DENSE VIDEO CAPTIONING BY UTILIZING AUXILIARY IMAGE DATA

YARDIMCI RESİM VERİLERİNİ KULLANARAK DETAYLI VİDEO ALTYAZILAMA

Emre BORAN

ASSOC. PROF. Nazlı İKİZLER CİNBİŞ

Supervisor

ASSOC.PROF.DR. Aykut ERDEM

Co-Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfillment to the Requirements

for the Award of the Degree of Master of Science

in Computer Engineering

2020

ÖZET

YARDIMCI RESİM VERİLERİNİ KULLANARAK DETAYLI VİDEO ALTYAZILAMA

Emre BORAN

Yüksek Lisans, Bilgisayar Mühendisliği
Danışman: Doç. Dr. Nazlı İKİZLER CİNBİŞ
İkinci Danışman: Doç. Dr. Aykut ERDEM
Temmuz 2020, 71 sayfa

Detaylı video altyazılama, uzun videolardaki olayları tespit etmek ve tespit edilen her olay için doğru ve tutarlı altyazı oluşturulmasını amaçlamaktadır. Altyazılar oluşturulurken, olaylar arasındaki zamansal bağımlılıklar ve olayların sıralamasının dikkate alınması ve bu kapsamda anlamlı ve akıcı bir paragraf oluşturulması gerektiğinden en zorlu altyazılama görevlerinden biridir ve önceki çalışmaların çoğu büyük ölçüde videolardan elde edilen özniteliklere bağımlıdır. Videoda yer alan her bir olayın ayrı ayrı altyazılanması ve uzun, tanımlayıcı bir paragraf oluşturulması gerektiğinden, metinsel altyazıların oluşturulması, yoğun video altyazılama görevi için oldukça zor bir iştir. Bu tezde, bu ağır yükü hafifletmenin bir yolunu arıyoruz ve bir videoda yer alan olaylar için uyumlu altyazılar oluştururken, yardımcı veri kaynağı olarak, videolara benzer resimlerin altyazılarından yararlanan, yeni bir detaylı video altyazılama yaklaşımı önerilmektedir. Önerilen model, görsel olarak benzer resimleri başarılı bir şekilde bulmakta ve videolara benzer nitelikteki resimlerin altyazılarında yer alan isim ve fiil tamlamalarını başarıyla kullanmaktadır. Yaratıcı ve seçici olarak adlandırabilecek

bir dizayn ve dikkat mekanizması tabanlı birleştirme tekniği ile resim altyazılarının, yoğun video altyazılama sürecinde dahil edilmesi sağlanmaktadır. Bir olay için en iyi üretilmiş altyazı, olaylar arasındaki zamansal ve anlamsal bağlantıları dikkate alan bir seçici tarafından seçilmektedir. Önerdiğimiz modelin başarımı, detaylı video altyazılama için önerilen ActivityNet Captions veri kümesi üzerinde gösterilmiş ve yaklaşımımız güçlü bir temel model ile kıyaslandığında otomatik metrikler ve nitel değerlendirmelerine göre daha iyi sonuçlar vermektedir.

Anahtar Kelimeler: Detaylı Video Altyazılama, Görüntü Açıklamaları, Yardımcı Veri, Bağlamsal Bilgi.

ABSTRACT

DENSE VIDEO CAPTIONING BY UTILIZING AUXILIARY IMAGE DATA

Emre BORAN

Master of Science, Computer Engineering Department
Supervisor: Assoc. Prof. Nazlı İKİZLER CİNBİŞ
Co-Supervisor: Assoc. Prof. Dr. Aykut ERDEM
July 2020, 71 pages

Dense video captioning aims at detecting events in untrimmed videos and generating accurate and coherent caption for each detected event. It is one of the most challenging captioning tasks since generated sentences must form a meaningful and fluent paragraph by considering temporal dependencies and the order between the events, where most of the previous works are heavily dependent on the visual features extracted from the videos. Collecting textual descriptions is an especially costly task for dense video captioning, since each event in the video needs to be annotated separately and a long descriptive paragraph needs to be provided. In this thesis, we investigate a way to mitigate this heavy burden and we propose a new dense video captioning approach that leverages captions of similar images as auxiliary context while generating coherent captions for events in a video. Our model successfully retrieves visually relevant images and combines noun and verb phrases from their captions to generating coherent descriptions. We employ a generator and a discriminator design, together with an attention-based fusion technique, to incorporate image captions as context

in the video caption generation process. We choose the best generated caption by a hybrid discriminator that can consider temporal and semantic dependencies between events. The effectiveness of our model is demonstrated on ActivityNet Captions dataset and our proposed approach achieves favorable performance when compared to the strong baseline based on automatic metrics and qualitative evaluations.

Keywords: Dense Video Captioning, Image Descriptions, Auxiliary Data, Contextual Information.

ACKNOWLEDGEMENTS

E. Boran received support from TUBITAK through 2210-A fellowship program. This work has support from the MMVC project funded by TUBITAK and the British Council via the Newton Fund Institutional Links grant programme (grant ID 219E054 and 352343575).

First and foremost, I would like to express my sincere gratitude to my advisors Assoc. Prof. Nazlı İKİZLER CİNBİŞ, Assoc. Prof. Aykut ERDEM, and Assoc. Prof. Erkut ERDEM who have always encouraged me and guided me with their valuable contributions and criticisms at all stages of my dissertation.

Besides I would like to thank my thesis committee members for insightful comments for this thesis.

Finally, I thank my beloved wife for all her love, support and patience. I would like to thank my daughter for being our miracle. I would also want to thank my family for their continued support throughout my educational life. I would like to thank my dear friend Hakan ERTEN for his help and support.

This work is supported by the Scientific and Technological Research Council of Turkey (TUBITAK) through 2210-A fellowship program.

CONTENTS

TABLES

FIGURES

1. INTRODUCTION

Video captioning can be described as the process of automatically generating natural language sentences for describing the content in a video [3]. A video captioning sample is seen in Figure-1.1.

In an ideal video captioning setup, objects, actions, scenes and the interactions between people, objects, actions and scenes must be recognized [14]. Following this recognition process, their time of arrival and temporal order must be learned. In the case of more than one event in the video, the interrelation between events must be considered as well. These steps should lead to generation of grammatically correct, coherent, visually related and human-understandable captions.

Although there has been significant interest in this topic with the emergence of new datasets [10, 15, 16] and techniques [10, 17–19], it remains a very challenging problem. Lack of diversity, redundancies and semantic inconsistencies in generated captions are the main issues. To overcome these issues, the majority of previous work [8, 18–23] has focused on generating captions for events in videos only based on visual cues. In [24] audio, in [25] speech, and in [26] both audio and speech are utilized along with visual cues as multi-modal information.

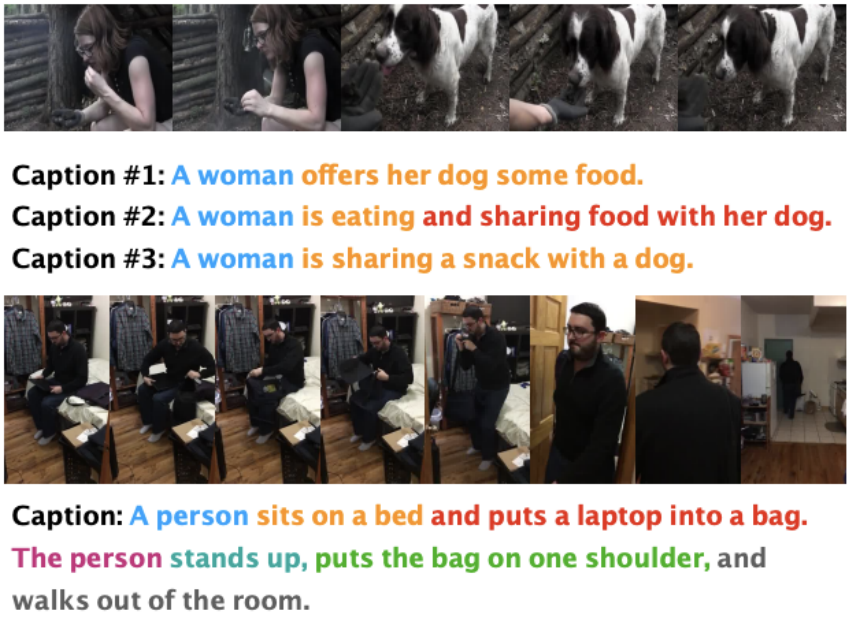

Figure 1.1. Two examples for video captioning from MSR-VTT dataset [2]. (Image is taken from [3].)

In video captioning, there are numerous major challenges [27]. First is the difficulty of generating fine-grained natural descriptions. While generating natural language sentences for event clips in a video, interactions between objects can be occluded or cannot be visible. There may be scene changes between events or reappearance of objects seen in different parts of videos. Due to the fact generation task is somehow dependent on visual information, these missing information and changes can drive the generation process to unrelated or unsatisfactory captions.

Second major challenge can be named as difficulty in learning intermediate representations between visual and text domains. Another main major challenge may be exposure to excessive amounts of objects, interactions of these objects, different activity categories, scenes, etc. in a video. Although we may access and learn these visual features from video data, ranking them according to their importance can be a difficulty while recounting visual contents.

Although these challenges may make video captioning task much more difficult, with the high-speed computing capacity of GPUs and better performance of deep learning methods, interest in video captioning in the community has increased. Generating coherent descriptions of videos for impaired people who have difficulty seeing or having natural language descriptions for video surveillance records are some of the important applications of this task.

To generate diverse, coherent and grammatically acceptable captions that are aligned with the visual context in a given video, different types of input combinations for models have been proposed. But as far as we know no other work has considered auxiliary textual information from image caption datasets in a pipeline as proposed in this thesis.

In this thesis, a dense video captioning approach that uses *image captions as auxiliary input* alongside video information is proposed. In this way, we aim to benefit from the additional diversity and richness of the image captions in generating more coherent descriptions for the videos.

Our proposed framework is inspired from the generator and hybrid discriminator design of [1], differs in using an existing image captioning dataset to collect useful information based on similarity of images and events. Our proposal exploits an attention mechanism to incorporate this auxiliary information with visual information.

Our main motivation is using captions from an already existing image captioning dataset to enhance the caption generation process of a selected video captioning model. Although there are not enough datasets when compared with image captioning, a novel way for fusion of

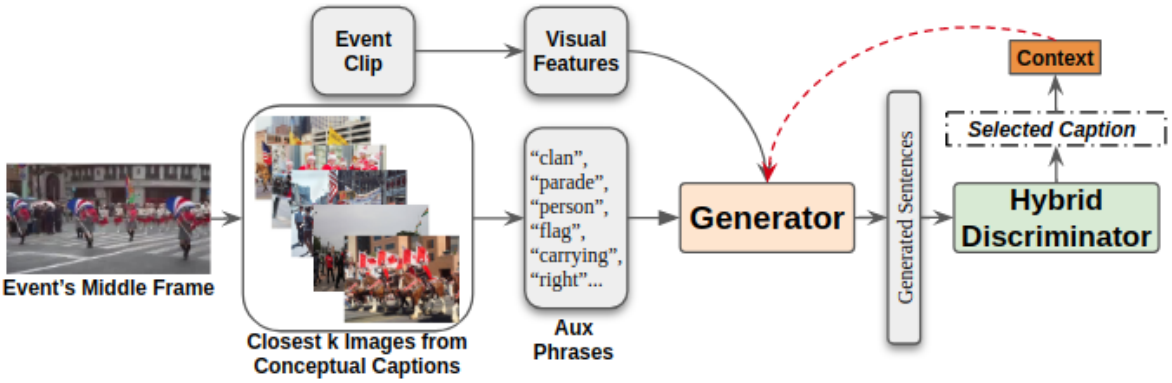

Figure 1.2. Summary for our proposed dense video captioning approach. (Best viewed in color.)

textual information gathered from an image caption dataset can be leveraged for generating more coherent captions for events in a video.

