Master Thesis

Multi-modal Video Understanding with Eye Gaze and Spoken Language

(視線と発話に基づく映像のマルチモーダル解析)

Medhe Mayanka Anil

48-206477

Department of Information and Communication Engineering

Graduate School of Information Science and Technology

the University of Tokyo

Thesis Supervisor: Yoichi Sato

August, 2022

Abstract

Humans use multi-modal cognitive information to perceive and interact with their surroundings, such as vision, hearing, touch, taste, and smell. Of all these senses, the most important cognitive sense to understand our surroundings is vision. This understanding is best conveyed in a deeper and well-organized manner by putting them into words. At the same time, we know that our brain cannot process all the visual input it receives but instead focuses on regions it finds interesting. This filtering of information is done through the medium of eye gaze. Therefore, we have used such visual and speech information to help machines understand the environment in a more holistic manner. In this thesis, we propose a new multi-modal dataset that tells us where and what are the regions of interest in a video. The modalities leveraged for this purpose are eye gaze, spoken language, and video. Many applications can be tackled using this dataset, such as spatially and temporally dense video captioning, grounded speech recognition, visual grounding, navigation, and video retrieval. The unique aspect of our dataset is the use of eye gaze information. The motivation behind its use is to understand human's underlying attention and observe the difference in eye fixations over time. The application that can be most benefited from this attention region filtering is dense video captioning as it helps in narrowing down important regions based on human attention. Therefore, we chose to tackle this task in our research. We demonstrate the effectiveness of our multi-modal information in improving the current video captioning tasks through various experimental settings.

Acknowledgments

I would like to express my gratitude to all the people who guided me throughout my research. I would like to give a special thanks to my respected advisor, Prof. Yoichi Sato for his continuous guidance and support throughout my master's program. His feedback has always been insightful and helped me become a better researcher. Besides my advisor, I would like to thank my senior, Yifei Huang. His ability to make difficult tasks seem manageable has helped me a lot to overcome the obstacles during my research. I would also like to thank our research associate Ryosuke Furuta for his constructive suggestions that helped refine my research further. I also extend my gratitude to our project researcher Yusuke Goutsu, whose advice has helped me a lot. I would like to thank each and every member of Sato/Sugano laboratory for their welcoming atmosphere and motivating spirit. I am forever indebted to my lab for all the knowledge I gained which has made me into the person I am today. Finally, I would like to thank my parents, Surekha Medhe and Anil Medhe, and my sister, Mitali Medhe for their constant motivation and emotional support throughout my degree.

Contents

Figure List

Table List

Chapter 1

Introduction

1.1 Background

Modal information refers to the information in the form of what we see, hear, smell, and feel. At any given point, we utilize one or more of this modal information to make decisions. This combination of different modal information is called multi-modal data. Our brain learns to give the appropriate weight to such multi-modal information over its course of development to better understand our surroundings. For example, when crossing the road, we rely more on our sense of seeing and hearing than other senses.

If we want to make computers understand its surroundings and make decisions like us we have to provide multi-modal information to better represent human-like cognitive state. Humans have five basic senses, namely, vision, hearing, touch, taste, and smell. Vision and hearing play a major role in the perception of our environment. The sounds of the surroundings are especially important when vision is obscure. Even within vision, there are only certain regions of interest that are important at any given time. These regions are innately identified by humans through their eye gaze. Using the gaze of the eye, we eliminate unnecessary background information and focus on the important activity that is happening. The importance of an activity is directly correlated with how long someone stared at it. This prolonged staring is called fixations, and it also gives us information about the most interesting activity for that person. Our brain processes this audio-visual

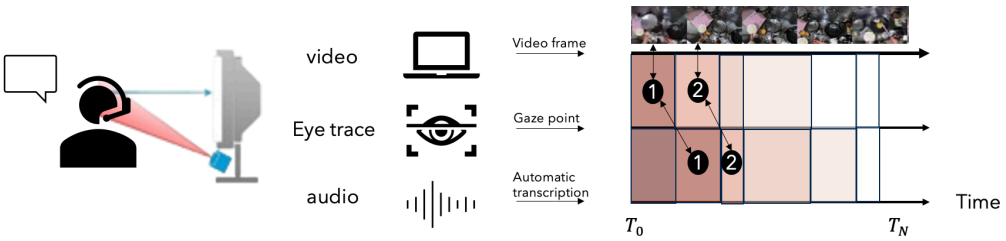

Fig. 1.1: The three streams of inputs: video, eye trace, and audio are aligned with each other. At any given instance, we have a video frame with a corresponding eye trace position and spoken language description.

information and translates them into thoughts that help us in making a decision. Therefore, it is important to have the data of what was the first impression of the person while watching a particular activity. It also helps us in the quantization of eye fixations and surrounding sounds into words. This quantization of the first impression of a particular scene can be best represented through spoken language where a person has to describe what they are seeing while they observe something.

Unfortunately, there is a lack of multi-modal datasets, including videos, eye gaze, and spoken-language description. The existing multi-modal video datasets focus only on some of the modalities, like ActivityNet [3] and YouCook2 [4] have 2 modalities: video and audio, while EGTEA Gaze+ [5] has 2 modalities: video and eye gaze. Therefore, we propose a multi-modal dataset including all 3 modalities: video, eye gaze, and spoken language description.

1.2 Approach

In this thesis, we plan to address the problem of lack of multi-modal video datasets to provide computers with more holistic and detailed information. Therefore, we are proposing a new multi-modal video dataset which includes eye gaze of different viewers along with the spoken-language description of the video.

The dataset is collected in the following manner. An annotator has to watch a

Fig. 1.2: Example of the collected data with eye gaze points marked in red on the video frames, original audio description by an annotator in Japanese, and translated audio description using DeepL language translator.

Table. 1.1: Tasks which can be solved using our multi-modal dataset

Video	**Gaze**	**Text**	**Speech**	**Tasks**
Input	-	Output	-	Video Captioning
Input	Input	Output	-	Dense Captioning
Input	Input	Output	Input	Grounded Speech Recognition
Input	Output	Output	-	Spatially and Temporally Dense Captioning
Input	Output	Input	-	Visual Grounding
Input	Output	-	Input	Grounding/Navigation via Speech
Output	-	Input	-	Video Retrieval

video. While they are watching the video, their eye gaze is continuously monitored using the screen eye tracker: tobii pro nano[1] and they have to simultaneously describe the video in their native language. We have chosen 2 different video datasets for this purpose: ActivityNet [3] and EPIC Kitchens [6]. We have collected 2.5 hours of data with 33 annotators in total. The spoken-language description is in 15 different languages. Fig. 1.1 describes the overall structure of our dataset.

Fig. 1.2 illustrates an example of our collected data for a video snippet from EPIC Kitchens [6]. The example shows the 3 modalities i.e., video, audio, and eye trace. Table 1.1 shows the tasks that can be solved using our multi-modal dataset. The table describes in detail the inputs and outputs to the model required for a

[1] `https://www.tobiipro.com/product-listing/nano/`

particular task. This shows the utility of each modality of the dataset and their interdependence on each other.

We are also introducing a new task which can be solved using our dataset called: spatially and temporally dense captioning. In this task, the model is given the input of a video and it gives the output of text description and gaze point. In simpler terms, given a video snippet, we will get pairs of captions and a gaze point. This task will basically be able to provide us with captions for videos that are localized in time and in the space of the video frame. The advantage of such dense captions is that it will not only provide us with captions for a time interval in the video but also give us its exact location on the video frame. This task will give us captions which are more detailed and richer in information. In this thesis, we have introduced this new task but its performance evaluation is a future work. Here, we instead focus on investigating the utility of multi-modal information in scene understanding to tackle dense video captioning. In this task, the model is given the input of video and gaze points and the output is a caption for different time intervals.

Another aspect which we have covered is the study of active and passive viewers. An active viewer is the one who has to actively describe a video while watching it, and a passive viewer is the one who does not have to describe the video and simply watch it passively. We analyzed the gaze information obtained from both cases to find out that active gaze is much more attentive and helps in capturing the main activity in the video than passive gaze.

Our contributions to this thesis are multifold:

1. We propose a novel multi-modal dataset consisting of video, eye trace, and spoken language description, which aims to provide deep learning models with more dense and holistic information.
2. We also propose a new task: spatially and temporally dense video captioning.
3. We conduct a performance evaluation of the dense video captioning task using additional multi-modal information of eye gaze.

4. We also discuss our study on visual attention by showing that the eye gaze data of active viewers are more attentive than passive viewers.

1.3 Organization of this Thesis

The rest of this thesis is organized in the following manner. In Chapter 2, we review related work on multi-modal image and video datasets and the unique tasks they solve. In Chapter 3, we describe in detail our proposed multi-modal dataset and explain the implementation of gaze modality to the existing video captioning models. In Chapter 4, we present results of the comparison of captions generated using our proposed method with the existing method and also show the results of the various ablation studies conducted. Finally, in Chapter 5, we conclude this thesis and propose future work.

Chapter 2

Related Work

2.1 Existing multi-modal datasets

In this chapter, we will highlight the major multi-modal datasets built using images and videos. The existing multi-modal datasets utilize a subset of additional modalities like eye gaze position, mouse pointer, bounding boxes, spoken language, and text description. We have also highlighted the advantages and limitations of using the selected modalities.

2.1.1 Multi-modal image datasets

The major multi-modal image dataset close to our dataset makes use of visual cues like eye gaze and mouse pointer location information for highlighting the important regions in the image. Using such visual cues helps in narrowing down the human attention into regions of interest in a particular scenario. Doing this helps in getting rid of unnecessary information and gives us an insight into how human attention works.

DIDEC dataset [7] is one of such datasets that consist of eye movements and spoken-language descriptions during an image inspection task. They chose the MS COCO image dataset [8] as their image stimuli which consists of more than 200k images with 5 descriptions per image in English. They chose 307 images from this dataset and collected eye gaze information using SMI RED 250 device and collected

natural language description using a headset microphone of 112 participants. They did a parallel study of collecting data while performing the description task and while performing free viewing of the image. They concluded that the eye tracking data are much more detail-oriented in the description task. In their studies, they used gaze in multiple ways to localize the language description like aggregating gaze of all annotators, using time-dependent annotator specific gaze saliency map, and hidden representation produced from a gaze LSTM model. But they only collected the spoken language in Dutch for their dataset. In one of their following works by van Miltenburg et al. [9], they collected data in both Dutch and English.

