermore, even if the generated caption for the video is not accurate, we can enhance it further by using crowdsourcing. We hope that our ideas can be reused and help inspire future work.

The remaining challenges are best illustrated with an example. Consider the image in Figure 7.1. As you can see, we need more frames to extract the pepper, and by having only a frame, we may not have an exact caption. However, we proved that the keyframes are a good method to caption for long videos. In addition, there is a trade-off between execution time for video captioning and the accuracy of the captions. Even if the video caption is not accurate, we can enhance it further by using crowdsourcing. Also, even if the keyframe does not give us a good result, we can add more keyframes like the example we had here. Experiments demonstrated that our method using keyframes could achieve competitive results for video captioning.

Figure 7.1: Video: a-LSTMs: a woman is slicing a tomato into pieces [44]; Only one frame: a person standing at a counter with a ball. [80]
GT:a woman is slicing a red pepper.

Bibliography

[1] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 487–12 496.

[2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[3] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text summarization techniques: A brief survey," *arXiv preprint arXiv:1707.02268*, 2017.

[4] S. Amirian, A. Farahani, K. Rasheed, T. R. Taha, and H. R. Arabnia, "The use of video captioning for fostering physical activity," in *2020 International Conference on IEEE CPS (IEEE XPLORE, Scopus), ISBN-13: 978-1-7281-7624-6*, 2020, pp. 611–614.

[5] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, "Image captioning with generative adversarial network," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, 2019, pp. 272–275.

[6] ——, "A short review on image caption generation with deep learning," in *The 23rd International Conference on Image Processing, Computer Vision and Pattern Recognition (IPCV'19), World Congress in Computer Science, Computer Engineering and Applied Computing (CSCE19), ISBN: 1-60132-506-1*, IEEE, 2019, pp. 10–18.

[7] ——, "Automatic generation of descriptive titles for video clips using deep learning," in *Springer Nature - Research Book Series: Advances in Artificial Intelligence and Applied Cognitive Computing, Transactions on Computational Science and Computational Intelligence; Springer ID: 89066307 (Book ID: 495585_1_En), ISBN #: 978-3-030-70295-3*, 2020, pp. 17–28.

[8] ——, "Automatic image and video caption generation with deep learning: A concise review and algorithmic overlap," *IEEE Access*, vol. 8, pp. 218 386–218 400, 2020. DOI: 10.1109/ACCESS.2020.3042484.

[9] S. Amirian, Z. Wang, T. R. Taha, and H. R. Arabnia, "Dissection of deep learning with applications in image recognition," in *Computational Science and Computational Intelligence; "Artificial Intelligence" (CSCI-ISAI); 2018 International Conference on. IEEE*, 2018, pp. 1132–1138.

[10] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*, Springer, 2016, pp. 382–398.

[11] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[12] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5561–5570.

[13] H. R. Arabnia, W.-C. Fang, C. Lee, and Y. Zhang, "Context-aware middleware and intelligent agents for smart environments," *IEEE Intelligent Systems*, vol. 25, no. 2, pp. 10–11, 2010.

[14] H. R. Arabnia and M. A. Oliver, "Fast operations on raster images with simd machine architectures," in *Computer Graphics Forum*, Wiley Online Library, vol. 5, 1986, pp. 179–188.

[15] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[16] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[18] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[19] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles, "Sst: Single-stream temporal action proposals," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2911–2920.

[20] F. Caba Heilbron, J. Carlos Niebles, and B. Ghanem, "Fast temporal activity proposals for efficient detection of human actions in untrimmed videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1914–1923.

[21] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, 2011, pp. 190–200.

[22] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[23] Q. Chen, X. Meng, W. Li, X. Fu, X. Deng, and J. Wang, "A multi-scale fusion convolutional neural network for face detection," in *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*, IEEE, 2017, pp. 1013–1018.

[24] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[25] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[27] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[28] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 578–10 587.

[29] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2970–2979.

[30] B. Dai and D. Lin, "Contrastive learning for image captioning," in *Advances in Neural Information Processing Systems*, 2017, pp. 898–907.

[31] L. Deligiannidis and H. R. Arabnia, "Parallel video processing techniques for surveillance applications," in *Computational Science and Computational Intelligence (CSCI), 2014 International Conference on*, IEEE, vol. 1, 2014, pp. 183–189.

[32] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 376–380.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: `1810.04805 [cs.CL]`.

[34] S. Ding, S. Qu, Y. Xi, and S. Wan, "A long video caption generation algorithm for big video data retrieval," *Future Generation Computer Systems*, vol. 93, pp. 583–595, 2019.