Figure 1.2. illustrates our overall approach. At training time, our model first fetches similar images to videos by comparing their visual features from a large image-captioning dataset. We then compare the corresponding closest "k" image captions against ground truth event captions, re-order them based on their sentence-level similarity. We extract the respective noun and verb phrases and use an attention mechanism to attentively select the useful parts of these auxiliary caption data. The generator module exploits this new enhanced input and uses it to generate novel captions that are more descriptive of the content in the video. Finally, a hybrid discriminator selects the optimal captions for events.

We evaluate our proposed model on ActivityNet Captions benchmark [13], and demonstrate that our proposed framework achieves qualitatively better captioning performance than [1]. Our results show that the proposed method utilizes auxiliary image captions effectively as additional context and combines them with visual information of videos.

1.1. Major Contributions of the Thesis

This thesis contribute towards a solution in generating coherent, meaningful captions for events in a video. In summary, our main contributions in this thesis are:

- We propose an auxiliary data enhancement pipeline that enables extracting meaningful phrases as auxiliary information to be used in caption generation process,
- We explore an attention mechanism to incorporate this auxiliary information with visual information,

- Our model performs comparably to a strong baseline when evaluated using automated metrics. We qualitatively show significant improvements with the generated video captions,
- Our pipeline is quite generic and our proposed pipeline and attention mechanism can be used in other dense video captioning models to enhance input diversity.

In this work, we propose using extracted noun and verb phrases as auxiliary input alongside visual information from videos. To lessen dependency over only visual information from videos, we propose utilizing usage of extracted phrases as lexical input to caption generation process in generator.

1.2. Structure of the Study

Overall structure for this thesis is as follows:
Chapter 2 presents the essential background knowledge. It starts off by presenting basic principles of Generative Adversarial Networks, Recurrent Neural Networks and provides an overview of Long-Short Term Memory. Then it briefly summarizes the Word Mover's Distance.
Chapter 3 summarizes main previous works on captioning tasks. We divide the related works to categories as: Image Captioning, Video Captioning, Dense Video Captioning and Image/Video Captioning models using Generative Adversarial Networks. This chapter also presents other prominent studies on the topic, ending with a discussion on the contributions of our work as compared other studies.
Chapter 4 describes the proposed pipeline for fetching relevant images and our model for incorporating auxiliary captions with adversarial inference. In particular, details of the proposed generator, the base hybrid discriminator and proposed hybrid discriminator architectures are given.
Chapter 5 introduces video and image captioning datasets used in the study. automated metrics and diversity metrics used for the evaluation of our proposed model are then defined. Visual attributes used in this study and our training and evaluation details for models are presented. Finally, the experimental results are discussed.
Chapter 6 includes a summary for our work in this thesis. At last, research directions for future work are detailed.

2. BACKGROUND

In this chapter, background knowledge which will help to follow the approaches and models represented at the rest of this study is introduced. First of all, basic information about Generative Adversarial Networks is given in Section 2.1.. Then, in Section 2.2., basic Recurrent Neural Networks are explained; and Long-Short Term Memory is introduced briefly in Section 2.3.. Finally, in Section 2.4., Word Mover's Distance is explained in detail.

2.1. Generative Adversarial Network

Generative Adversarial Nets (GANs) are proposed by [28] as a new kind of deep generative networks. GANs have been successfully applied to various different problems (e.g. generating realistic human faces [29], image-to-image translation [30], style transfer [31], or to our interest, generating captions for events in videos [1]). They use a game-theoretic framework in that generative modeling is formulated as a game between a generator network and a discriminator network. The generator network is specifically designed for the task at hand whereas the discriminator network scores how realistic are the samples generated by the generator network. For example, when applied to text generation, the generator learns to generate novel sentences, which are evaluated by the discriminator which learns to distinguish natural sentences from fake ones.
In more detail, the GAN framework involves two interconnected tasks, which are illustrated in Figure 2.1.:

- The generator learns to generate truthful examples, which are used as negative training data in training of the discriminator network. At the beginning of the training the quality of these samples is low.
- The discriminator network learns to distinguish fake data generated by the generator from ground truth data. In fact, at this stage, the discriminator gives feedback to generator for implausible generated data.

2.1.1. The Discriminator

The discriminator network serves basically as a classifier that tries to appoint real data and fake data from generator to two different classes. Hence, the network architecture of discriminator is based on data to be classified.

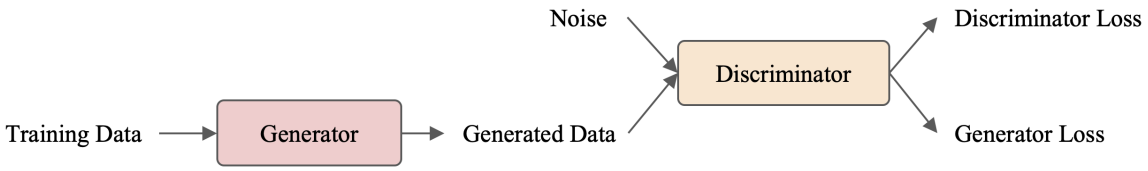

Figure 2.1. Basic Generative Adversarial Network design.

The discriminator and the generator losses are combined in the full training objective. However, discriminator training uses only discriminator loss to update its weights through back-propagation. Hence, generator loss is not directly used in discriminator training.
The discriminator:

- learns to separate output of generator as fake input and real input,
- updates its weight as stated with back-propagation [32] from only discriminator loss,
- penalizes learned weights through discriminator loss emerging from misclassified real input and fake input instead of one another.

2.1.2. The Generator

The other component of GANs, the generator network tries to produce fake input to discriminator without being realized by the discriminator [28]. To this end, generator learns how to generate realistic fake data by the supervisory signal from the discriminator's feedback. The main aim of the generator is fooling the discriminator to classify its outputs as real. This feedback mechanism forces a tight alliance between the discriminator and generator training. The generator network of a GAN needs the following components:

- a random noise vector as input,
- a generator network that learns to generate output as training data,
- a discriminator network to assess its generation,
- a generation loss to update its weights through back-propagation to learn how to fool discriminator with fake output.

One of the main advantages of Generative Adversarial Networks is that the update of generator is not based on the training samples it has access to, but with the feedback from the

discriminator. Moreover, it provides a generic framework so that there is a larger variety of functions that can be integrated into the GAN architecture [28].
Discriminator is trained with [28] the objective of maximizing the probability for setting true labels to training examples and generated samples of G as follows:

$$min_G max_D V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z)))] \quad (1)$$

$V(G,D)$ is value function, $p_{data}(x)$ is probability distribution over data x, $p_z(z)$ is probability distribution over noise z, $D(x)$ is probability that x is from data. While D is maximized, G is simultaneously trained to minimize $log(1-D(G(z)))$.

2.2. Recurrent Neural Networks

Recurrent neural networks (RNNs) are neural networks which are tailor fit to process sequential data. An RNN network has hidden state(s) and allow previous outputs to be used as inputs along with these hidden states, which allows for processing inputs with any length. Hence, while processing inputs, an RNN has a history based on its hidden states.
As mentioned before, compared to conventional artificial neural networks that process each input separately, RNNs have shared weights in processing an input based on previous inputs. This sequential processing allows model to learn context of input data. Although these are clear advantages, there are some drawbacks of RNNs as well. Computations in RNN are slow, and although RNN has hidden state to carry historical information, difficulty of accessing information from a long time ago persists. Vanishing or exploiting gradients are the main reason behind difficulty of capturing long term dependencies. Gradients can exponentially decrease or increase with respect to the number of layers. The architecture of a vanilla RNN is shown in Figure-2.2.. Commonly used activation functions in RNN are $sigmoid$, $tanh$, and $RELU$. In Eq.2, on the left sigmoid, in the middle tanh and on the right, RELU activation functions are shown.

$$g(x) = \frac{1}{1+e^{-z}}, \qquad g(x) = \frac{e^z - e^{-z}}{e^z + e^{-z}}, \qquad g(x) = max(0,z). \quad (2)$$

h^{t-1} is hidden state at timestamp $t-1$, h^t is hidden state at timestamp t, x^t is the input for RNN.

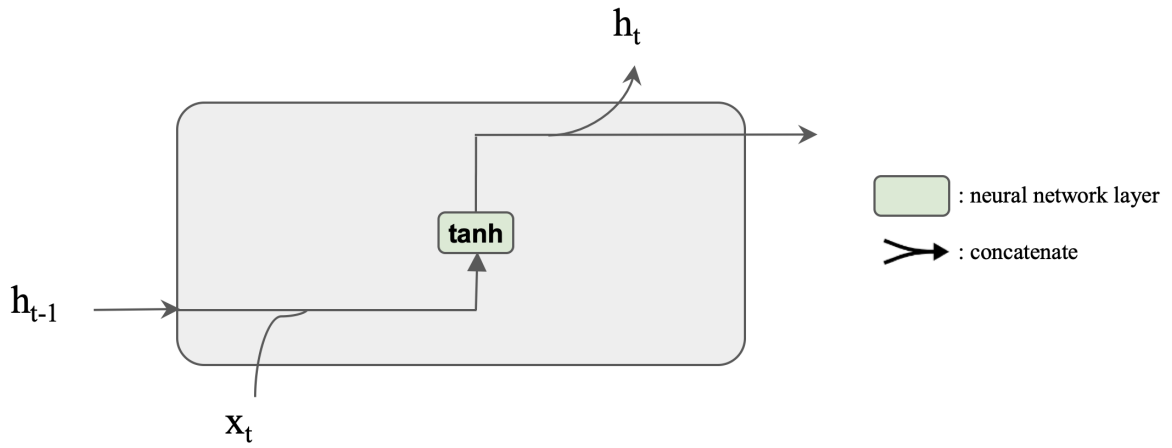

Figure 2.2. A repeating module in a basic Recurrent Neural Network.