The SNAG dataset [10] also collects eye movement and spoken-language description similar to the previous dataset. They selected 100 general-domain images from MS COCO image dataset [8]. They collected their multi-modal dataset on 30 participants who were native speakers of American English using a similar SMI RED 250 device for gaze data and TASCAM DR-100MKII recorder for spoken description. The key point of their work was to show the usefulness of eye gaze and spoken language in visual-linguistic annotation framework. In their work, they identified the fixations and saccades of each annotator and used them individually to label their images. They also proposed a multi-modal alignment method to align the eye gaze and spoken description. This work consisted of non-experts describing the images. In one of their previous works Vaidyanathan et al. [11], they collected similar data on medical images with skilled annotators specializing in that particular medical domain.

All these datasets proposed image annotations which helped to localize spoken language in image space through gaze and are pivotal in understanding the way humans observe and understand images. They however didn't propose image captioning models for their dataset. The multi-modal dataset introduced in He et al. [12] also proposes a captioning model. They collected eye gaze and spoken language description on 1,000 images selected from Pascal-50S dataset [13] and 3,000 images selected from MS COCO image dataset [8] using Tobii X2-30 eye-tracker on 16 subjects. They also performed a study on difference in fixations between free viewing and image description task and came to a similar conclusion that eye

gaze is much more attentive while describing than free viewing. They also came to the conclusion that machine attention and human attention in top-down image captioning does not affect the quality of the caption. Another work proposing image captioning using multi-modal dataset is Takmaz et al. [14] which makes use of the DIDEC dataset [7]. Their work opens doorway to a cognitive approach to a computer vision task. They were able to show different ways to use the gaze modality in current image captioning tasks to further focus on regions which are attractive to human attention.

One important thing missing from both captioning works was the lack of showcasing the importance of time-dependent change in captions based on human attention information in describing images. The work done in Localized Narratives [15] overcomes this limitation by producing controlled captioning. Controlled captioning [16] is a task where the user can specify the image regions to be described and the order of description of the regions. In Localized Narratives [15], they have proposed a multi-modal annotation tool on image datasets in which an annotator hovers their mouse around the image region they are describing at the given moment. This tool is useful for visual grounding of spoken words in the image in a time-dependent pattern. This time dependence in visual grounding tells us a lot about regions which attract a person's attention the most and how it varies depending on people belonging to different areas of expertise, age group, and gender.

However, all the above datasets consists of images. But the real-world environment is better described through videos. The next section discusses the multi-modal datasets that leverage videos and highlights their advantages and limitations.

2.1.2 Multi-modal video datasets

There are many multi-modal video datasets that consider additional modalities such as audio cues and eye gaze. One of the most popular kitchen datasets available currently is EPIC Kitchens [6]. It is the largest egocentric video dataset which has audio coming from the background while the video is being recorded. The dataset is 55 hours long collected by 32 participants belonging to 10 different nationalities

collected by a head-mounted Go-Pro camera. Participants were asked to capture their entire kitchen activity and then narrate them afterwards for action labeling.

Another major video dataset is the ActivityNet dataset [3]. It is a 849 hours long video dataset consisting of a wide range of activities like Eating and drinking Activities; Sports, Exercise, and Recreation; Socializing, Relaxing, and Leisure; Personal Care; and Household Activities. Videos have been collected from online video sharing websites and annotated using crowd sourcing.

YouCook2 dataset [4] is another such web instructional video, which is 176 hours long in total. It contains about 2,000 videos of 89 different recipes spanning the major cuisines of Africa, the Americas, Asia, and Europe. All the video datasets described above, however, only makes use of visual and audio (background noise and spoken description). They do not use specialized visual cues like eye gaze or mouse to further narrow down important regions.

EGTEA Gaze+ dataset [5] overcomes this limitation by introducing eye gaze modality in their egocentric video dataset. It is a 29 hours long dataset consisting of 86 unique sessions. The 32 subjects in their study are asked to cook seven different recipes while their eye gaze is continuously monitored using SMI eye tracking glasses. They still lack detailed description of the video and mainly focus on action recognition and gaze estimation tasks.

Based on our survey on multi-modal video datasets, we have come to the conclusion that there is a dire lack of video datasets with additional modalities like eye gaze, mouse pointer position, or spoken language description. Without such multi-modal datasets, it is not possible to incorporate a human-like cognitive state into computer vision tasks. This inspired us to build a novel multi-modal dataset with 3 modalities: video with their innate audio, spoken language description, and eye gaze trace.

2.2 Utilizing the new modality: Gaze

Gaze provides an important visual cue in computer vision tasks. The egocentric gaze is in the form of the x and y coordinates of the viewer, which gives us information about the regions of the environment that captured their attention. This visual cue can be used in many ways, depending on the task. Some of the common ways to encode eye gaze positions are listed below.

1. Eye fixation: It is the region in the visual field where the person gazed at for longer than the usual period of time.
2. Scan path: It is the trajectory of the eye movement of a person over a period of time.
3. Attention map: It is a 2-dimensional matrix representation of the visual field where the eye position was detected at a given instance of time.
4. Gaze-dedicated deep learning module: It is the hidden representation produced from deep learning models to capture the change in gaze position over time for different viewers.

Different methods use one or more types of gaze encoding methods depending on the task they are solving. For instance, Vaidyanathan et al. 2016 [11] and Vaidyanathan et al. 2020 [17] uses fixations of all participants over time to annotate different regions of the image. In He et al. [18], they made use of eye fixations both individually and as an aggregate over all viewers for their multi-modal image dataset. Analyzing eye fixations individually helped in understanding the differences in eye movements in different types of viewers whereas aggregating them helped in image captioning task by using more than one important region in the image.

In the task of activity prediction, methods like Yang et al. [19] make use of time dependent eye position to identify the regions where the person paid attention as they play an important role in detecting and prediction the activity going on. The

joint gaze estimation and action prediction works done by Li et al. [5] and Huang et al. [20] make use of raw gaze points and encode them as attention maps.

In Boccignone et al. [21], they made use of the scan path of a viewer to understand their eye gaze pattern in different social situations and subsequently use this information for scan path prediction in the video. The same task of predicting the scan path has been solved using GANs in Assens et al. [22].

In the work by Takmaz et al. [23], they trained their description generation model using different types of gaze encoding: aggregating gaze points over participants, time-dependent saliency map for every participant, and encoding gaze as a hidden representation using the LSTM model.

The task we are trying to solve in our work is that of dense video captioning. In such a task, we must preserve the trajectory of gaze points as they give us deeper insight into the change in eye fixations and in turn change in attention regions in the video. That is why, we propose to use raw gaze points in two broad encoding types: soft attention map and hard attention map. When gaze is encoded using a soft attention map, every image feature is weighted by giving more weight to areas near the eye gaze point. In the case of hard-attention map, we completely focus on the regions near the gaze point and mask the rest of the regions.

Chapter 3

Data Collection

The aim of this thesis is to build a novel multi-modal dataset and provide a new cognitive perspective on computer vision tasks. We have shown the utility of multi-modal information, such as gaze and spoken language, in improving existing tasks in Table 1.1. In this thesis, we will focus on the dense video captioning task on two different datasets, namely ActivityNet [3] and EPIC-Kitchens [6].

3.1 Overview

We have already established the importance of eye gaze and spoken language in a multi-modal dataset. In this section, we will focus on discussing the method in which we collect and use this multi-modal information. The following chapter can be divided into 4 parts- multi-modal data collection, preprocessing, and comparison with existing datasets.

3.2 Multi-modal data collection

This section will describe our dataset collection in details. Our setup consists of 3 data components- video stream, eye trace, and audio description. We are using videos from the ActivityNet [3] and EPIC-Kitchens [6] datasets, as they cover a range of videos such as outdoor activities, cooking, sports, and socializing

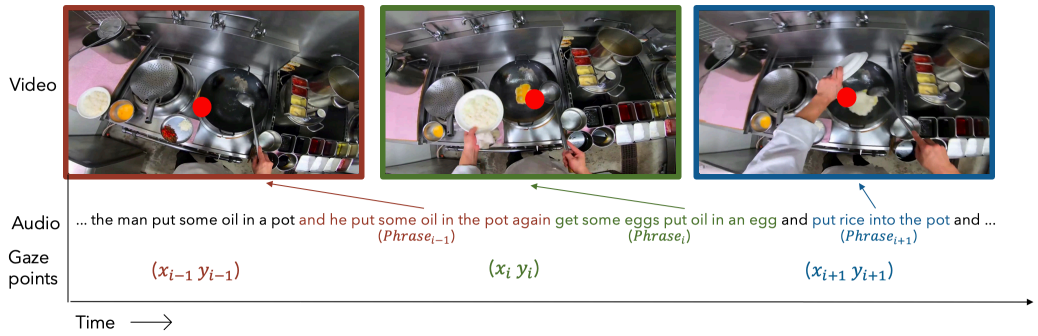

Fig. 3.1: The 3 information stream in our collected data are video, audio, and gaze (represented by red circles on the video)

Table. 3.1: Summary of our dataset

Video Dataset	Total video length(mins)	Number of active annotators	Number of passive annotators	Number of languages
ActivityNet	75	15	18	8
EPIC-Kitchens	75	33	0	15

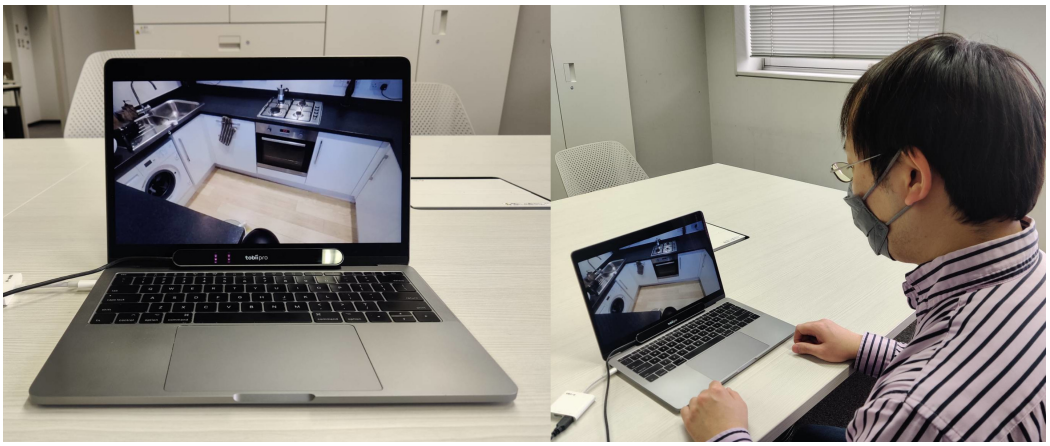

Fig. 3.2: Data collection setup. Front view of the laptop setup with eye tracker on the bottom of the screen (left), snapshot of the data collection environment with the annotator (right).

activities. We use the Tobii Pro Nano eye tracker[1] to collect the eye trace. Both left and right eye data are captured and averaged to obtain the gaze point in any particular instance. The video is played at a frame rate of 60 fps, and the eye tracker also collects samples at a rate of 60 fps. A snippet of the collected data can be visualized in Figure 3.1. As you can observe there is a certain lag between the eye gaze position and the corresponding speech phrase. This is because of the reaction time of the annotator. In most cases, this reaction time is very small and can be ignored for computational purposes.