[35] ——, "Stimulus-driven and concept-driven analysis for image caption generation," *Neurocomputing*, vol. 398, pp. 520–530, 2020.

[36] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[37] P. Dubey, *Text Summarization*, https://github.com/edubey/text-summarizer, accessed 2020-04-04.

[38] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, ACM, 2012.

[39] S. Ehandarkar and H. R. Arabnia, "Parallel computer vision on a reconfigurable multiprocessor network," *IEEE Transactions on Parallel and Distributed Systems*, vol. 8, no. 3, pp. 292–309, 1997.

[40] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *arXiv preprint arXiv:2010.03978*, 2020.

[41] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *European conference on computer vision*, Springer, 2010, pp. 15–29.

[42] J. Fowler and S. Amirian, "Integrated plant growth and disease monitoring with iot and deep learning technology," in *Springer Nature - Research Book Series: Transactions on Computational Science & Computational Intelligence; Series Title: Advances in Data Science & Information Engineering, Springer ID: 89066304 (Book ID: 495582_1_En)*, 2020, pp. 389–395.

[43] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2321–2334, 2017.

[44] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[45] S. Gella, M. Lewis, and M. Rohrbach, "A dataset for telling the stories of social media videos," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 968–974.

[46] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[47] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[48] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, 2. MIT press Cambridge, 2016, vol. 1.

[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[50] R. Goutham, *Simple abstractive text summarization with pretrained T5- Text-To-Text Transfer Transformer*, https://towardsdatascience.com/simple-abstractive-text-summarization-with-pretrained-t5-text-to-text-transfer-transformer-10f6d602c426, accessed 2020-06-06.

[51] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller, "The case for process fairness in learning: Feature selection for fair decision making," in *NIPS Symposium on Machine Learning and the Law*, 2016.

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, IEEE, 2017, pp. 2980–2988.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[54] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-vae: Learning basic visual concepts with a constrained variational framework.," *ICLR*, vol. 2, no. 5, p. 6, 2017.

[55] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[56] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks.," in *CVPR*, vol. 1, 2017, p. 3.

[57] InfluxData, *Telegraf 1.13 documentation*, accessed 2019-12-30. [Online]. Available: `https://docs.influxdata.com/telegraf/v1.13/`.

[58] T. Iqbal and H. Ali, "Generative adversarial network for medical images (mi-gan)," *Journal of medical systems*, vol. 42, no. 11, p. 231, 2018.

[59] R. Jafri, S. A. Ali, and H. R. Arabnia, "Computer vision-based object recognition for the visually impaired using visual tags," in *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2013, p. 1.

[60] R. Jafri and H. R. Arabnia, "Fusion of face and gait for automatic human recognition," in *Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on*, IEEE, 2008, pp. 167–173.

[61] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, ACM, 2014, pp. 675–678.

[62] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth, "Fairness in learning: Classic and contextual bandits," in *Advances in Neural Information Processing Systems*, 2016.

[63] N. Jouppi, C. Young, N. Patil, and D. Patterson, "Motivation for and evaluation of the first tensor processing unit," *IEEE Micro*, vol. 38, no. 3, pp. 10–19, 2018.

[64] D. Justus, J. Brennan, S. Bonner, and A. S. McGough, "Predicting the computational cost of deep learning models," in *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, pp. 3873–3882.

[65] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in neural information processing systems*, 2014, pp. 1889–1897.

[66] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv: 1412.6980*, 2014.

[68] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[69] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 706–715.

[70] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`.

[72] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International Conference on Machine Learning*, 2015, pp. 957–966.

[73] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[74] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, 2011, pp. 220–228.

[75] T. Lillesand, R. W. Kiefer, and J. Chipman, *Remote sensing and image interpretation*. John Wiley & Sons, 2014.

[76] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[77] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[78] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks.," in *ICML*, 2016, pp. 507–516.

[79] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[80] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich, "Discriminability objective for training descriptive captions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6964–6974.

[81] D. Luper, D. Cameron, J. Miller, and H. R. Arabnia, "Spatial and temporal target association through semantic analysis and gps data mining.," in *IKE*, vol. 7, 2007, pp. 25–28.

[82] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[83] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, "Streamlined dense video captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6588–6597.

[84] O. M. Nezami, M. Dras, S. Wan, C. Paris, and L. Hamey, "Towards generating stylized image captions via adversarial training," in *Pacific Rim International Conference on Artificial Intelligence*, Springer, 2019, pp. 270–284.

[85] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 971–10 980.

[86] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 2002, pp. 311–318.

[87] E. Parcham, N. Mandami, A. N. Washington, and H. R. Arabnia, "Facial expression recognition based on fuzzy networks," in *Computational Science and Computational Intelligence (CSCI), 2016 International Conference on*, IEEE, 2016, pp. 829–835.