2.3. Long-Short Term Memory

Long-Short Term Memory (LSTM) is specialized RNN and proposed by [33]. Although RNNs propose to handle long-term dependencies in long sequences, [34] show they are not practically efficient at remembering and considering more than a few previous timestamps due to incapability of basic RNN cell to memorize long sequences. Main difference between a basic repeating module in RNN and LSTM cell is memory cells. These cells are proposed as a solution for long-term dependency problem of RNNs. They take an input sequence and process this sequence considering input at previous timestamps. Long-Short Term Memory(LSTM) is used in both generator and discriminator parts of our model. A basic LSTM repeating module is shown in Figure-2.3..
An LSTM cell can be defined with the following equations:

$$f_t = \sigma(W_f \cdot z + b_f), \tag{3}$$

$$i_t = \sigma(W_i \cdot z + b_i), \tag{4}$$

$$\bar{C}_t = tanh(W_C \cdot z + b_C), \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t * \bar{C}_t, \tag{6}$$

$$o_t = \sigma(W_o \cdot z + b_t), \tag{7}$$

$$h_t = o_t * tanh(C_t). \tag{8}$$

In this equations, t is the timestamp, x_t is the input vector, h_t is the hidden state of the cell. f_t denotes the forget gate, i_t is the input gate, o_t is the output gate, C_t is the memory cell

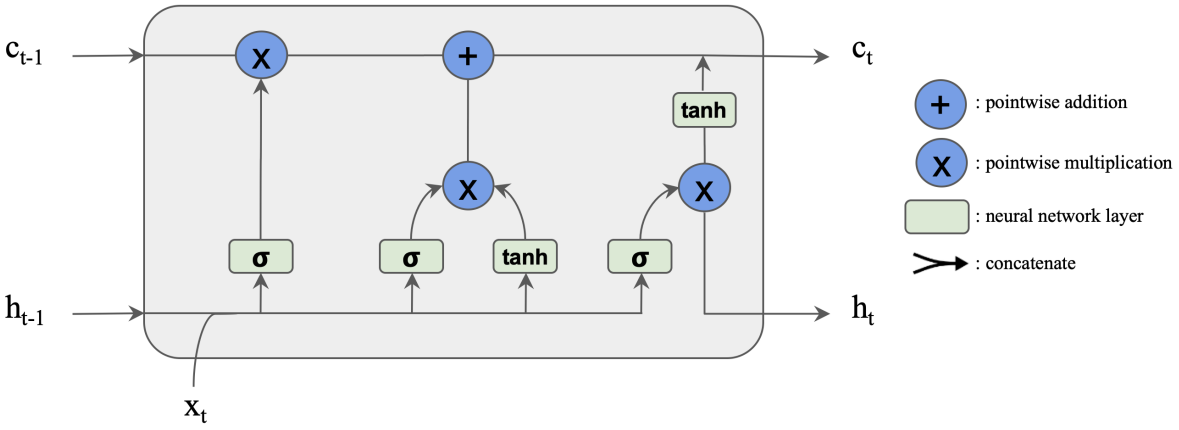

Figure 2.3. An overview of a basic LSTM cell.

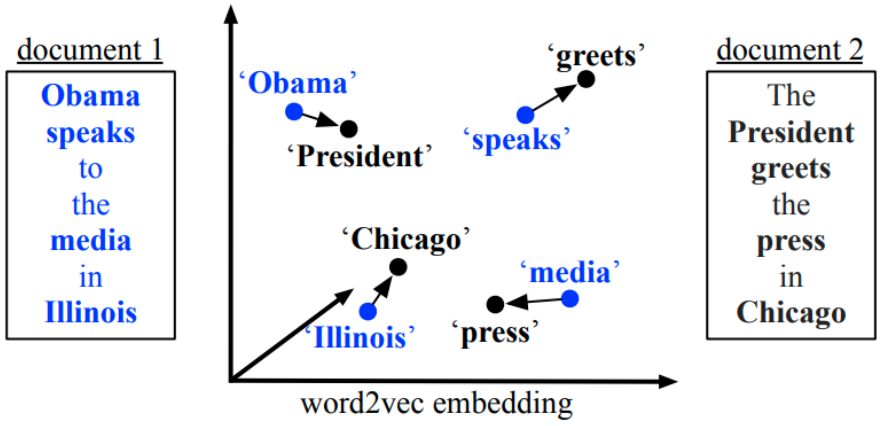

Figure 2.4. An illustration of WMD. Words of two sentences are embedded, then distance between sentences are calculated in a cumulative manner. (Image is taken from [4].)

state. $\bar{C}_t$ is candidate state value that scaled according to i_t. W is the weight matrix of for gates f, i, C, o. z is concatenated vector form of x_t and h_{t-1}, b is a bias. X is used for the element-wise product of vectors. σ stands for the sigmoid function, $tanh$ is the hyperbolic tangent function.

2.4. Word Mover's Distance

Word mover's distance (WMD) [4] measures level of dissimilarity between two documents or sentences. The idea is to first embed words of two text documents into a common semantic space (usually done by using pretrained word vectors) and the the distance between two documents is defined as the travel cost from one document to the other. An illustration of WMD is seen in Fig.-2.4.

WMD is a straight-forward technique to implement and has no hyperparameters to train. In Figure-2.4., bold words are non-stop words in document one and two. All non-stop words are embedded to word2vec[35] space. Representations for words in sentences are learned by a word2vec model designed for this task. A neural network which model that consists an input, a projection and an output layer is used for mapping words in text documents to word2vec space in WMD. In a cumulative manner, distance to travel from all word embeddings of sentences one to sentence two are added. Then this distance is used to compare all sentences and to find most similar ones.

In this section, we provided background on the models and methods used in our work. As this thesis mainly focuses on Generative Adversarial Network (GAN) based dense video captioning model, in the following sections, we will give details about our proposed idea to enhance baseline model, experimental details and results. Additionally, in this section, we also provided a brief summary of the semantic distance metric we used in our study.

3. RELATED WORK

Every day, around the world, thousands of videos and photos from various sources (e.g. personal devices, survaillence cameras, news articles, web sites) are created. Video surveillance devices, social media, video sharing platforms, and phone applications have increased the interest in taking pictures and recording videos. These images and videos are shared on different platforms and used for different purposes. For humans, time is a limited source and having reasonable summarizations for these excessive visual sources is a need emerged lately.

For this reason, image and video captioning tasks have been popular for computer vision field. Image captioning is generating syntactically and semantically correct sentences for an image by learning salient objects, scenes and their relations in that image [36]. As mentioned before, videos have been an important source in daily life, this need further expanded and led to video captioning task. Video captioning is the process of generating a natural language sentence for the action in a video automatically [3]. Furthermore, with longer videos, generating informative sentence that summarizes all events and connections between them became harder. To this end, dense video captioning task emerged [10]. Apart from video captioning, dense video captioning task aims detecting events in a video, and generating a sentence for each one of these detected events.

3.1. Image Captioning

Machines, even in some cases humans, cannot understand what is important or happening in a given image. But, by having advanced models, tasks once thought as impossible has become applicable. Image captioning works on image understanding and creating a description for a given image [36]. A sample output of an image captioning model is shown in

Figure 3.1. An example for video captioning task. (Image is taken from [5].)

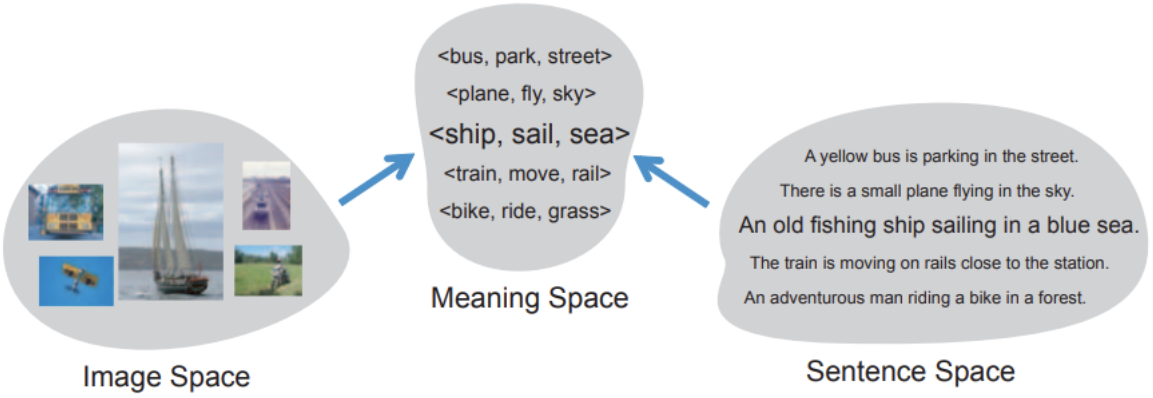

Figure 3.2. Spaces for image caption generation. (Image is taken from [6].)

Figure-3.1.. To understand an image, recognition, and detection of objects, understanding the scene they are in, interactions of them with each other must be learned by created models. At the same time, it must come up with words that can describe different aspects of an image and fuse them into a sentence or a paragraph [37]. Describing images with a single natural language sentence has recently received increasing interest. Advances in deep learning have significantly accelerated collaboration between Computer Vision (CV) and Natural Language Processing (NLP) field [27].

In Figure-3.2. a sample to combine two different projections, image and sentence spaces are shown. [6] argues to learn projections for these two spaces. By this, models can learn to caption an image with a sentence.

Image captioning heavily depends on gathering image features. Techniques can be grouped into two categories as hand-crafted features and learning frameworks, aka. shallow approaches and deep learning techniques. Techniques for hand-crafted features can be exampled as Scale-Invariant Feature Transform(SIFT) [38], Histogram of Oriented Features(HOG) [39], Local Binary Patterns(LBP) [40] which are widely used. Then, these features are fed into a classifier, i.e. Support Vector Machines(SVM), to classify a detected object. The feature extraction process of hand-crafted models is not that much applicable on large scale datasets when compared with deep learning techniques.

On the other hand, the feature extraction phase is done internally and automatically with deep learning techniques. With MSCOCO challenge [41], a new platform to evaluate results of image captioning is provided. [42] proposed a model, Natural Image Captioning (NIC), which uses a CNN as an image encoder, that is pre-trained on an image classification task and feeds the hidden layer of CNN as input to a RNN decoder that generates the captions

for a given image. [5] proposed Multi-modal Recurrent Neural Network architecture that leverages images and descriptions of them, learns connections between image sub-regions and these descriptions. These two architectures surpassed shallow approaches and reached the state-of-arts.
Other works have contributions with attentive attention images [43]. Combining CNN and RNN with LSTM to boost image captioning with attributes [44] added attention mechanism to image captioning tasks. With improvements on image captioning, video captioning task has been on the spot, lately.

3.2. Video Captioning

Due to success in object recognition on videos, a task to create a sentence that defines visual semantics of an event on a video has attracted interests recently. The main goal of video captioning is to generate a natural language description of a given short video sequence summarizing the most important actors, their actions and their interactions seen in the video [45]. A sample output of a video captioning model is shown in Figure-3.3..
Before the deep learning era, the earlier works which tackle video captioning generally generally employ template-based language generation strategies [46–48]. These methods utilize existing detectors and classifiers to separately detect subjects, verbs, and objects in the given videos and use a template to join these information to form a sentence. In particular, to improve the description coherence, [46] suggested to use a large text corpora for selecting the best subject-verb-object triplet over the detected entities and actions. [47] learned semantic relationships between subjects, verbs, and objects and used these semantic hierarchies while forming a sentence. Interestingly, [48] was the first who formulated video captioning as a machine translation problem. Specifically, they suggested to extract an intermediate semantic representation from a given video and consider it as the source while generating a natural language description by using methods borrowed from Statistical Machine Translation.
In the recent years, the developments in deep learning led to significant progress in video captioning [7, 49–54]. Specifically, [49] borrowed techniques from neural machine translation to generate video descriptions, and used recurrent neural networks to encode the visual features and then transform them into a sequence of words. Within an end-to-end learning framework, [7] proposed to additionally employ a common visual-semantic embedding space for the descriptions and the videos while training a video captioning model to improve

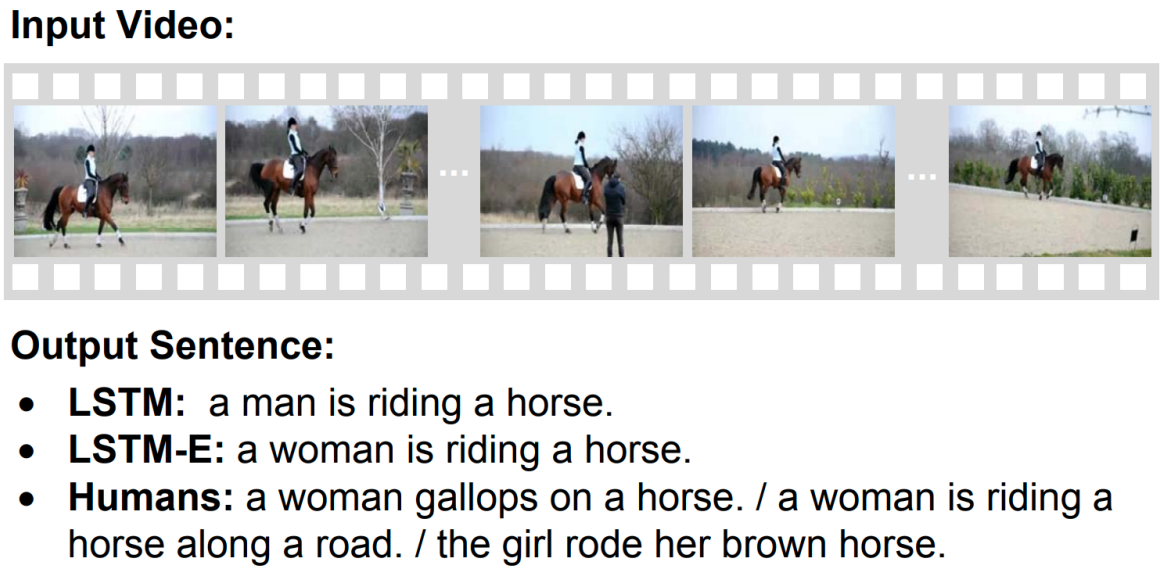

Figure 3.3. An example for video captioning task. Under the video frames, generated caption from two different video captioning models, *LSTM* and *LSTM-E* are shown. At the bottom, alternative ground truth captions for video is shown. (Image is taken from [7].)

the alignment between the visual and the textual domains. [50] suggested a temporal attention mechanism to select most important temporal segments in videos during the description generation process.