Data collection is carried out as follows. We begin the data collection by showing videos to the annotator. This helps them to get a general idea of the video and gives them an opportunity to look up the vocabulary of difficult words. After this, the calibration of the eye tracker is done. After that, the collection begins by showing the annotator the video.

There are two different types of annotators in our collection. The first group of annotators has to actively describe the video while it is being played. They are referred to as active annotators in the rest of the thesis. The second group has to passively watch the video without saying anything. They are referred as passive annotators in the rest of the thesis. The active annotator has to describe the video to the best of their knowledge. They do not necessarily need to have expert knowledge about the ongoing task. We have chosen task-oriented videos for this purpose as they involve many objects and sub-tasks. During the collection, we have recording the eye gaze of the annotator continuously. The data collection setup is shown in Figure 3.2.

For 75 minutes of EPIC-Kitchens dataset [6] we had 33 active annotators. For 75 min of ActivityNet dataset [3] we had 15 active annotators and 18 passive annotators. Therefore, in total we have collected 2.5 hours of multi-modal data on both the video datasets. The audio data were collected in 15 different languages with the distribution being as follows. 13 annotators used Japanese, 7 annotators used English, 6 annotators used Chinese, 3 annotators used Hindi, 2 annotators used Per-

[1] `https://www.tobiipro.com/product-listing/nano/`

sian, 2 annotators used Spanish(Spain), 2 annotators used Spanish(Columbia), 2 annotators used Urdu, 1 annotator used Arabic(Jordan), 1 annotator used French, 1 annotator used Gujarati, 1 annotator used Indonesian, 1 annotator used Italian, 1 annotator used Portuguese(Brazil), and 1 annotator used Turkish. The summary of our dataset is shown in Table 3.1.

3.3 Preprocessing

This section describes the preprocessing of eye trace and audio data. The eye gaze is collected for the left and right eyes. After verifying the accuracy of both eye data, we came to the conclusion that in most cases the most accurate eye position was given by the average of the gaze positions of the left and right eye.

We devised a simple experiment for this purpose. We made a video of a red dot moving around a black screen at predefined positions. The experimenters had to follow the dot around the screen with their eyes. After collecting the eye trace of 10 people on such a video, we analyzed the variance between the eye trace collected and the ground truth position of the red dot. Therefore, we settled on using the average of the left and right eye trace. But during the actual data collection, we observed that in some cases, one of the eye trace data was erroneous showing null value most of the time. The reason behind it was that the eye was not tracked properly by the tracker. In such cases, we checked if the other eye trace data was correct and used it directly.

The audio was pre-processed in the following manner. We transcribed the audio files using Google Cloud Speech API which converted the speech into text along with the timestamp of each phrase. The model gives us the transcription along with the confidence value for each translated phrase. We discard the data where the cumulative confidence is below 0.3 as the translation is very erroneous at that point. After transcription, the files are sent to the respective annotators to check for transcription errors. We observed that it took about 15 minutes to correct a file corresponding to a 5-minute video. After receiving the corrected transcript,

we translated the files using the DeepL translator. At this stage, the groundtruth captions for videos are ready.

3.4 Comparison with existing datasets

Table. 3.2: Comparison between different multi-modal datasets

Dataset	Image/ Video	Size of dataset	Visual cues	Audio cues	Annotation type	Number of annotators
Localized Narratives [15]	Image	849k images	Mouse	Yes	Dense	156
SNAG [10]	Image	100 images	Gaze	Yes	Regular	30
DIDEC [7]	Image	307 images	Gaze	Yes	Regular	112
ActivityNet [3]	Video	849 hrs	-	Yes	Regular	-
YouCook2 [4]	Video	176 hrs	-	Yes	Regular	-
EGTEA Gaze+ [5]	Video	28 hrs	Gaze	No	Regular	32
Ours	Video	2.5 hrs	Gaze	Yes	Dense	33

The comparison of our dataset with existing multi-modal dataset can be seen in Table 3.2. Localized Narratives [15] has the largest image dataset in the table with 849k images annotated by 156 professional annotators who worked on it full-time during the data collection phase. However, the visual cue used is through a mouse, which may create a delay between when the person was seeing the object and when they bought the mouse near that area. The annotation type they produce are dense. SNAG [10] and DIDEC [7] are somewhat smaller image datasets with only 100 and 307 images, respectively, and 30 and 112 annotators each. The visual cue used by both of them is gaze which is a better indicator of the person's attention. All three datasets use audio cues in the form of spoken language description.

Video datasets like ActivityNet [3] and YouCook2 [4] are huge with 648 hrs and 176 hrs of videos respectively. They do not use any visual cues and only produce regular annotations on videos. The EGTEA Gaze+ [5] dataset is a little smaller with 28 hr of videos and 32 annotators, but they have collected gaze data of people recording the videos in first-person view. The annotations they produce are still regular. The audio cues used by them are part of the video in the form on

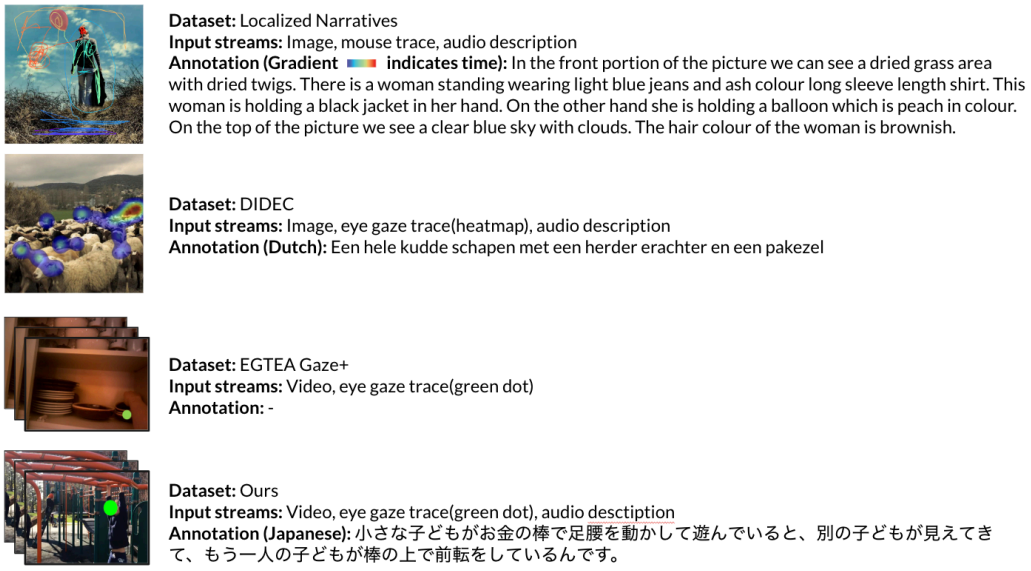

Fig. 3.3: Comparison of our dataset with existing multi-modal datasets. Modalities in different datasets are listed as follows- (From top) Localized narratives: image, mouse trace, audio description; DIDEC: image, eye gaze trace, audio description; EGTEA Gaze+: video, eye gaze trace; Ours: video, eye gaze trace, audio description.

audio associated with step-by-step instructional videos or audio coming from the ambience.

On comparison, our dataset is smaller in size with only 2.5 hrs of videos but it has been annotated by 33 participants. We also collected the participants gaze data as a visual cue and the audio description as an audio cue. These multi-modalities of gaze and audio help us to create dense annotation in space and time.

Chapter 4

Proposed Method

In this chapter, we will first describe the implementation details of the gaze modality in our dense captioning model. Then we will move on to discuss in detail different ways to encode eye gaze for this task.

4.1 Overview

After demonstrating the benefits of multi-modality in computer vision tasks, we'll move on to implement multi-modality in dense video captioning tasks. The architecture of our dense captioning model is composed of two parts, one is a pre-existing bi-modal dense captioning model, and the other is the additional gaze model added to it. The details of both parts of our model are described in detail in the next sections.

4.2 Dense captioning model

This section describes the dense video captioning model utilized in this work. The dense video captioning model that we have chosen is a bi-modal transformer model called BMT [1]. The 2 modalities it considers is video features extracted from I3D [2] and audio features extracted from VGGish [24]. We propose a modification to their model by introducing a third modality, Gaze, in the input. The overall flow of the model is shown in Figure 4.1.

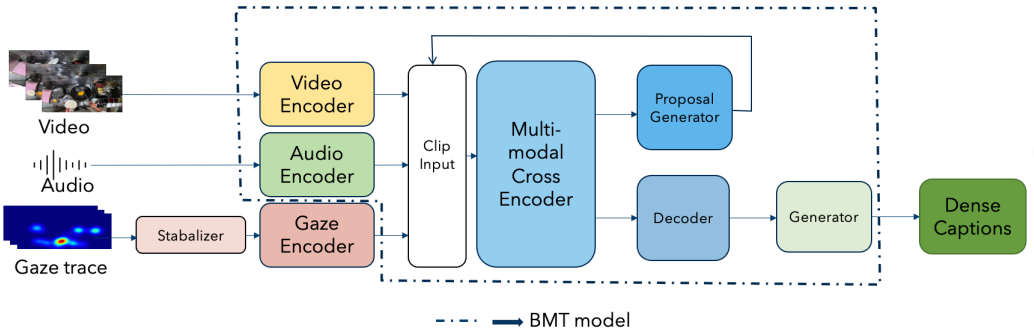

Fig. 4.1: Our proposed dense captioning model is inspired from the BMT model [1]. The doted box represents the original BMT model [1]. We proposed to add a 3rd input modality of eye gaze trace which will help in clipping the video features and focus only on the important details to human eye.

The task of dense captioning in multi-modal datasets have been addressed by methods like [25, 26, 27, 1]. The overall task can be solved by dividing it into 2 parts - event proposal generator to detect events followed by the captioning model. The detecting of events is done using different visual and audio cues. For example, in SDVC [27], they simply use basic visual features. In DVC [26], they use visual features with optical flow to capture the temporal change in events. The method proposed by Masked transformer [25] takes a novel approach of masking certain parts of proposal event to restrict its attention. All these methods, however, focus only on visual features. The method proposed by BMT [1], takes a step further into the importance of multi-modal information by proposing a bi-modal transformer architecture for dense captioning using the bi-modal input of video and audio.