[88] Particle.io, *Particle Photon*, accessed 2019-12-30. [Online]. Available: `https://docs.particle.io/photon/`.

[89] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," *arXiv preprint arXiv:1705.04304*, 2017.

[90] M. Pérez-Ortiz, S. Jiménez-Fernández, P. A. Gutiérrez, E. Alexandre, C. Hervás-Martınez, and S. Salcedo-Sanz, "A review of classification problems and algorithms in renewable energy applications," *Energies*, vol. 9, no. 8, p. 607, 2016.

[91] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[92] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[93] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[94] ——, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[95] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 25–36, 2013.

[96] M. Research, *Microsoft research*, accessed 2019-04-04. [Online]. Available: `https://caption.ai`.

[97] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3202–3212.

[98] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 94–120, 2017.

[99] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 1194–1201.

[100] B. Romera-Paredes and P. H. S. Torr, "Recurrent instance segmentation," in *European Conference on Computer Vision*, Springer, 2016, pp. 312–329.

[101] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[102] R. K. Roul, S. Mehrotra, Y. Pungaliya, and J. K. Sahoo, "A new automatic multi-document text summarization using topic modeling," in *International conference on distributed computing and internet technology*, Springer, 2019, pp. 212–221.

[103] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," *Computer vision and image understanding*, 2001.

[104] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth moverś distance as a metric for image retrieval," *International journal of computer vision*, 2000.

[105] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[106] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv preprint arXiv:1605.06211*, 2016.

[107] F. Shenavarmasouleh and H. R. Arabnia, *Drdr: Automatic masking of exudates and microaneurysms caused by diabetic retinopathy using mask r-cnn and transfer learning*, 2020. arXiv: `2007.02026 [cs.CV]`.

[108] F. Shenavarmasouleh, F. G. Mohammadi, M. H. Amini, and H. R. Arabnia, *Drdr ii: Detecting the severity level of diabetic retinopathy using mask rcnn and transfer learning*, 2020. arXiv: `2011.14733 [eess.IV]`.

[109] S. Shi, Q. Wang, P. Xu, and X. Chu, "Benchmarking state-of-the-art deep learning software tools," in *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, IEEE, 2016, pp. 99–104.

[110] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*, Springer, 2016, pp. 510–526.

[111] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[112] N. Soans, E. Asali, Y. Hong, and P. Doshi, "Sa-net: Robust state-action recognition for learning from observations," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 2153–2159.

[113] J. Sung Park, M. Rohrbach, T. Darrell, and A. Rohrbach, "Adversarial inference for multi-sentence video description," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6598–6608.

[114] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[115] A. P. Tafti, A. Baghaie, M. Assefi, H. R. Arabnia, Z. Yu, and P. Peissig, "Ocr as a service: An experimental evaluation of google docs ocr, tesseract, abbyy finereader, and transym," in *International Symposium on Visual Computing*, Springer, 2016, pp. 735–746.

[116] A. Torabi, C. Pal, H. Larochelle, and A. Courville, "Using descriptive video services to create a large data source for video annotation research," *arXiv preprint arXiv:1503.01070*, 2015.

[117] M. Toutiaee, S. Amirian, J. A. Miller, and S. Li, "Stereotype-free classification of fictitious faces," *arXiv preprint arXiv:2005.02157*, 2020.

[118] M. Toutiaee, A. Keshavarzi, A. Farahani, and J. A. Miller, "Video contents understanding using deep neural networks," *arXiv preprint arXiv:2004.13959*, 2020.

[119] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[120] H. Valafar, H. R. Arabnia, and G. Williams, "Distributed global optimization and its development on the multiring network," *Neural, Parallel & Scientific Computations*, vol. 12, no. 4, pp. 465–490, 2004.

[121] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[122] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko, "Improving lstm-based video description with linguistic knowledge mined from text," *arXiv preprint arXiv:1604.01729*, 2016.

[123] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7622–7631.

[124] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional attentive fusion with context gating for dense video captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7190–7198.

[125] L. Wang, A. Schwing, and S. Lazebnik, "Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space," in *Advances in Neural Information Processing Systems*, 2017, pp. 5756–5766.

[126] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1711–1720.

[127] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Yang Wang, "Video captioning via hierarchical reinforcement learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4213–4222.

[128] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 203–212.

[129] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Googleś neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[130] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5288–5296.

[131] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[132] X. Yang, X. Yang, M.-Y. Liu, F. Xiao, L. S. Davis, and J. Kautz, "Step: Spatio-temporal progressive learning for video action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 264–272.

[133] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.

[134] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[135] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.