To improve the performance, [51] extended the previous models with a hierarchical recurrent neural encoder which exploits temporal structure of videos while reducing the computations and in this temporal information plays a crucial role. To better encode the visual content, [52] utilized two different video features, one depending on the objects and their attributes and the other relying on the motion and the action of the objects, that are processed by two different sub-networks where another network model evaluates the descriptions generated by these sub-networks and picks the best one. [53] proposed a joint model that employs spatio-temporal attention along with frame-level image classifiers to learn to detect subjects, verbs and objects and generates the video description accordingly. Finally, [54] suggested a captioning model named LSTM-TSA that combines the semantic attributes(TSA) extracted from frames and the video feature within an encoder-decoder architecture.

3.3. Dense Video Captioning

Dense video captioning aims to generate multiple descriptions summarizing all the events that are temporally detected in a given video. Hence, temporal localization of such events

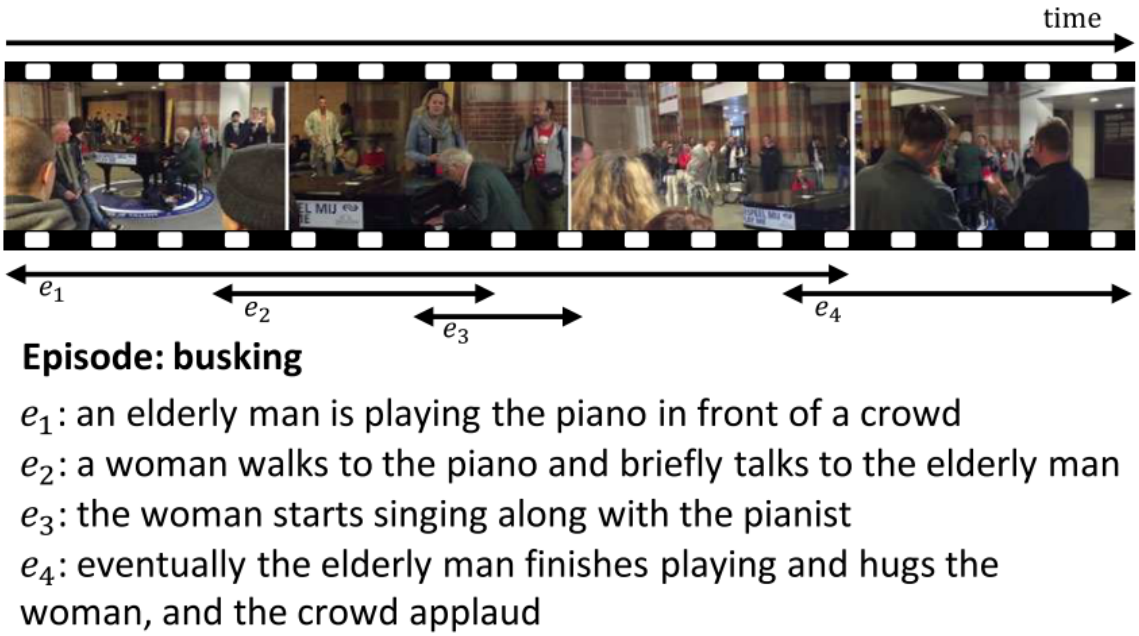

Figure 3.4. An example for dense video captioning task. e_* stands for events in videos. With overlapping events, this task becomes harder. (Image is taken from [8].)

are as important as generating a coherent set of natural language descriptions. A sample output of a dense video captioning model is shown in Figure-3.4.. For instance, [17] proposed a two-staged approach where at first simple descriptions are generated using temporal and spatial attention from every short video segments, and then these generated descriptions are combined to form a coherent paragraph. [10] suggested another dense captioning model that detects all events in a single pass and proposed a captioning module that utilizes surrounding events as context and generates captions for all detected events. [18] developed an transformer [55]-based model which jointly performs event proposal and description generation within a single framework that is trained in an end-to-end manner. [19] proposed a method which generates temporal event proposals from a given video and employs the events from both forward and backward directions as additional context information along with the current video features to generate a natural video description for a video segment. [56] developed a dense video captioning model called Memory-Augmented Re-current Transformer (MART) model that utilizes an extra memory module where the previous states and the descriptions of the events captioned before are stored in order to eliminate repetitive and incoherent descriptions. Recently, there are some other works [24–26] that utilize multiple modalities such audio or speech along with the input video while generating video descriptions.

3.4. Image/Video Captioning with Adversarial Learning.

Lately, some researchers developed both image captioning techniques [57, 58] and video captioning [1, 59, 60] models that employs Generative Adversarial Networks (GANs) [28]. Apart from the previous studies, these works mainly aim to implicitly learn the distribution of the groundtruth video captions so that when a video is given, they can generate a diverse set of descriptions that are indistinguishable from the descriptions written by humans. To do that, the training process generally involves a min-max game between two different networks, a generator network which is responsible from generating descriptions, and a discriminator network that is in control of distinguishing real descriptions from the machine generated ones.

In particular, in their image captioning framework, [57] combined adversarial training with Gumbel [61] sampling to match the distributions of descriptions by humans and machines. [58] proposed a conditional GAN model for image captioning, which is trained using policy gradient reinforcement learning.

For dense video captioning, [59] was the first that employed adversarial learning to learn a model that is capable of generating multiple descriptions from a given video, giving a precise summary of all of the events seen in that video. Their model which is called Recurrent Topic-Transition Generative Adversarial Network (RTT-GAN), in particular, uses a structured paragraph generator which forms descriptions in a recurrent manner by considering both textual and spatial attention mechanisms at each step, and multi-level paragraph discriminators that evaluate (sentence-level) plausibility and (paragraph-level) coherency of the descriptions. [1] proposed another GAN-based dense video captioning model, however, their formulation differs from that of [59] in that it applies adversarial methods during inference time. That is, the discriminator in their formulation, which consists of multiple sub-discriminator networks, is mainly responsible from assessing the quality of the descriptions sampled from the generator. Particularly, within these sub-discriminator networks, the authors considered different evaluation criteria that measure the visual relevance, the consistency and the distinctiveness of the generated descriptions.

In this thesis, we propose using image captions as auxiliary input alongside video information. We differ in this from all previous works. To lessen dependency over only visual information from videos, we propose utilizing usage of extracted phrases as lexical input to caption generation process. Furthermore, our pipeline is quite generic and extracted phrases can be used any dense video captioning model.

4. MODEL

Our main hypothesis in this thesis is that dense video captioning can benefit from auxiliary textual information from captions of visually similar images. Our model starts fetching similar images to event middle frames from ActivityNet dataset. Then features are extracted and compared. Accordingly, highest k scores are used to find most similar images. Captions for these k similar images are than used in caption generation process.

4.1. Fetching relevant images

Our framework starts with the auxiliary data retrieval, which is shown in Figure 4.1.. First, middle frames of each event in a video is used to find closest k images from the Conceptual Captions (CC) [9] dataset. This is done by computing the cosine similarity, $sim(f^l, f^v)$ between video middle frames and CC images, which is computed as:

$$sim(f^l, f^v) = \frac{f^l \cdot f^v}{\|f^l\| \ \|f^v\|} \tag{9}$$

where $sim(f^l, f^v)$ is the similarity between visual features (f^l) of the l'th image from CC corpora and f^v represents the features of the j'th event in the video. Top-k closest images corresponding to each event in a video are collected using Eq 9. An example for fetched images are shown in Figure-4.2.

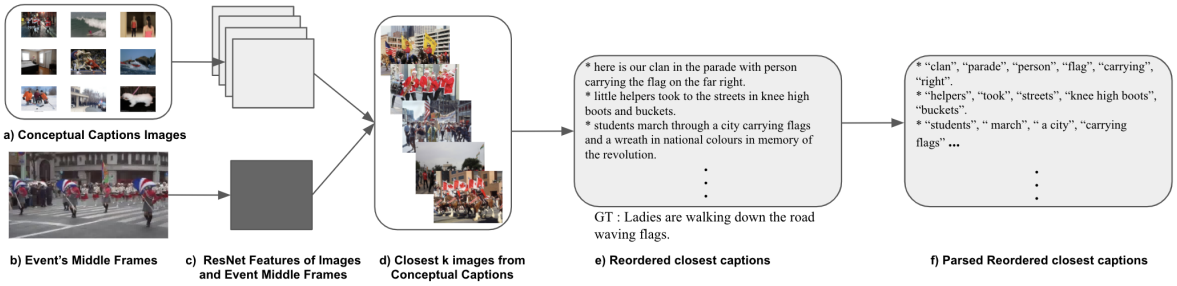

Figure 4.1. Pipeline for extracting noun and verb phrases from closest k image captions. (a),(b). Middle frames of each event are compared with images. (c) Visual features are extracted. (d) Closest k images are found. (e) Closest k captions are reordered according to their WMD [4] scores. (f) Resulting noun and verb phrases of closest k captions are used as auxiliary words in event caption generation.(Best viewed in color.)

Figure 4.2. An example for image fetching pipeline is shown. On left hand side, a middle frame of an event, and on the right, closest images and their captions from Conceptual Captions[9] dataset is seen. Our pipeline fetches images similar to event in the video.

We then reorder these images based on their caption similarities. We use Word Mover's Distance [4] (WMD) computed over non-stop words of two given sentences. In this computation, words of both sentences are first embedded into a word embedding space, then the cumulative word distance to travel from one sentence to another is evaluated. We use WMD to reorder the closest k images, according to similarity between their captions and ground truth event caption.

More formally, $WMD(c_{gt}, c_i)$ is calculated between ground-truth event caption c_{gt} and retrieved image i's caption c_i. The most similar captions to the event's ground truth caption is used as auxiliary data in the further steps. At inference, we do not use c_{gt}, we reorder according to the similarity of retrieved captions to have the most common captions at the top. We select one retrieved caption and compute similarity score between that caption and remaining captions. We sum these scores, repeat this for every caption and reorder them.