We propose to introduce the additional modality of eye gaze to further narrow down the location of important events in the video in the dense captioning task. This can be done by encoding gaze into video features in different ways. We discuss it in detail in Section 4.3. To address this, we are taking help from one of the existing bi-modal dense captioning models and modifying it for our task.

One of the best bi-modal dense captioning models present today is the BMT

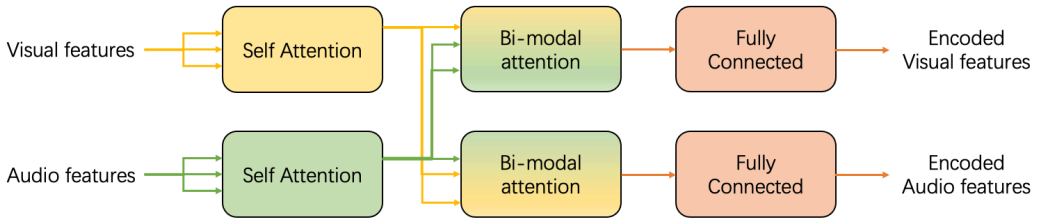

Fig. 4.2: Encoder module of the BMT model [1]. Input are the visual and audio features and output are the bi-modal encoded visual and audio features.

model [1] as it is one of the few models that considers both audio and visual features of equal importance. That's why we believe that extending it to a tri-modal model will further help us focus on the important video regions. The BMT [1] captioning model can basically be divided into 2 parts. The first part is the bi-modal transformer to which the audio and video features are given as input. After passing the N layers of features through the bi-modal encoder, they are passed through the multi-headed proposal generator. The bi-modal transformer consists of a bi-modal encoder, bi-modal decoder, and a generator. The bi-modal encoder encodes the audio and video features into 2 different ways: visual attended audio features and audio attended visual features. The bi-modal encoder used by BMT [1] is special in the sense that it uses 2 input streams of visual and audio features. All N layers of the encoder have three sub-layers: self-attention, bi-modal attention, and position-wise fully connected layers. The network flow of the encoder is described in Figure 4.2. The layers can be defined as follows:

Let the audio features at the $n-1^{th}$ layer be denoted by A_{n-1}^{fc} and visual features at the $n-1^{th}$ layer be denoted by V_{n-1}^{fc}. At the n^{th} layer, these audio and visual features of the previous layer are encoded into feature maps A_n^{self} and V_n^{self} via the self-attention layer:

$$\begin{aligned} A_n^{self} &= MultiHeadAttention(A_{n-1}^{fc}, A_{n-1}^{fc}, A_{n-1}^{fc}) \\ V_n^{self} &= MultiHeadAttention(V_{n-1}^{fc}, V_{n-1}^{fc}, V_{n-1}^{fc}) \end{aligned} \tag{4.1}$$

Then, the two outputs of the self attention layer of visual features denoted by

V_n^{self} and one output of the self attention layer of the audio features denoted by A_n^{self} are passed through one of the bi-modal attention layers to get visual attended audio features A_n^{mm}:

$$A_n^{mm} = MultiHeadAttention(A_n^{self}, V_n^{self}, V_n^{self}) \tag{4.2}$$

Similarly, the two outputs of the self attention layer of audio features A_n^{self} and one output of the self attention layer of the visual features V_n^{self} are passed through one of the bi-modal attention layer to get audio attended visual features of V_n^{mm}:

$$V_n^{mm} = MultiHeadAttention(V_n^{self}, A_n^{self}, A_n^{self}) \tag{4.3}$$

Finally, the outputs of the audio attended visual features V_n^{mm} and visual attended audio features A_n^{mm} are encoded by passing through fully connected layers to obtain the encoded audio features A_n^{fc} and encoded video features V_n^{fc}:

$$\begin{aligned} A_n^{fc} &= TwoFullyConnected(A_n^{mm}) \\ V_n^{fc} &= TwoFullyConnected(V_n^{mm}) \end{aligned} \tag{4.4}$$

The output of the encoder layer is then sent to the proposal generator. The proposal generator head is a fully-convolutional network with 3 layers. This design is inspired by YOLO object detector [28, 29, 30]. The first layer has a kernel size of k followed by a size 1 kernel size in the second and third layer. For every input sequence of T features, predictions are made at each time step in the interval [1, T] and at every segment length anchor in the set Ψ. For each proposal, the temporal boundaries and the confidence are calculated as follows.

$$\begin{aligned} center &= p + \sigma(c) \\ length &= anchor \cdot exp(l) \\ confidence &= \sigma(o) \end{aligned} \tag{4.5}$$

where $\sigma(c)$ is the sigmoid function bounded between values [0, 1] of the segment center relative to a position p in a sequence. $exp(l)$ is the coefficient for an anchor, and $\sigma(o)$ is the objectiveness score.

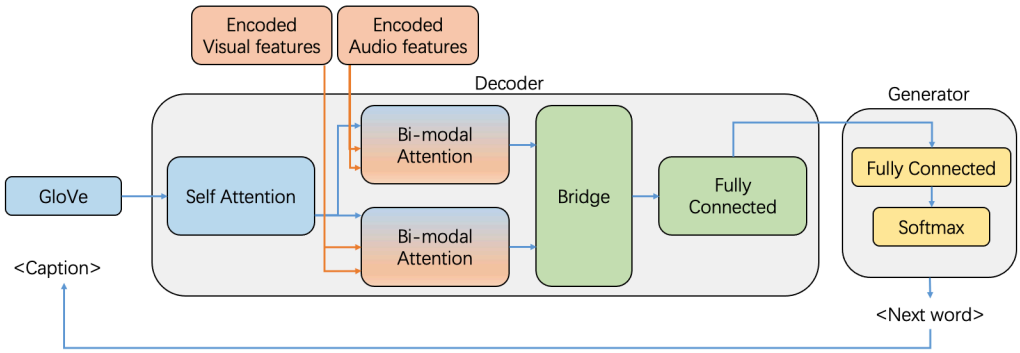

Fig. 4.3: Decoder and generator module of the BMT model [1]. Input are the bi-modal encoded visual and audio features and output is the next word in the caption.

The center and grid are converted to seconds by multiplying by the cell size of the temporal span of the feature. The total proposals generated at this stage are $3 \cdot (T_a \cdot K_a \cdot |\Psi_a| + T_v \cdot K_v \cdot |\Psi_v|)$ where T_a and T_v are the length of the audio and visual input features, K_a and K_v are the kernel sizes of the audio and video features, and $|\Psi_a|$ and $|\Psi_v|$ are the anchor sizes of the audio and visual modalities. From these, the top 100 proposals are selected based on the confidence score calculated above. To maintain fairness between different modalities, the values of $T_a \cdot |\Psi_a|$ and $T_v \cdot |\Psi_v|$ are set to almost equal. In addition, the kernel sizes K_a and K_v are set to the same values. The generated proposals at this stage are then used to clip the input features which are passes through the encoder again.

At this stage, the output of the encoder layer is sent to the decoder layer. Each layer of the N-layered decoder module then uses the output of the encoder layer. The decoder takes the multi-modal representation and models these features with the previous caption word at any given time. The output of this model is a representation which is employed to model a distribution of the next word in the caption. Each decoder layer can be sub-divided into 4 sub layers: self-attention, bi-modal encoder-decoder attention, bridge, and position-wise fully-connected layers. The layers can be defined as follows:

Let the caption features be denoted by C. At the n^{th} layer, the caption features of the previous layer denoted by C_{n-1}^{fc} are encoded through the caption self-attention layer:

$$C_n^{self} = MultiHeadAttention(C_{n-1}^{fc}, C_{n-1}^{fc}, C_{n-1}^{fc}) \tag{4.6}$$

Then the output of the caption self attention layer C_n^{self} and the two outputs of the encoded audio features A_n^{fc} are passed through the bi-modal attention layer to obtain audio-visual attended previous caption denoted by $C_n^{A^v}$:

$$C_n^{A^v} = MultiHeadAttention(C_n^{self}, A^v, A^v) \tag{4.7}$$

Similarly, the output of the caption self attention layer C_n^{self} and the two outputs of the encoded visual features V_n^{fc} are passed through the bi-modal attention layer to obtain the visual-audio attended previous caption denoted by $C_n^{V^a}$:

$$C_n^{V^a} = MultiHeadAttention(C_n^{self}, V^a, V^a) \tag{4.8}$$

Then the outputs of both the bi-modal attention layers $C_n^{A^v}$ and $C_n^{V^a}$ are passed through the bridge followed by a fully connected layer to get the caption output features of the decoder layer denoted by C_n^{fc}:

$$\begin{aligned} C_n^{mm} &= OneFullyConnected([C_n^{A^v}, C_n^{V^a}]) \\ C_n^{fc} &= TwoFullyConnected(C_n^{mm}) \end{aligned} \tag{4.9}$$

This output is sent to the event proposal generator to generate the next word. The word tokens are embedded with the pretrained GloVe model [31]. The structure of the generator is a fully-connected layer with softmax activation which outputs the distribution for the next word in the caption c_{t+1} for a given decoder output $C_t^{av} \in \mathbb{R}^{t \times d_c}$ where d_c is the dimension of the caption feature. The network flow of the decoder and generator module to generate the next word in the caption is represented in Figure 4.3.

The beginning of the model is where the extraction of audio and visual features takes place. The audio features are extracted from the input video with a pre-trained VGGish model [24]. The video features are extracted with a pre-trained

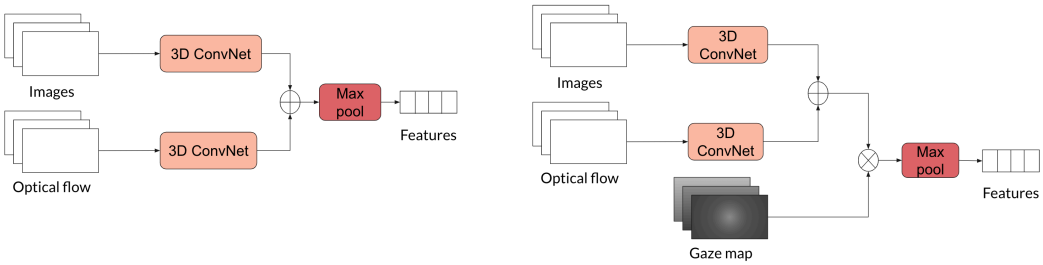

Fig. 4.4: Architecture of the video features extraction model. On the left, is the original 2 stream 3D-ConvNet I3D model [2]. The streams of Images and Optical flow are individually passed through 3D-ConvNet networks, combined, and then max pooled to get the final features. On the right, is our modified 3 stream 3D-ConvNet I3D model. The encoded gazemap is multiplied with the combined 3D-ConvNet output and then max pooled to get the final features.