After reordering the captions, we extract noun and verb phrases from each image caption. For this, part of speech tags are combined with regular expressions and grammar based heuristics are exploited. We then extract corresponding noun and verb phrases. The resulting noun and verb phrases of a sample caption from CC corpora is shown in Figure 4.1.(f).

4.2. Using auxiliary captions with adversarial inference

We conjecture that these retrieved image captions can be incorporated into many of the video captioning models to improve the generated captions. In this thesis, we follow adversarial inference based approach, as these approaches are well-suited to incorporate the auxiliary image information.

Therefore, following the recent work of [1], we select Adversarial Inference(AdvInf) as our base model. This model uses a generator and discriminator design, where the generator and discriminator are pre-trained and updated jointly. The generator is responsible for generating captions, which will then be scored by the discriminator.

Our proposed framework improves the generator with the use of the retrieved auxiliary captions. We let the generator use *both* the captions from the input videos and textual information from the visually similar images. We use an attention based mechanism to select the auxiliary phrases and learn a no-bias option to prevent the model from using auxiliary phrases if required. Below, we describe the details of our framework.

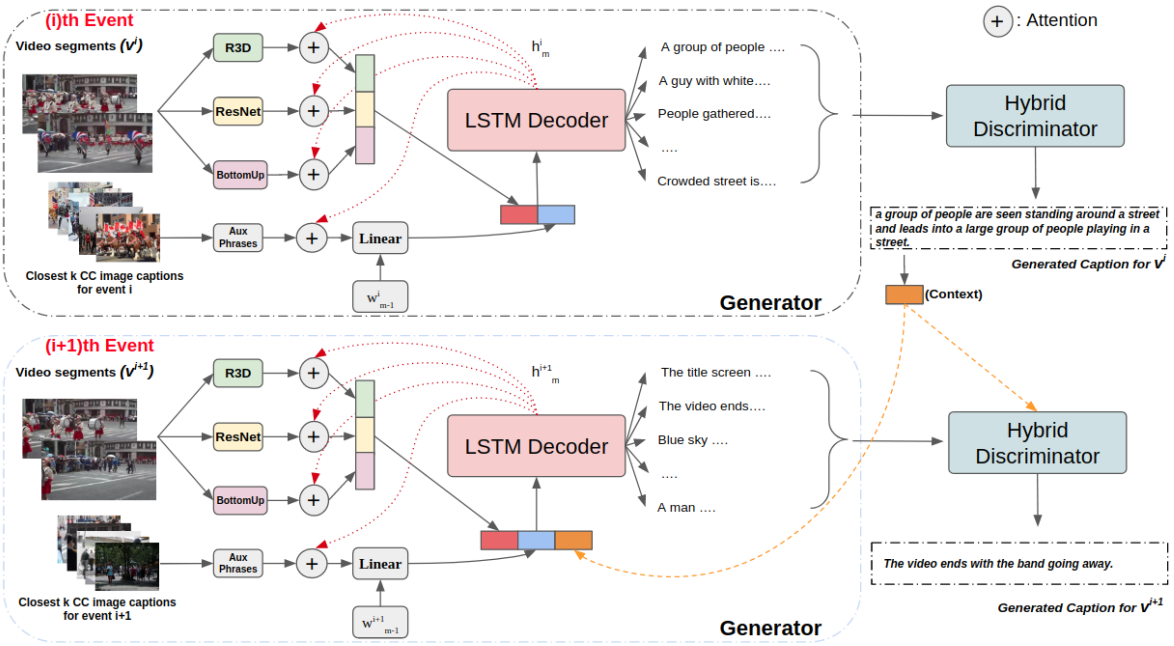

Figure 4.3. The overview of our proposed approach. Generator generates candidate captions for events in a video by using visual attributes, auxiliary phrases, previous word, w, and the previous selected sentence as input. The Hybrid Discriminator selects best captions for events among generated samples. Overall process is shown on top for ith event, and on bottom for $i+1$th event. (Best viewed in color.)

4.2.1. Generator

For a video, which has L event clips $[v^1, v^2, ..., v^L]$, generator G generates L sentences $[s^1, s^2, ..., s^L]$, where each sentence s^i is generated in accordance with corresponding event clip v^i. The generator in AdvInf is based on LSTMs [33]. It has three inputs for each timestamp; visual features (f_m^i), previous ground truth word (w_{m-1}^i), last hidden state of generated caption (h^{i-1}). Visual features consist of motion-, RGB, and object-level features concatenated to form f_m^i (Section 5.). The base LSTM based decoder of the generator is thus formulated as:

$$h_m^i = LSTM(f_m^i, w_{m-1}^i, h^{i-1}) \quad with \quad h^0 = 0. \tag{10}$$

Dense video captioning system might benefit from not only visual information but also other modalities such as textual information from an image captioning dataset. To this extend, we propose enhancing generator with phrases selected to be auxiliary context. One of the main advantages of this proposal is being able to implement this idea easily to other dense video captioning models, as well.

To make use of auxiliary image information, we extend this base generator with the caption

from the closest images. In order to have a fixed length representations for auxiliary phrase list used while generating a caption for an event, we first extract noun and verb phrases from closest caption. We extract a maximum of j phrases in total, and if the extracted number of phrases is less than j, then we zero-pad to a fixed length j. We then use the word-embedding layer of [1] to embed the extracted phrases to a common representation and and concatenate them to form the auxiliary vector f_{aux}.

In order to boost the performance of context selection, temporal attention mechanism [50] is applied over f_{aux}. Unnormalized relevance score for an auxiliary phrase from f_{aux} is calculated with respect to the previous hidden state as follows:

$$e_m^i = \psi^T tanh(W h_{m-1}^i + U f_{aux} + b) \tag{11}$$

where ψ, W, U and b are parameters estimated during the learning phase. After calculating relevance scores for all auxiliary phrases, normalized attention weights are obtained by:

$$\alpha_m^i = \frac{exp\Big\{e_m^i\Big\}}{\sum_1^j exp\Big\{e_m^i\Big\}} \tag{12}$$

These normalized attention weights are then multiplied with auxiliary vector to obtain the attentive auxiliary vector a_m^i as:

$$a_m^i = \alpha_m^i f_{aux} \tag{13}$$

We define a linear layer with inputs of previous ground truth word, w_{m-1}^i, and attentive auxiliary vector, a_m^i as follows:

$$\rho_m^i = (W_c[w_{m-1}^i, a_m^i] + b_c) \tag{14}$$

where W_c, b_c are parameters. Output of this layer is used in decoding LSTM of the proposed generator $\tilde{G}$:

$$h_m^i = LSTM(v_m^i, \rho_m^i, h^{i-1}) \quad with \quad h^0 = 0. \tag{15}$$

With attention over the auxiliary vector, our model learns whether to use auxiliary phrases or not and with Eq.14, how much information will be added to input from w_i^{m-1}, a_m^i along with visual features and previous context vector. While experimenting with attention mechanism, if there are multiple zero-padded phrases in f_{aux}, we use the attention weights of only single zero-padded element and normalize the remaining weights. This single 'zero-phrase' stands

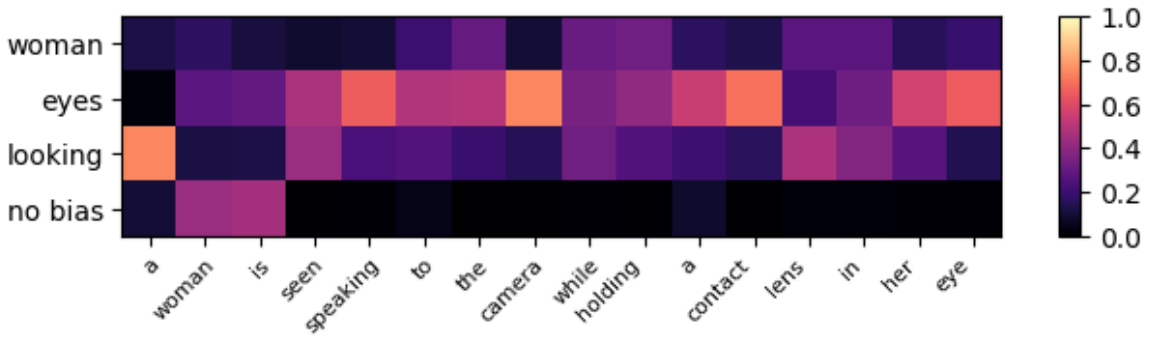

Figure 4.4. Visualized attention weights over auxiliary phrases extracted from caption of closest image. The caption of closest image was in this case: *a young woman with large eyes looking down to up*. The phrase *no bias* on y-axis means not using any of extracted auxiliary phrases. Generated caption is seen on x-axis. (Best viewed in color.)

for the *no-bias* option [62], meaning not using any of the auxiliary noun and verb phrases while generating captions.

An example of visualized attention weights over auxiliary phrases are shown in Figure 4.4.. Phrases extracted from the caption of the closest image are seen on the y-axis and the generated caption is seen on x-axis. Our model learns to back-off if no additional phrase is needed while generating next word.

4.3. Hybrid Discriminator

We use hybrid discriminator[1] for scoring all candidate generated captions choose best caption based on its visual relevance to video features, linguistic semantics and consistency with the previous caption. The hybrid discriminator has three sub-modules named as visual, language and pairwise discriminators. Overview of hybrid discriminator is shown in Figure-4.5.. Then we will give details about our proposed sub-module namely Similarity Discriminator. This discriminator is proposed to evaluate similarity between generated caption and fetched images caption.

4.3.1. Visual Discriminator

Visual Discriminator, D_V, assesses whether generated caption is aligned with event features or not. While this process, D_V does not check fluency or grammar of the generated sentence s^i for event in a given video.

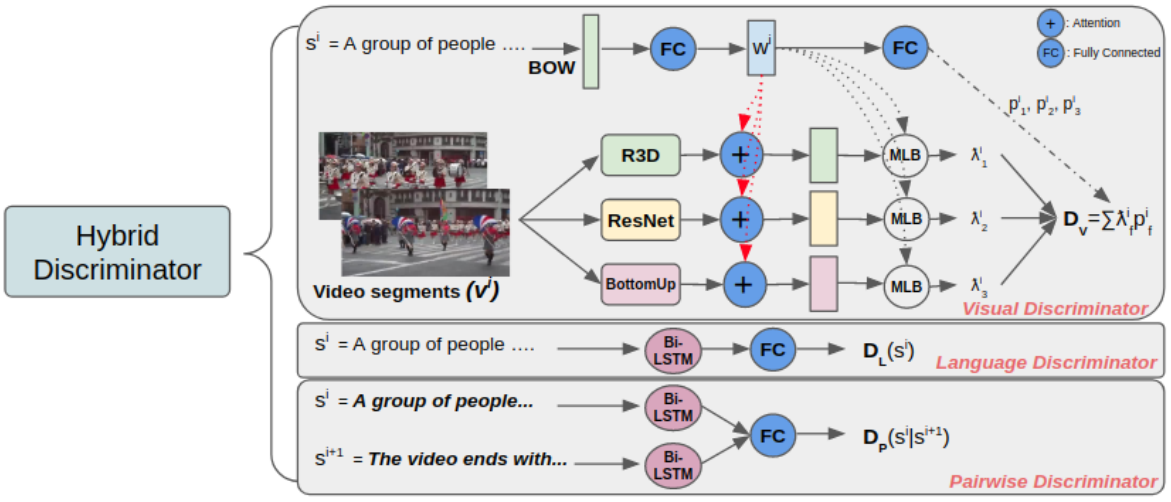

Figure 4.5. The overview of Hybrid Discriminator. Main goal of discriminator is to score generated captions for events in a video. s^i and s^{i+1} are consecutive captions generated by generator for two consecutive events, i and $i+1$. v^i stands for visual features for ith event in a video. This discriminator is referenced as $HybDis$ in our experiments.