I3D model [2]. We introduce our gaze modality at this step. The gaze features are multiplied by the visual features right before the pooling step.

Figure 4.4 shows this implementation in a pictorial way. The image on the left shows the original implementation of the I3D model [2]. The image on the right shows the scenario when gaze features are introduced in the visual features. The gaze features introduced in the I3D feature extraction step are of 3 types as discussed in the previous section, namely the attention map, the attention map with fixations, and the mask. Hence, our tri-modal input is converted into bi-modal in order to be processed by the bi-modal encoder of the BMT model [1].

4.3 Adding a new modality: Gaze

This section describes the different ways in which we propose to use the eye gaze modality in the dense video captioning task. As discussed in section 2.2, gaze can be used as an aggregate of many annotators or a single annotator. It can also be averaged over time to get a sense of overall important regions in a scene, or it can be considered in a time-dependent fashion to get a sense of priority of different important regions to the human eye.

In our work, we plan to use gaze trace of a single person at a time in a time dependent manner. This happens because the scene in a video changes over time and depends on the individual who watches and describes the video. The format in which we receive the eye trace is very straightforward. We obtain the x and y coordinates of the left and right eyes of the person. After testing our tracker on 10 different annotators we came to the conclusion that we should average the coordinate positions of left and right eyes for the best estimate of where the person was looking.

These coordinates are encoded as attention maps which are applied to our video features. The definition of an attention map as described in Jetley et al. [32] is a scalar matrix that represents the relative importance of layer activations at different spatial 2D locations with respect to the target task. The motivation behind using attention maps is to eliminate background information and focus on the regions where the eye position was recorded.

Given an input matrix $x \in \mathbb{R}^{H \times W}$, a feature matrix $z \in \mathbb{R}^{h \times w}$, an attention matrix $a \in [0,1]^{h \times w}$, output matrix $y \in \mathbb{R}^{h \times w}$, and an attention network with parameters ϕ as $f_\phi(x)$, attention can be represented as-

$$\begin{aligned} a &= f_\phi(x) \\ y &= a \odot z \end{aligned} \tag{4.10}$$

where $\odot$ is element-wise multiplication. y gives us the output matrix after applying an attention mechanism.

In our work, we propose to use these gaze position coordinates in 3 different attention manners-

1. Encoding gaze as a hard attention map using masking
2. Encoding gaze as a soft attention map
   (a) Soft attention map on every frame

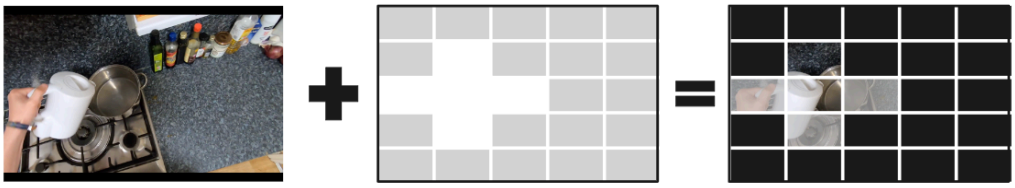

Fig. 4.5: Visual representation of adding mask to an image in our hard attention map encoding of gaze.

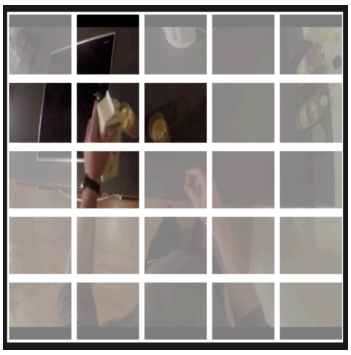

Fig. 4.6: (a) Hard attention map using mask

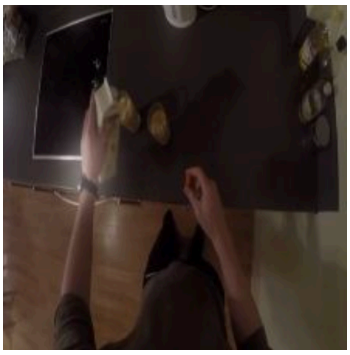

Fig. 4.7: (b) Soft attention map

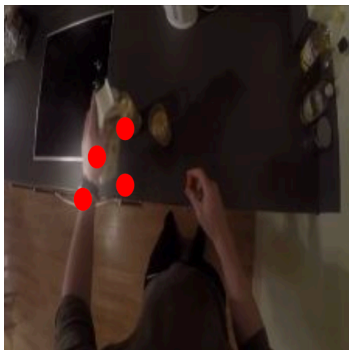

Fig. 4.8: (c) Soft attention map with fixations

Fig. 4.9: Different eye gaze encoding methods in our dense captioning task- (a) Hard attention map using mask, (b) Soft attention map using a Gaussian map, (c) Soft attention map using a Gaussian map with fixations

(b) Soft attention map on frame where eye fixations are detected

4.3.1 Encoding gaze hard attention map using masking

In this subsection, we will talk about how to encode gaze as a hard attention map. In Equation (4.10), if the attention matrix a can only take values 0 or 1 then it becomes a hard attention mechanism. In simpler terms, we are basically blocking out some parts of the image completely and only using the features of some certain parts. We have implemented this using a mask whose diagram is

shown in Figure 4.5. A mask is a grid like representation of an image and can have values either 0 or 1. We have assigned the value 1 near the gaze position and 0 everywhere else. The visual representation of this masking can be seen in Figure 4.6. The cells in the mask that have 0 values have been blurred in the image for visualization purposes. In the actual computation, that region is completely omitted.

4.3.2 Encoding gaze as a soft attention map

In this subsection, we will talk about how to encode gaze as a soft attention map. In a soft attention map, unlike in hard attention, the attention matrix from equation. (4.10) can take any value between 0 to 1. It basically blurs out some of the parts without completely removing them. We implement this soft attention using Gaussian filters around the gaze position of the eye. The equation of Gaussian function is given by-

$$G(x, y) = \frac{1}{2\pi\sigma^2}\epsilon^{-\frac{x^2+y^2}{2\sigma^2}} \tag{4.11}$$

where σ is a tuning parameter. We have used multiple values of σ in our experiments to achieve the best performance.

Soft attention map on every frame

The soft attention mechanism mentioned above can be utilized in many ways. The most straightforward way is to apply it to every frame in the video based on the eye gaze value in that frame. This is what we are doing with this method. This method will be abbreviated as GMσ where σ is the value of the parameter in the Gaussian function, in the rest of the thesis. The visual representation of this method can be seen in Figure 4.7. The Gaussian map helps to focus on the regions where the gaze was detected and darkens the rest of the image.

Soft attention map on frame where eye fixations are detected

In this method, we use the Gaussian soft attention map only in certain situations. We have observed that the entire eye trace of a person may contain a lot of unnecessary and redundant values. For example, when a person is moving their gaze from one object to another they don't necessarily process the regions in between even though their eye gaze may be found in that area. Also, there are some involuntary eye movements from time to time which leads to error in deciding if that region was indeed important or not. To avoid these problems, we perform eye fixation detection to find out clusters of eye gaze positions. In this way, we eliminate the transitory gaze between fixations and involuntary eye movement. Once we detect these fixations, we apply our soft Gaussian attention map only on the frames where the fixations were detected. This method is referred as Fixσ, where σ is the value of the parameter in the Gaussian function, in the rest of the thesis. Figure 4.8 shows how the Gaussian map is applied when fixation is detected in a certain video frame.

Chapter 5

Experiments

In this chapter, we will discuss the qualitative and quantitative results of our dense video captioning task. We have used the dense captioning model proposed by BMT [1] as the core model and modified it to incorporate the 3rd modality proposed by us, which is gaze. First, we will do the comparison on ActivityNet [3] followed by EPIC-Kitchens [6]. Finally, we will discuss the ablation study in which we showcase the performance improvement in adding gaze to captions on a subset of our collected dataset. We have performed a caption comparison using the evaluation code proposed by Krishna et al. [33]. We are considering the following metrics for comparison- BLEU@1-4 [34], METEOR [35], ROUGE_L [36], and CIDEr [37]. All these metrics are averaged over every video.

5.1 Comparison of captions on ActivityNet

Our dataset contains 75 minutes long videos from ActivityNet [3] which includes multitude of categories like outdoor activities, socializing, household activities, leisure, and recreation activities, to name a few. Each video had at least 2 annotators including active and passive annotators. So in total, there were 15 active and 18 passive annotators. We generated the captions using our proposed modified BMT model. An example of the generated captions along with the collect spoken language description, and ActivityNet groundtruth can be visualized in Figure 5.1.

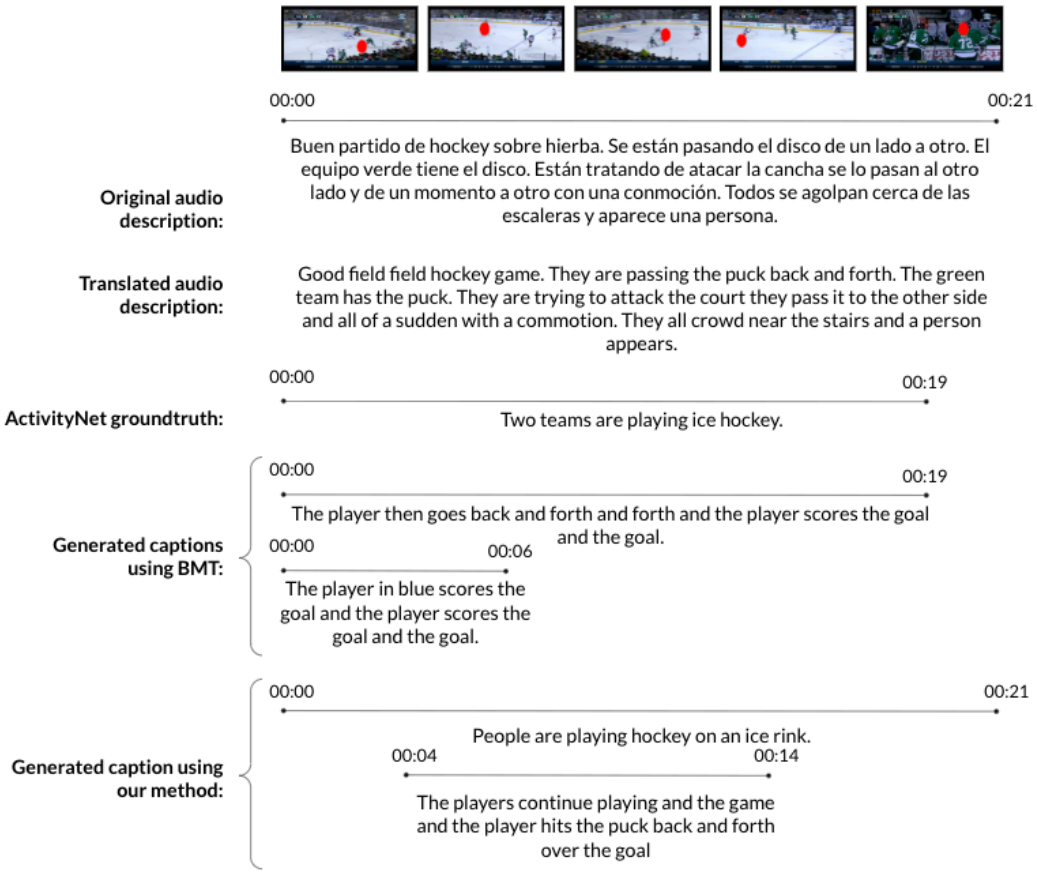

Fig. 5.1: Example of the collected data with eye gaze points marked in red on the video frames, original audio description by an annotator in Spanish(Columbia), translated audio description using DeepL language translator, ActivityNet groundtruth, generated captions using BMT model, and generated captions using our proposed method.

5.1.1 Comparison with groundtruth of ActivityNet

In this section, we will show the numerical comparison of our generated captions with existing dense captioning models. Table 5.1 shows this complete comparison. The method proposed by Masked Transformer [25] makes use of visual features with a mask on their proposal event module to generate captions. DVC [26] makes use of visual features and optical flow to generate captions. SDVC [27] simply uses

Table. 5.1: Comparing quality of captions generated from different dense captioning models. The top 2 values of each metric are highlighted for emphasis.

Method	BLEU@1	BLEU@2	BLEU@3	METEOR	ROUGE_L	CIDEr
Masked Transformer [25]	9.96	4.81	2.42	4.98	-	9.25
DVC [26]	12.22	5.72	2.27	6.93	-	12.61
SDVC [27]	**17.92**	**7.99**	**2.94**	**8.82**	-	**30.68**
BMT [1]	**13.48**	**7.04**	**3.49**	**7.56**	**12.75**	**26.32**
Ours	10.79	5.18	2.51	6.26	**10.34**	12.72

basic visual features to generate captions. BMT [1] is the only model out of all the works mentioned in the Table 5.1 that uses multiple modalities, as it uses both the encoded audio and visual features separately. Our proposed model goes one step further and uses the third input modality of eye gaze.

The best performance is shown by SDVC [27] even though it only uses basic visual features. The second best performance is exhibited by BMT [1] even after using multiple modalities. Our method performs better than the mask transformer [25] and DVC [26]. We believe that such behaviour is observed due to 2 main reasons. Firstly, the size of the dataset used by other methods for testing is much larger than ours. The second reason is that our proposed method uses the pretrained BMT model which is trained on visual and audio features. When our method gets rid of some of the visual features by focusing on areas based on gaze information, the pre-trained BMT model gets less visual information than it is trained on and hence degrades the quality of the caption. We believe that this can be solved by retraining the model with all 3 modalities. But due to the small size of our dataset, it will lead to overfitting and lose its generality.

Table. 5.2: Comparing captions generated from different types of gaze encoding, namely- BMT(baseline), BMT+Mask, BMT+GM, BMT+Fix on different values of parameter λ. The top 2 values are highlighted in each metric for emphasis.

Method	BLEU@1	BLEU@2	BLEU@3	METEOR	ROUGE_L	CIDEr
BMT [1]	**13.48**	**7.04**	**3.49**	**7.56**	**12.75**	**26.32**
BMT + Mask	10.59	4.83	2.10	6.04	**10.42**	**14.48**
BMT + GM140	**10.79**	**5.18**	**2.51**	**6.26**	10.34	12.72
BMT + GM70	10.21	4.44	1.74	5.70	10.00	12.03
BMT + GM35	8.80	3.07	0.95	4.70	8.53	7.70
BMT + Fix140	9.92	4.75	2.00	5.66	9.56	11.90
BMT + Fix70	10.08	4.43	2.04	5.93	9.85	13.33
BMT + Fix35	9.60	4.31	1.91	5.41	9.30	11.67

5.1.2 Comparison of captions generated from different gaze encoding

Numerical metrics

In this subsection, we will highlight the difference in performance of our proposed method using 3 different types of eye gaze encoding with different parameters. The 3 ways are- using gaze as a hard attention map with mask, using gaze as a soft attention map, and using gaze as a soft attention map during eye fixation. The numerical comparison can be seen in Table 5.2.

The best performance is shown by the original BMT method because of the reason mentioned in the previous section. Of all the ways to encode gaze, the best performance is shown when gaze is encoded as a soft-attention Gaussian map with the parameter $\lambda = 140$. This is followed by the method in which gaze is encoded as a mask in every video frame.

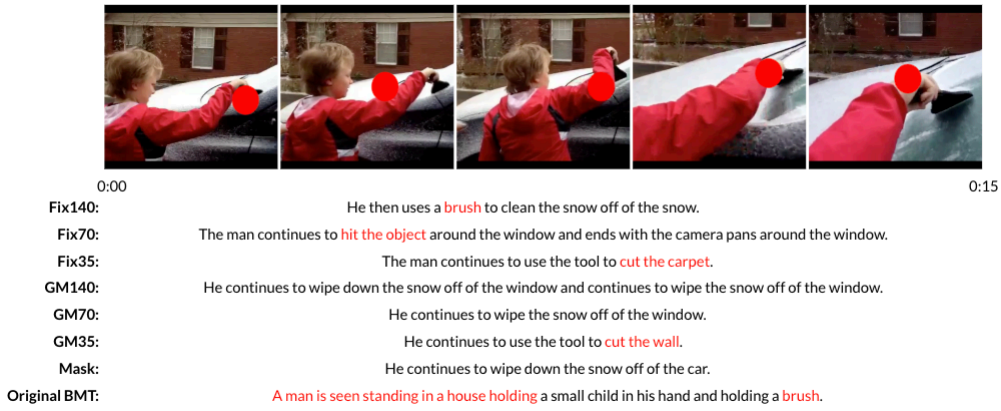

Fig. 5.2: Example of captions generated from different gaze encoding methods on ActivityNet video. The red dots on the video frame correspond to the eye gaze position. The incorrect words in the captions are highlighted in red.

Qualitative comparison

Figure 5.2 shows the comparison of captions generated from different gaze encoding techniques on an ActivityNet video. The eye gaze position of each video frame is shown with a red dot on the frame. The incorrect parts in the caption are shown in red font for emphasis. The caption generated from the original BMT method is the most erroneous. The method GM35 also produced an incorrect caption. The rest of the methods got somewhat correct captions with different degrees of correctness.

5.1.3 Comparison of captions of Active and Passive annotators

In this subsection, we will compare the difference in captions generated from the eye gaze of active and passive annotators. Our hypothesis is that the gazes of active annotators are more attentive than those of passive annotators. This is because active annotators pay extra attention as they have to describe the video while watching it. However, passive annotators can get distracted as they simply watch the video without describing it. We are comparing the generated captions

from both in 2 different ways- through numerical metrics to get a quantitative idea and through human evaluation to get people's perspective.

Numerical metrics

Table. 5.3: Comparison of active and passive captions generated from different eye gaze encoding. The top two values are highlighted for each metric.

Method	Annotator type	BLEU@1	BLEU@2	METEOR	ROUGE_L	CIDEr
BMT + GM70	Active	5.51	1.57	3.16	5.28	6.23
	Passive	6.28	2.86	**4.20**	6.25	**11.07**
BMT + GM35	Active	**6.32**	**3.11**	**4.40**	**6.28**	6.48
	Passive	5.81	2.80	3.87	5.61	5.82
BMT + Fix70	Active	**7.46**	**3.20**	4.19	**7.30**	**11.43**
	Passive	6.05	2.56	3.60	5.76	5.92
BMT + Fix35	Active	5.63	2.54	3.28	5.38	9.46
	Passive	5.98	2.48	3.68	5.78	6.57

Table 5.3 shows the BLEU@1-2, METEOR, ROUGE_L, and CIDEr scores for active and passive captions generated using different gaze encoding methods. From the table, it is evident that in most cases the higher metric value is found to be for active captions. This is in line with our hypothesis that active captions are more attentive to details than passive captions. Also, it should be noted that the captions generated from the method BMT + Fix70 show the best performance out of all the methods. This further strengthens our argument that eye fixations help in narrowing down the important regions in a video.

Human evaluation

In this subsection, we will discuss the performance of captions generated from active and passive captions using the human evaluation survey. The survey was carried out on 23 participants, consisting of 3 women and 20 men who belong to the age group of 20-30 years. Figure 5.3 shows a sample question of the survey. There

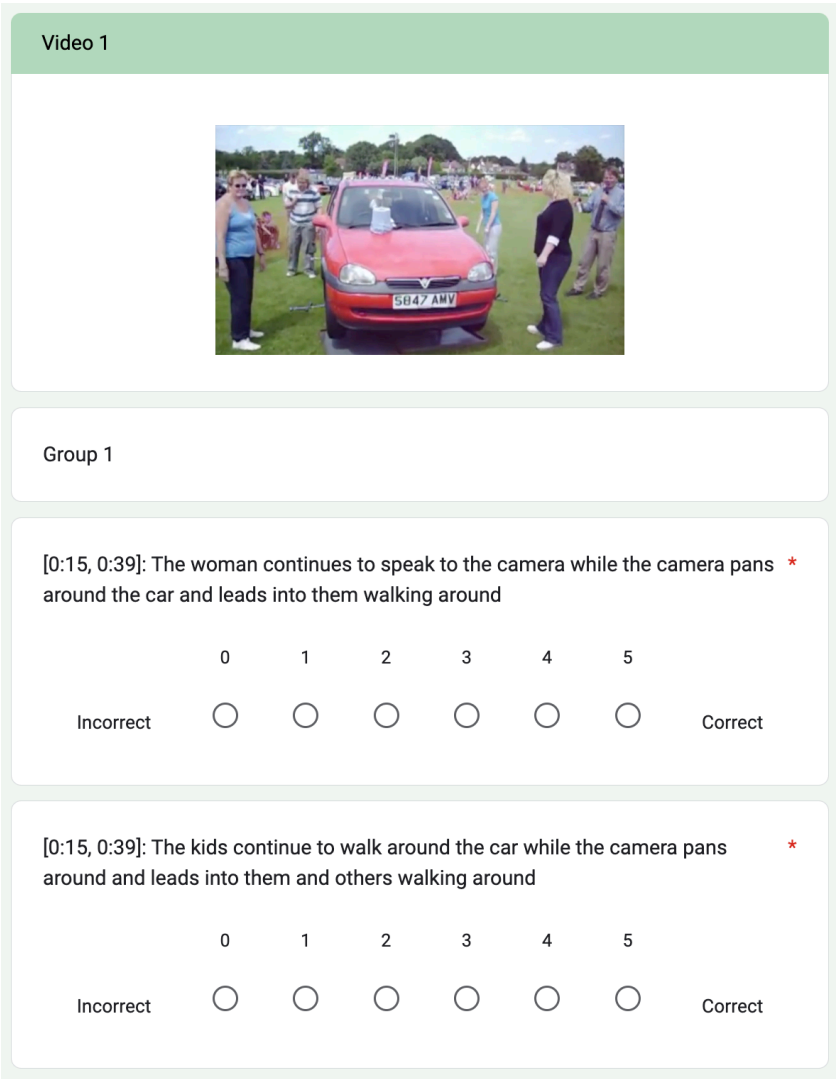

Fig. 5.3: Example of one of the caption comparison question in the human evaluation survey. There are many pairs of captions for each video, one generated from active and one generated from passive annotator, placed in a random order. Both captions belong to about the same time interval. The annotator has to rate both captions between 0 to 5 based on the accuracy of the sentence and the correctness of the time interval.

are multiple groups for each video, and each group contains a pair of captions. One of the captions is active, the other is passive, and they are placed in a random order.