There are two types of negatives as inputs to visual discriminator D_V; mismatched ground truth and mismatched generated captions. In first two epoch, mismatched negatives are randomly selected, starting from third epoch, they are selected from captions of videos which have same activity with target caption. There is one point that needs attention: visual discriminator, D_V, does not use captions of generator, with the idea that early at training step of the generator, one input to generator was visual attributes, v_m^i of an event i. They will be already aligned with visual features of subject video, v_m^i as input.

To be able to determine only visual alignment, generated captions are encoded with Bag of Words (BOW) instead of Long-Short Term Memories (LSTMs). Whole sentence is encoded with BOW, then BOW vector is embedded by a linear layer. Resulting sentence encoding is named as w_i. Visual discriminator uses visual attributes consisted of video, image and object label features. To fuse these features(f) together, each attribute is encoded with attention mechanism [50] based on sentence embedding, w^i, resulting as v_f^i. Multi-modal Low-rank Bilinear Pooling (MLB) [63] is used over these two representations, w^i and v_f^i(for each attribute separately). Score of each visual attribute and sentence representation is calculated as:

$$p_f^i = \sigma(tanh(U^T v_f^i) \odot tanh(V^T w^i)), \tag{16}$$

In Eq.-16, σ is sigmoid, $\odot$ is Hadamart product, U,V are linear layers. Weights for each p_f^i is learned based on the sentence representation, w^i as follows:

$$\begin{aligned} \lambda_f^i &= \frac{e^{a_f^T w^i}}{\sum_j e^{a_j^T w^i}}, \\ D_V(s^i|v^i) &= \sum_f \lambda_f^i p_f^i, \end{aligned} \tag{17}$$

a_j are learned parameters and D_V is weighted, λ_f^i, sum of scores, p_f^i.

4.3.2. Language Discriminator

Language Discriminator, D_L, is used for ensuring diversity and fluency for generated captions. Generator G, has no structure for checking language structure in generation process. Although pointing out difference between diversity of real and fake sentences can be achieved with a single discriminator, to capture missing fluency and repetitive N-grams in generated captions, a separate discriminator named as language, D_L, is used[1]. Negative inputs of this discriminator are mixture of randomly mixed words and repeating phrases at the same sentence. This sentence, s^i, is encoded with a bi-directional LSTM and both hidden states are concatenated as $\bar{h}^i$, then a fully-connected and sigmoid layer follows bi-directional LSTM.

$$D_L(s^i) = \sigma(W_L \bar{h}^i + b_L), \tag{18}$$

4.3.3. Pairwise Discriminator

Pairwise discriminator, D_P, is used to score diversity of two consecutive generated event captions, s^i and s^{i-1}. At the end of caption generation process, a caption paragraph will be created. While checking for diversity, D_P ensures coherency of these generated captions that forms a paragraph. For negative inputs to D_P, order of sentences in caption paragraph are shuffled and two random sentences are selected from them.

To obtain D_P score, both sentences are encoded with bi-directional LSTM, their hidden states, $\bar{h}^i$ and $\bar{h}^{i-1}$ are concatenated. If s^i is first sentence, no score is calculated. D_P score is calculated as:

$$D_P(s^i|s^{i-1}) = \sigma(W_P[\bar{h}^{i-1}, \bar{h}^i] + b_P). \tag{19}$$

W_P, b_P are parameters and σ is a sigmoid layer.

4.3.4. Proposed Similarity Discriminator

Our main goal in this part is adding a sub-discriminator to hybrid discriminator which will enable discriminator to score similarity between generated caption, s^i and closest caption, g^i fetched by our proposed pipeline. For this purpose, we define a sub-discriminator namely Similarity Discriminator, D_S, similar to pairwise discriminator. But this time, instead of checking diversity of two consecutive generated event captions, our similarity discriminator scores similarity between s^i and g^i.
To obtain D_S score, both s^i and g^i are first encoded with bi-directional LSTM, their hidden states, $\bar{h}_s^i$ and $\bar{h}_g^i$ are concatenated, accordingly. We propose computing D_S score as follows:

$$D_S(s^i|g^i) = \sigma(W_S[\bar{h}_s^i, \bar{h}_g^i] + b_S). \tag{20}$$

By help of this sub-module, we aim evaluating similarity between closest caption and generated caption. If the similarity score is close to 0, we penalize this by similarity discriminator, D_S while selecting best scored caption from generator. On the other hand, if similarity score is high, we encourage hybrid discriminator to select that generated captions which are similar with closest caption. This is somehow dependent on fetched captions similarity with generated caption. To compensate this, we use a hyper-parameter and use this hyper-parameter as a weight to control contribution from similarity discriminator to Hybrid Discriminator.

4.4. Adversarial Inference with Basic Hybrid Discriminator

Basic Hybrid Discriminator has three sub-discriminators as in AdvInf[1]: Visual Discriminator, Language Discriminator and Pairwise Discriminator.
Amongst the K sampled sentences from the proposed generator $\tilde{G}$, the best sentence is selected by finding the sentence that yields the maximum hybrid discriminator score such that:

$$s_*^i = s^i_{argmax_{n=1..K} D(s_n^i|v^i, s_*^{i-1})} \tag{21}$$

where s_*^{i-1} is previous best sentence, and $D(.)$ is the final score of the hybrid discriminator, calculated by weighted combination of individual discriminators:

$$D(s_n^i|v^i, s_*^{i-1}) = \beta D_V(s_n^i|v^i) + \gamma D_L(s_n^i) + \theta D_P(s_n^i|s_*^{i-1}) \tag{22}$$

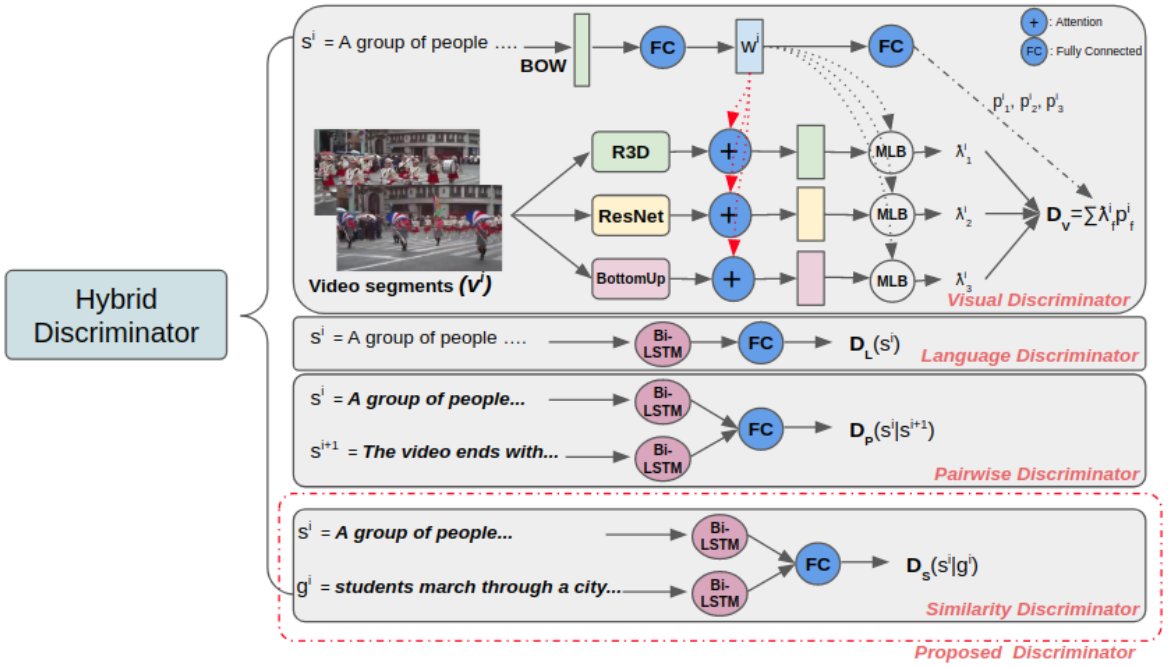

Figure 4.6. The overview of proposed Hybrid Discriminator with Similarity sub-discriminator. s^i is generated sentence by $\tilde{G}$, g_i is closest caption from Conceptual Captions[9] dataset. Our goal is scoring similarity between s_i and g_i, then adding it to overall score of Hybrid Discriminator. This discriminator is referenced as $HybDis + Sim$ in our experiments.

where β, γ, θ are hyperparameters and s_n^i is nth sampled sentence from $\tilde{G}$. We will reference this model as *HybDis* in our experiments.

4.5. Adversarial Inference with Proposed Hybrid Discriminator

Our Proposed Hybrid Discriminator has four sub-discriminators instead of three: Visual Discriminator, Language Discriminator, Pairwise Discriminator and proposed Similarity Discriminator. Overview of *Hybrid Discriminator with Similarity sub-discriminator* is presented in Figure-4.6..

Instead of formula for hybrid discriminator given in Eq.22 we propose adding similarity score between generated caption and fetched closest caption with our proposed pipeline from Conceptual Captions dataset as in Eq.23 as follows:

$$D(s_n^i|v^i, s_*^{i-1}) = \beta D_V(s_n^i|v^i) + \gamma D_L(s_n^i) + \theta D_P(s_n^i|s_*^{i-1}) + \phi D_S(s_n^i|g^i) \tag{23}$$

where $\beta, \gamma, \theta, \phi$ are hyperparameters, s_n^i is nth sampled sentence from $\tilde{G}$, g^i is closest caption fetched. We will reference this model as *HybDis+Sim* in our experiments.

5. EXPERIMENTS AND RESULTS

5.1. Datasets

We evaluate the effectiveness of the proposed approach on ActivityNet Captions [13] video captioning dataset, where we utilize auxiliary phrases extracted from Conceptual Captions [9] image captioning dataset. In the following, we briefly describe these two datasets and discuss our experimental setup.

The ActivityNet Captions contains 10K videos in train split and 4.9K videos in validation split. Each training video is annotated with a single reference paragraph, whereas the videos in the validation set have two reference paragraphs, each provided by a different annotator. As done in prior works [1, 15, 18], we used the videos in the validation set for both validation and testing – utilizing the first reference description for assessing the test performance and the second one for development and training of the models. Our approach involves retrieving visually similar images from Conceptual Captions dataset, hence we need RGB frames of the videos in ActivityNet so we downloaded them by their URLs. Due to permission issues or videos deleted from YouTube, we could not download the whole dataset and only managed to obtain 91% of the videos. An illustration of dense event video captioning sample from

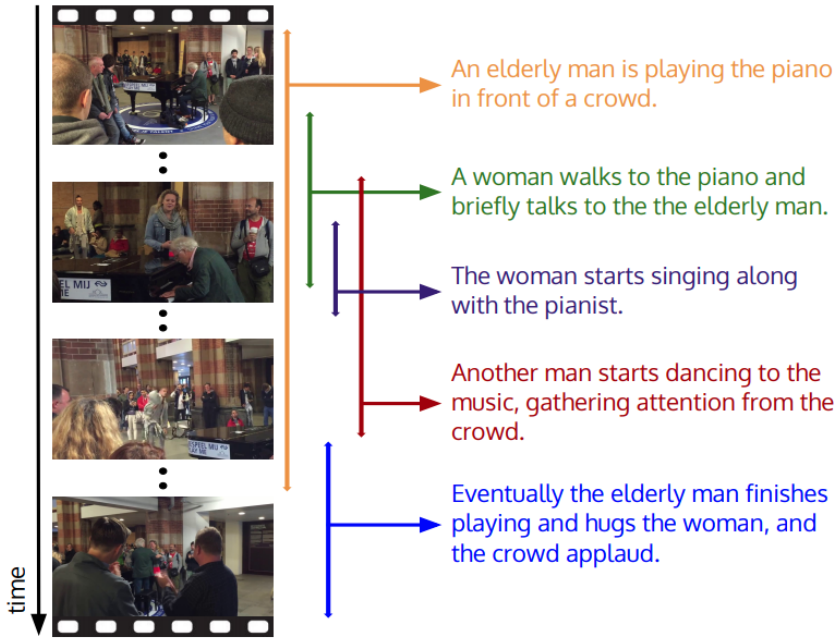

Figure 5.1. An illustration of Dense-event captioning sample from ActivityNet Dataset. (Image is taken from [10].)

Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

Conceptual Captions: a worker helps to clear the debris.

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Figure 5.2. Image and image description examples from the Conceptual Captions dataset. (Image is taken from [9].)

ActivityNet Dataset is shown in Figure-5.1.

Conceptual Captions [9] dataset is collected from the web with the objective of having large-scale in-the-wild image captions. The dataset contains around 3.3M web images for training, 28K for validation, and 22.5K for testing. Each image has a single description, which correspond to cleaned and hypernymized alt-text html attributes associated with the images. We deliberately select Conceptual Captions dataset as a means to provide additional context because of its diversity and size. In particular, in our experiments, we utilize all the images in the training set of Conceptual Captions and extract auxiliary phrases from the captions of images that are visually similar to the input video clip, as explained before. Image and image description examples from the Conceptual Captions dataset are shown in Figure-5.2..

5.2. Implementation Details

5.2.1. Metrics

We use the evaluation tool provided in [1]. This benchmark suite evaluates the generated captions at paragraph-level [1, 22, 56]. Moreover, it does not focus on the event detection task and employs ground truth event intervals, which also allows a fair comparison with the baseline method [1]. For quantitative evaluation, we employ the commonly used METEOR [64], BLEU-4 [65] and CIDEr-D [66] metrics. In addition to these metrics, Div-1, Div-2 [57] and RE-4 [22] are also evaluated. Div-1 and 2 metrics used to evaluate unique N-grams ratio compared to total words of generated captions. RE-4 metric on the other hand, evaluates number of N-gram that are repeated in generated captions and number of N grams

are set as 4. The latter metrics are used to evaluate lexical diversity and to detect repetition of phrases in generated captions.

5.2.2. Visual Features

We follow the steps described in [1] and encode each video with three different visual features, namely video-, image- and object-level features.

Video features are 8192-dimensional 3D convolution based features, denoted as R3D [11], which are extracted by a ResNext-101 model pre-trained on Kinetics dataset [67]. A basic block of ResNext model is shown in Figure-5.3.. Image features correspond to 2048-dimensional ResNet-152 features [12] obtained by a model pre-trained on the Imagenet dataset [68]. A basic ResNet block model is shown in Figure-5.4.. While image features are extracted at every 16 frames, video features are obtained by setting the temporal resolution to 16. Finally, each video clip is divided into 10 regular intervals and then the extracted features are mean pooled [22, 69].

For the object features, we also use the strategy by [1] and detected objects at the start, end and middle frames of a clip. These objects are detected by Faster R-CNN detector [70] trained with Visual Genome [71] from as [72]. Top-16 detection labels are then encoded with standard bag of words features, which are additionally weighted by their detection scores.

Finally, these video-, image- and object-level features are concatenated to obtain a combined visual representation, denoted with v_m^i. Moreover, to obtain a more contextualized representation during decoding phase, we apply temporal attention mechanism proposed in [50], as explored in previous work [1, 22, 69].

5.2.3. Training and Evaluation Setup

For training, we use 16 as batch size. Adversarial inferences main components, generator and discriminator, are trained with cross entropy loss. The weight for negative input weights used in the discriminator are used as 0.5. The weights for the visual discriminator, the language discriminator and the pairwise discriminator are empirically set to 0.8, 0.2 and 1.0, respectively. The weight for proposed similarity discriminator is set to 0.5, if used. We use ADAM [73] optimizer and set learning rate as 5×10^{-4}. During training, we set the temperate for sampling as 1.0 whereas at the inference, this parameter is set to 0.2. We generate a total of $K = 100$ samples at inference. Moreover, we use the maximum number of noun or verb phrases from closest captions (if more then one closest caption is used) as $k_{aux} = 10$ where we zero pad this vector to 15, as detailed in Section 4.2.1..

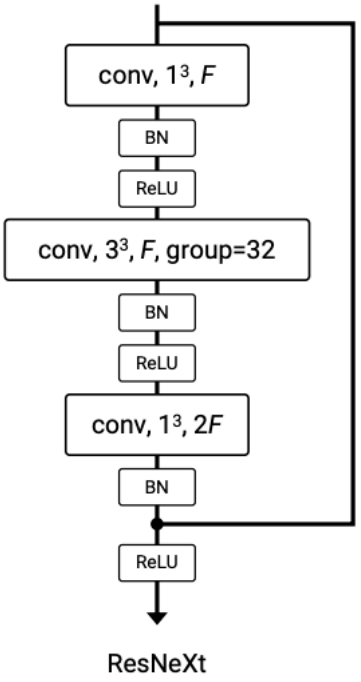

Figure 5.3. Block of ResNext. (Image is taken from [11].)

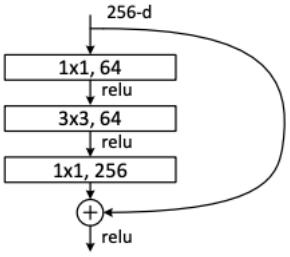

Figure 5.4. Block of ResNet. (Image is taken from [12].)

5.3. Experimental Results

In our experiments on ActivityNet Captions, we test two alternative versions of our model, referred to as a) Ours($k_{aux} = 1$, $HybDis$) and b) Ours($k_{aux} = 10$, $HybDis$). In the first one, we use the phrases from the caption of the single most visually and conceptually closest image caption as auxiliary data in the proposed caption generation framework. On the other hand, in the latter, we exploit a wider auxiliary context by considering a larger number of phrases extracted from the captions of $k_{aux} = 10$ similar images. There is a trade-off here since enlarging the neighborhood size might introduce noise (irrelevant phrases) whereas using only too few phrases could enforce a large bias and reduce diversity. Again, we note that

Table 5.1. Comparison against the base model of AdvInf [1]. Both of our models give better performances in terms of both automatic metrics, METEOR, BLUE-4, CIDEr-D. But there is a catch, the generated descriptions are on par linguistic diversity compared to AdvInf, as we integrate auxiliary lexical context into the generation process.

Method	Per video			Overall		Per video		
	METEOR	BLEU-4	CIDEr-D	Vocab Size	Sent Length	Div-1	Div-2	RE-4
AdvInf[1]	13.94	8.88	17.56	2340	12.35	**0.626**	**0.780**	**0.052**
Ours ($k_{aux} = 1$, $HybDis$)	14.05	**9.20**	**17.94**	1905	12.29	0.619	0.775	0.055
Ours ($k_{aux} = 10$, $HybDis$)	**14.46**	8.88	15.54	1776	14.30	0.574	0.747	0.075

here we employ ground truth segments for events and only focus on decoding better paragraph captions from these provided segments. To have a fair comparison with the baseline AdvInf model [1], we further trained it with our downloaded version of the dataset, where some videos are missing, as detailed in Section 5..

As can be seen from Table 5.1., our models both outperform AdvInf in terms of METEOR, BLEU-4 and CIDEr-D metrics and on par in terms of the diversity metrics. Also we note that one of the main advantages of our approach is that the proposed auxiliary data pipeline is quite generic and in fact the extracted phrases can be used to boost the performance of any dense video captioning model. It allows a model to have access to additional lexical input to increase the certainty of visual information.

In Figure 5.5.,5.6.,5.7., we show examples of paragraph descriptions generated by our ($k_{aux} = 1$, $HybDis$), AdvInf [1] and ground-truth captions. These qualitative examples show that our model generates more diverse and informative captions when compared with AdvInf. As in these examples, our model generates more coherent and linked captions than baseline. In first example, paragraph captions generated by our model start with the location; *'standing in a gym'* and *'a large gymnasium'*, continues with the exercise; *'spins', 'flips' and 'twirls'* and ends with the *'walking away'* action. It is more diverse and coherent and with a clear continuity of actions. In second example, although AdvInf also generates meaningful sentences, our captions include more information with additional words *'lifts'* and *'drops'*.

5.3.1. Ablation Studies

In Table-5.2., we present an ablation study in which we evaluate the impact of various design choices on the captioning performance. In addition, we evaluate our proposed Hybrid Discriminator with similarity sub-discriminator, D_S in this study. Specifically, we have tested the following variants:

Our Model ($k_{aux} = 1$, $HybDis$): A woman is seen standing in a *gym* holding a pose and holding a baton and flipping around *a large gymnasium*. She twirls her baton and twirls around the baton. She does *a baton routine*. The girl *spins* her baton, *flips* and twirls the baton. She continues to dance around the baton and *ends by walking away*.

AdvInf [1]: A woman is standing in a room. She is twirling a baton and twirling her baton. She twirls and twirls around and twirls the batons. She starts twirling batons and the baton. She does a flip and twirls the batons.

Ground Truth: The video starts with a lady walking to the center of a court. The lady starts doing a rendition and spinning a stick around while doing fancy tricks with it. The lady continues her rendition and spins the stick using her neck. The lady throws the stick into the air and catches it. The lady ends the rendition on the floor.

Figure 5.5. Qualitative comparison of our proposed model with the Adversarial Inference [1] for a sample video from ActivityNet Captions dataset [13] with action of gymnastic. This video includes a dense caption involving five sentences one for each event. Our model ($k_{aux} = 1$) generates more coherent and diverse captions for events as seen in blue colored parts. (Best viewed in color.)

Our Model ($k_{aux} = 1$, $HybDis$): *A woman* is seen speaking to the camera and *leads into lifting a weight*. *The woman* lifts the weight and *lifts it up and then drops it over* his head. She then lifts the weight over his head and then *drops it down* and then *walks back* to the ground.

AdvInf [1]: A man is seen bending over a bar. The man lifts the weight above his head. He lifts the weight above his head.

Ground Truth: A female weight lifter bends at the knees.She lifts a barbell to her chest. She then lifts it over her head before dropping it heavily to the ground.

Figure 5.6. Qualitative comparison of our proposed model with the Adversarial Inference [1] for a sample video from ActivityNet Captions dataset [13] with action of weight lifting. This video includes a dense caption involving three sentences, one for each event. Our model ($k_{aux} = 1$) generates more coherent and diverse captions for events as seen in blue colored parts. (Best viewed in color.)

Table 5.2. Ablation study showing the effect of different design choices for our method on the captioning performance.