Table. 5.4: Rating of captions generated from different gaze encoding for active and passive annotators. The rating is between 0-5, 0 being the lowest and 5 being the highest. The top 2 values are highlighted for emphasis.

	Active				Passive			
	BMT + Fix70	BMT + Fix35	BMT + GM70	BMT + GM35	BMT + Fix70	BMT + Fix35	BMT + GM70	BMT + GM35
Rating(0-5)	**4.12**	3.65	**3.71**	3.48	3.61	3.2	3.58	3.64

The participant has to rate both the captions from 0-5 where 0 corresponds to completely incorrect caption and 5 corresponds to completely correct. Correctness is based on the accuracy of the caption and the correctness of the time interval corresponding to the caption.

Table 5.4 shows the overall rating of the captions of the survey participants. As is evident, the rating for active captions is higher than for passive captions except for one case. Even within active captions, the rating is highest for the method using gaze as fixations and $\lambda = 70$. This is similar to our observations in the metric comparison in the previous section.

Qualitative comparison

In this subsection, we will discuss some example captions of active and passive cases. It will help us in understanding the difference in captions generated from the 2 cases. Figure 5.4 shows some of the cases where active captions are much more descriptive than passive captions. The first example is a clear indicator of this behavior. The second example shows how the passive captions are erroneous towards the end. On the other hand, there are some cases where the captions generated from active and passive are both partially correct. This is illustrated in Figure 5.5. In this case, both captions are partially correct, so it is difficult to compare the quality in such cases.

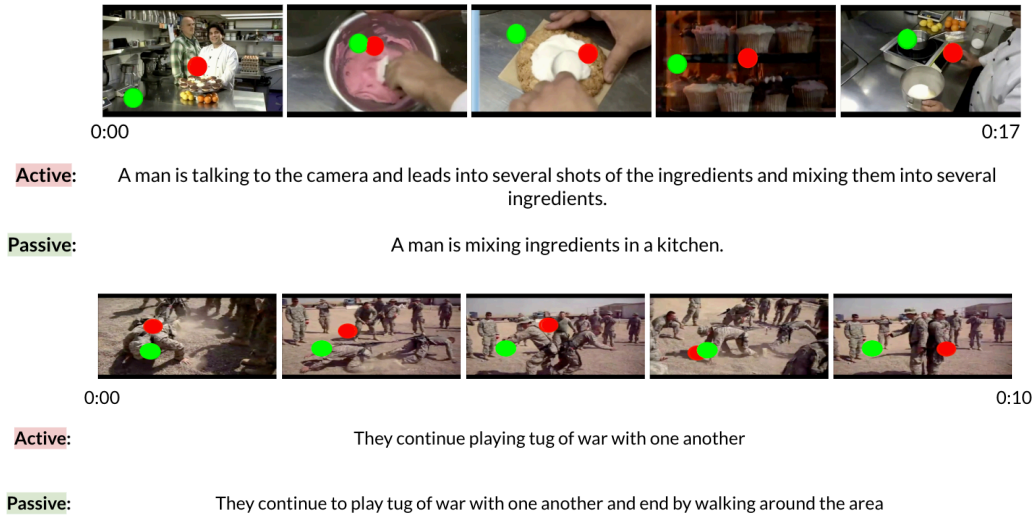

Fig. 5.4: Some examples of captions generated from the gaze of active and passive viewers where the active captions are better. The red dot represents the gaze position of the active viewer and the green dot represent the gaze position of the passive viewer. The captions generated from Active viewer's gaze is more accurate and capture more details of the video.

Fig. 5.5: Examples of caption generated from the gaze of active and passive viewers where both active and passive capture different parts of the video and are partially correct. The red dot represents the gaze position of the active viewer and the green dot represent the gaze position of the passive viewer.

5.2 Comparison of captions on EPIC-KITCHENS

In this section, we will discuss the performance of our proposed dense video captioning model on EPIC-Kitchens [6]. Figure 5.6 shows an example of a short

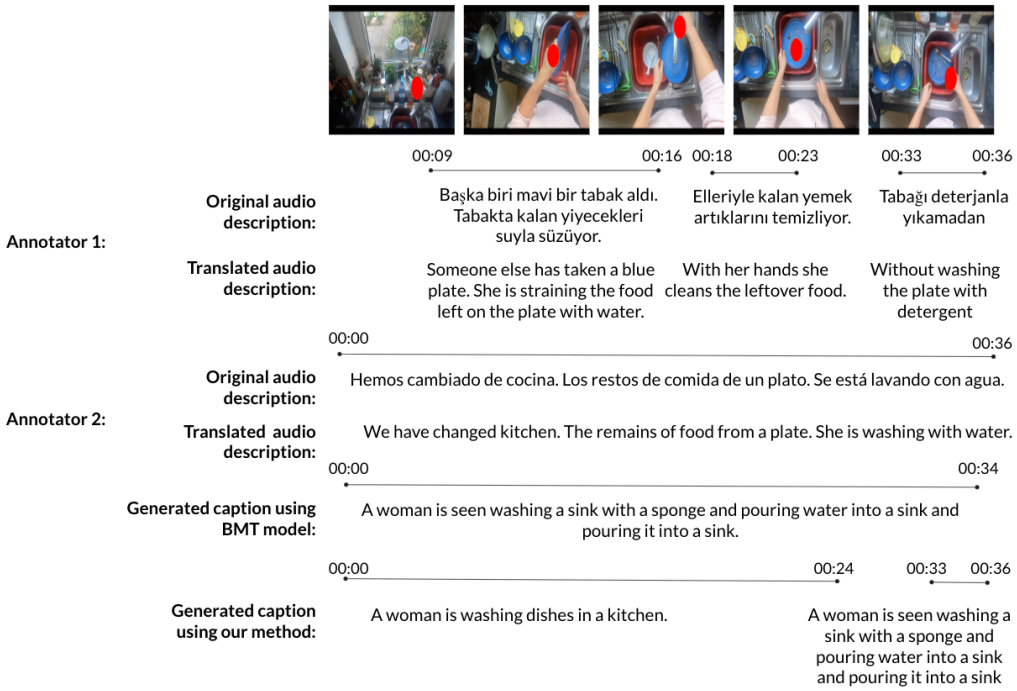

Fig. 5.6: Example of the collected data with eye gaze points marked in red on the video frames, original audio description by 2 annotator in their native languages (Annotator 1 in Turkish and Annotator 2 in Spanish(Spain)), translated audio description using DeepL language translator, generated captions using the BMT model, and generated captions using our proposed method.

video clip with the eye gaze positions of the collected data and shows the different natural language descriptions collected followed by some of the captions generated by the original BMT model [1] and our proposed modified model.

5.2.1 Numerical metrics

In this subsection, we will highlight the numerical comparison of the generated captions from different methods using gaze modality and compare them with the natural language descriptions collected during our data collection. We perform the comparison with the natural language description because there is no groundtruth captions available for EPIC-Kitchens [6] yet.

Table. 5.5: Comparison of captions generated from different gaze encoding methods. The top 2 values are highlighted for emphasis.

Method	BLEU@1	BLEU@2	METEOR	ROUGE_L	CIDEr
BMT	2.89	0.84	2.38	1.81	1.37
BMT + Mask	**3.61**	1.02	**2.81**	2.45	1.68
BMT + GM140	3.35	**1.06**	2.72	2.40	**1.73**
BMT + GM70	3.46	0.94	2.70	2.42	1.40
BMT + GM35	3.31	0.85	2.51	**2.47**	1.38
BMT + Fix140	**3.56**	**1.08**	2.75	2.37	1.68
BMT + Fix70	3.47	0.92	**2.76**	2.36	**1.71**
BMT + Fix35	3.14	0.82	2.32	**2.46**	1.63

Table 5.5 shows the BLEU@1-2, METEOR, ROUGE_L and CIDEr scores for the captions generated from different gaze encoding methods. The best performance is the BMT + Mask method followed by BMT + Fix140 method. All the gaze encoded methods outperformed the baseline BMT method for our dataset consisting of EPIC-Kitchens video.

5.2.2 Qualitative comparison

In this subsection, we will qualitatively compare the captions generated from different gaze encoding methods. Figure 5.7 shows an example of captions generated from different gaze encoding methods. The gaze position on the corresponding video frame is shown in red. The mistakes in the captions are shown in red font. The GM35 method produced the most incorrect caption of all. The other methods caught the general information of the video and produced somewhat correct captions. In the case of the Mask method, it generated the most detailed caption but made the mistake of identifying the gender of the person in the video. The main reason behind it could be the fact that the person was never fully visible in the video. The nuances captured by different gaze encoding methods as shown in

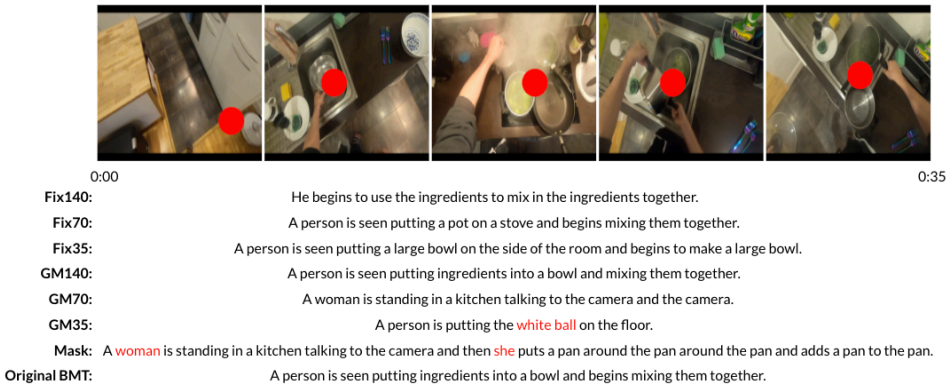

Fig. 5.7: Example of captions generated from different gaze encoding methods on EPIC-Kitchens video. The red dots on the video frame correspond to the eye gaze position. The incorrect words in the captions are highlighted in red.

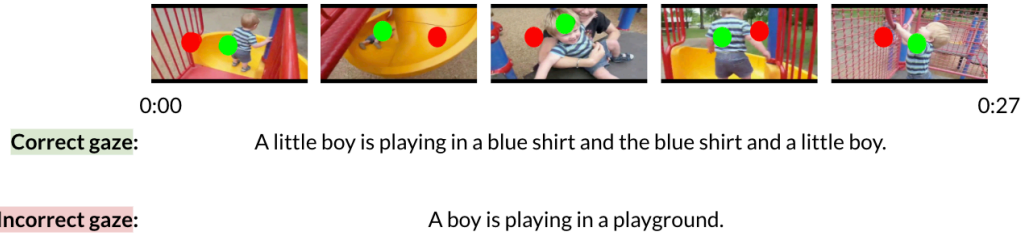

Fig. 5.8: This example illustrates the difference in captions generated using correct and incorrect gaze positions. The green circle corresponds to the correct gaze and the red circle corresponds to the incorrect gaze positions on the video. As you can observe the caption generated with correct gaze is much more descriptive in nature.

the figure are what makes it difficult to quantitatively compare one caption against another.

Table. 5.6: Comparison of captions generated from correct and incorrect eye gaze data. The larger values are highlighted for emphasis.

BMT+Fix70	BLEU@1	BLEU@2	BLEU@3	METEOR	ROUGE_L	CIDEr
Correct gaze	**9.88**	**3.94**	**1.86**	**5.67**	**9.86**	**12.35**
Incorrect gaze	7.53	2.75	1.44	4.44	7.30	10.00

5.3 Ablation study: Verification of the usefulness of gaze modality

In this section, we will highlight the importance of using the additional modality of gaze for dense video captioning task. In order to achieve this, we have taken a subset of our dataset which includes ActivityNet videos with a total length of 30 mins and used incorrect gaze position to generate captions. Our aim is to show that the captions generated by using incorrect gaze are worse than the captions generated by using correct gaze using both numerical metrics and qualitative analysis.

Figure 5.8 shows an example of the captions generated using the correct and incorrect gaze data. As you can observe, the incorrect gaze position (marked with a red dot) is usually not near the subject of the video. On the other hand, the correct gaze position (marked by the green dot) is focused on the subject of the video, which is the little boy. This difference in areas of attention is well reflected in the generated captions. The caption generated using incorrect gaze can get the gist of the video, but it fails to notice small details like the fact that the boy is little and that he is wearing a blue shirt. These details are captured by the captions generated using the correct gaze.

Table 5.6 shows the numerical comparison of the captions generated from the correct and incorrect gaze using our proposed model with fixations and the parameter $\lambda = 70$. The values clearly reflect that the captions generated from correct gaze have a larger overlap with the groundtruth of the ActivityNet dataset and are hence more accurate.

Chapter 6

Conclusion

In this thesis, we have proposed a novel multi-modal video dataset that consists of video, eye gaze, and spoken language description. Our dataset is 2.5 hours long with videos of equal length selected from ActivityNet [3] and EPIC Kitchens [6]. They were annotated by 33 participants with audio description in 15 different languages. We have highlighted the multitude of tasks such as dense captioning, visual grounding, video retrieval, and navigation that can be improved with the help of our dataset.

We have introduced a new task called: spatially and temporally dense video captioning. The output of this task are captions that can be localized on the video in time and space. In simpler terms, we can say that each caption belongs to a specific time interval and corresponds to a particular pixel region in the video frames. Performance evaluation of this task is a future work. We have focused on implementing gaze modality in the task of dense video captioning to investigate the importance of additional modality in video understanding. We have discussed the different ways to encode gaze in the dense video captioning task. Our proposed gaze encoding can be broadly divided into hard attention map and soft attention map. In a hard attention map, the regions are selected in a binary manner by only including regions near eye gaze. Whereas in a soft attention map, the regions are given weights between 0-1 by giving higher values to regions near eye gaze. Within a soft attention map, we also incorporated eye fixations to selectively choose regions where the annotator's fixation was observed.

We have compared our generated captions with the available groundtruth and the collected audio transcription. There was an improvement in captions observed for EPIC Kitchens videos by using gaze but not for ActivityNet videos. We believe that this can be rectified by retraining the model appropriately. For ActivityNet among the different gaze encoding methods, BMT+GM140 and BMT+Mask showed top performance. For EPIC Kitchens, BMT+Fix140 and BMT+Mask showed top performance. In our study between active and passive viewers, we have highlighted the difference in their attention regions and concluded that eye gaze data collected from active viewers is much more richer. Captions generated from active viewer's data exhibited higher scores. We also did a separate study in which participants had to rate these captions based on the accuracy of the caption and the correctness of the time interval. Higher performance was observed for active captions in this study. To reinforce the importance of eye gaze data, we also generated captions using random gaze and compared them with our captions. We could establish the usefulness of eye gaze as it exhibited better performance than using random incorrect gaze.

In conclusion, we have highlighted the importance of multi-modal datasets in computer vision. We have shown the effectiveness of our one of a kind dataset in dense captioning task in various experimental settings. We have also analyzed the difference in the attention regions of active and passive viewers in a video.

Future Work

Future work can be carried out in two directions. Firstly, the proposed dense captioning model can be re-trained on our collected data in a smart way by using self-supervised learning. This will be helpful as the size of our dataset is small. We believe that retraining will improve captioning quality. The other way to make a model more adaptive to our dataset will be by using knowledge distillation. In knowledge distillation, we basically transfer knowledge from a larger model to a smaller model to fit the smaller dataset better.

The second direction in which this work can be extended is to tackle the other tasks mentioned in Table 1.1 by introducing gaze and spoken language as additional modalities. There are different ways in which these modalities can be added to these tasks. The general inputs and outputs of these models are specified in the table. It is important to understand the task carefully to be able to judge the usefulness of the additional modalities. The performance evaluation of our newly introduced task of spatially and temporally dense captioning can be the first step in this direction.

References

[1] V. Iashin and E. Rahtu, "A better use of audio-visual cues: Dense video captioning with bi-modal transformer," in *British Machine Vision Conference (BMVC)*, 2020.

[2] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[3] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[4] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 7590–7598.

[5] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[6] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision (ECCV)*, 2018.

[7] E. van Miltenburg, Á. Kádár, R. Koolen, and E. Krahmer, "Didec: The dutch image description and eye-tracking corpus," in *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2018, pp. 3658–3669.

[8] 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 *Computer Vision – ECCV 2014*. Springer International Publishing, 2014, pp. 740–755.

[9] E. van Miltenburg, R. Koolen, and E. Krahmer, "Varying image description tasks: spoken versus written descriptions," in *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. Association for Computational Linguistics, 2018, pp. 88–100.

[10] P. Vaidyanathan, E. T. Prud'hommeaux, J. B. Pelz, and C. O. Alm, "SNAG: Spoken narratives and gaze dataset," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 132–137.

[11] P. Vaidyanathan, E. Prud'hommeaux, J. B. Pelz, C. O. Alm, and A. R. Haake, "Fusing eye movements and observer narratives for expert-driven image-region annotations," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research amp; Applications*, ser. ETRA '16, 2016, p. 27–34.

[12] S. He, H. R. Tavakoli, A. Borji, and N. Pugeault, "Human attention in image captioning: Dataset and analysis," 2019.

[13] R. Vedantam, C. L. Zitnick, and D. Parikh, "Collecting image description datasets using crowdsourcing." ArXiv, 2014.

[14] E. Takmaz, S. Pezzelle, L. Beinborn, and R. Fernández, "Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4664–4677.

[15] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[16] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8299–8308, 2019.

[17] P. Vaidyanathan, E. Prud ' hommeaux, C. O. Alm, and J. B. Pelz, "Computational framework for fusing eye movements and spoken narratives for image annotation," *Journal of Vision*, vol. 20, no. 7, pp. 13–13, 2020.

[18] S. He, H. R. Tavakoli, A. Borji, and N. Pugeault, "Human attention in image captioning: Dataset and analysis," 2019.

[19] Y. Shen, B. Ni, Z. Li, and N. Zhuang, "Egocentric activity prediction via event modulated attention," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[20] Y. Huang, M. Cai, Z. Li, F. Lu, and Y. Sato, "Mutual context network for jointly estimating egocentric gaze and action," *IEEE Transactions on Image Processing*, vol. 29, pp. 7795–7806, 2020.

[21] G. Boccignone, V. Cuculo, A. D'Amelio, G. Grossi, and R. Lanzarotti, "On gaze deployment to audio-visual cues of social interactions," *IEEE Access*, pp. 1–25, 2020.

[22] M. Assens, X. G. i Nieto, K. McGuinness, and N. E. O'Connor, "Pathgan: Visual scanpath prediction with generative adversarial networks," 2018.

[23] E. Takmaz, S. Pezzelle, L. Beinborn, and R. Fernández, "Generating image descriptions via sequential cross-modal alignment guided by human gaze," in *EMNLP (1)*, 2020, pp. 4664–4677.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Rep-*

resentations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[25] L. Zhou, Y. Zhou, J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," 2018, pp. 8739–8748.

[26] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Jointly localizing and describing events for dense video captioning," 2018, pp. 7492–7500.

[27] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, "Streamlined dense video captioning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2019, pp. 6581–6590.

[28] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517–6525.

[29] R. Joseph and F. Ali, "Yolov3: An incremental improvement." ArXiv, 2018.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection." ArXiv, 2015.

[31] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1532–1543.

[32] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn to pay attention," *ArXiv*, vol. abs/1804.02391, 2018.

[33] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos." ArXiv, 2017.

[34] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," 10 2002.

[35] 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. Association for Computational Linguistics, 2005, pp. 65–72.

[36] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out.* Association for Computational Linguistics, 2004, pp. 74–81.

[37] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.