Method	Per Video			Overall		Per Video		
	METEOR	BLEU-4	CIDEr-D	Vocab Size	Sent Size	Div-1	Div-2	RE-4
Ours ($k_{aux} = 1, Concat$)	14.39	**9.50**	17.92	1909	13.14	0.57	0.75	0.07
Ours ($k_{aux} = 1, AvgPool$)	14.41	9.38	17.54	2007	12.85	**0.62**	0.76	**0.06**
Ours ($k_{aux} = 1, AvgPool + Concat$)	13.97	9.20	**18.51**	1835	12.21	**0.62**	**0.78**	0.07
Ours ($k_{aux} = 1, AvgPool + Linear + Concat$)	14.39	9.39	18.41	1705	13.42	0.58	0.75	0.08
Ours ($k_{aux} = 1, Attention, HybDis$)	14.05	9.20	17.94	1905	12.29	**0.62**	**0.78**	**0.06**
Ours ($k_{aux} = 10, Attention, HybDis$)	**14.46**	8.88	15.54	1776	14.30	0.58	0.75	0.08
Ours ($k_{aux} = 1, Attention, HybDis + Sim$)	14.32	9.27	16.69	2104	13.19	0.60	0.76	**0.06**

Our Model ($k_{aux} = 1$, $HybDis$): A man is seen speaking to the camera and leads into a person riding down a hill. The man is then shown of the camera and leads into a man riding down a hill. The man is *snowboarding down a hill*. He then *goes down the mountain* going down the *mountain slope*.

AdvInf [1]: A man is seen riding down a snowy hill and leads into him speaking to the camera. The man continues to ride around on the skis while looking back to the camera and leads into him riding down a hill. The people continue riding down the hill and ends with him moving around and looking off into the distance. The man then begins skiing down the hill while looking off into the distance.

Ground Truth: A man is skiing down some snow at a very fast speed. Snow is building up all over his face, it looks extremely cold. He kind of almost falls down but continues moving. He is skiing fast through trees, passing by other skiers, he continues to go so fast.

Figure 5.7. Qualitative comparison of our proposed model with the Adversarial Inference [1] for a sample video from ActivityNet Captions dataset [13] with action of skiing. This video includes a dense caption involving four sentences, one for each event. Our model ($k_{aux} = 1$) generates more coherent and diverse captions for events as seen in blue colored parts. (Best viewed in color.)

1. *Ours ($k_{aux} = 1, Concat, HybDis$):* We concatenate one of auxiliary phrase word embedding with previous ground-truth word embedding in every timestamp and then feed this vector to generator instead of only previous ground-truth word (or the ground-truth word during training) embedding.

2. *Ours ($k_{aux} = 1, AvgPool, HybDis$):* We first average pool all the auxiliary noun phrases and verb phrases, and concatenate this embedding vector with the embedding of the previously generated word (or the ground-truth word during training) at each time step of decoding.

3. ***Ours*** *($k_{aux} = 1, AvgPool + Concat, HybDis$):* We first average pool all the auxiliary noun phrases and verb phrases separately, and concatenate these two embeddings with the embedding of the previously generated word (or the ground-truth word during training) at each time step of decoding.

4. ***Ours*** *($k_{aux} = 1, AvgPool + Linear + Concat, HybDis$):* After retrieving auxiliary noun phrases and verb phrases, we again apply average pooling independently to the noun and verb embeddings, but this time, we introduce an additional linear layer, in a fashion similar to DAN encoder [74]. We then concatenate the output of this layer with the embedding of the previously generated word (or the ground-truth word during training) at each time step of decoding.

5. ***Ours*** *($k_{aux} = 1, Attention, HybDis$):* This is basically the version of our approach described in Section 4., which exploits phrases extracted from the single most visually similar image to the middle frame of a given video event and uses an attention mechanism to focus on different auxiliary phrases at different times of decoding.

6. ***Ours*** *($k_{aux} = 10, Attention, HybDis$):* This is the version of our approach described in Section 4. with $k_{aux} = 10$. In this version, we consider a larger number of phrases extracted from the captions of 10 similar images and use an attention mechanism to focus on different auxiliary phrases at different times of decoding.

7. ***Ours*** *($k_{aux} = 1, Attention, HybDis + Sim$):* This is the version of our approach described in Section 4. as $k_{aux} = 1$ and different than other variant hybrid discriminator with usage of Similarity Discriminator, D_S as forth sub-module. We name this discriminator design as $HybDis + Sim$. In this version, we consider phrases extracted from the captions of closest image and use an attention mechanism to focus on different auxiliary phrases at different times of decoding.

We show loss plots for models aforementioned in between Figure-5.8.-5.14.. All figures in this part, show generator loss on the left, discriminator loss on the right. These loss plots suggest that each model variant is a slight variant of the other. In fact, when we evaluate scores in Table-5.2., we see slightly better or similar scores between variants.

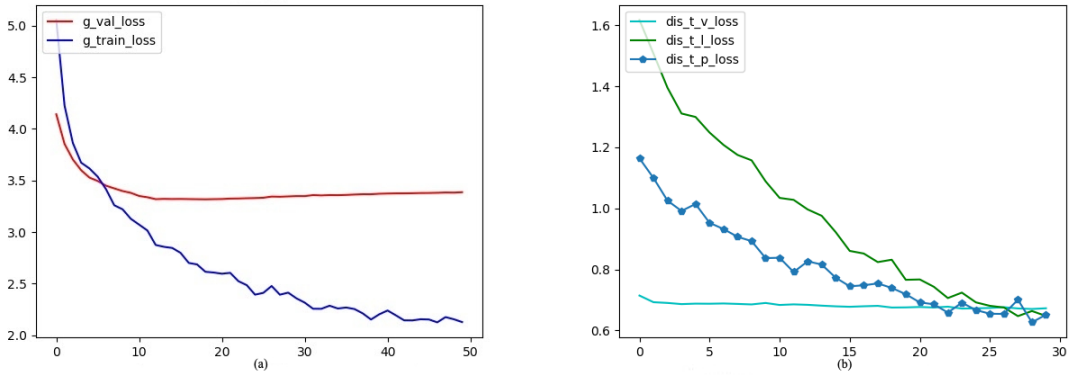

Figure 5.8. Loss plots for model, Ours ($k_{aux} = 1, Concat, HybDis$).

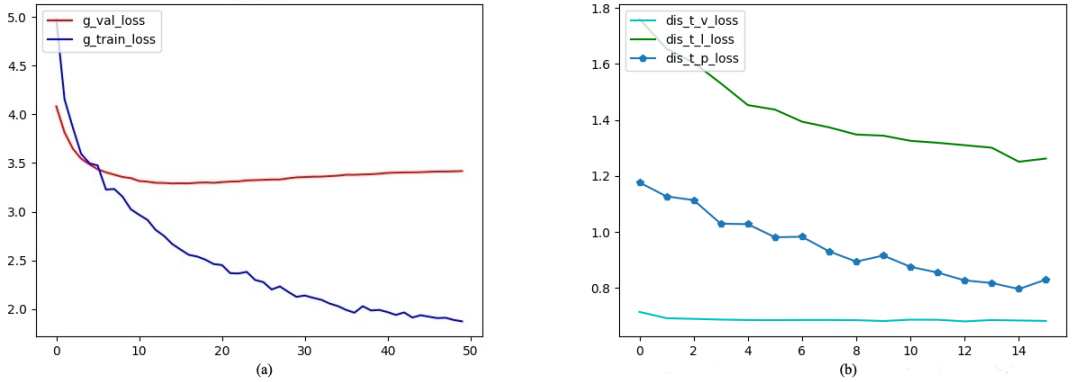

Figure 5.9. Loss plots for model, Ours ($k_{aux} = 1, AvgPool, HybDis$).

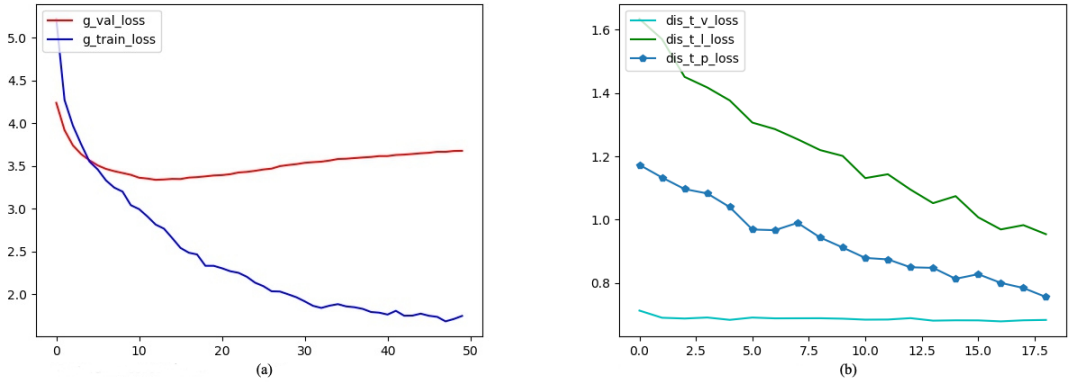

Figure 5.10. Loss plots for model, Ours ($k_{aux} = 1, AvgPool + Concat, HybDis$).

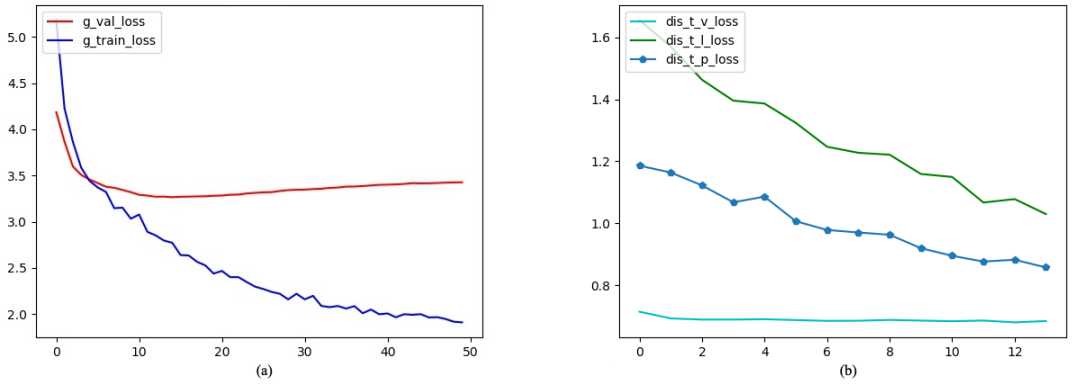

Figure 5.11. Loss plots for model, Ours ($k_{aux} = 1, AvgPool + Linear + Concat, HybDis$).

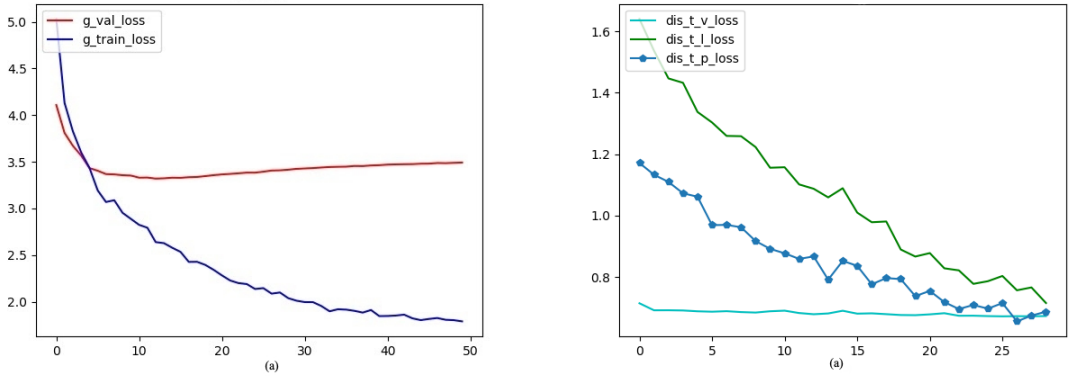

Figure 5.12. Loss plots for model, Ours ($k_{aux} = 1, Attention, HybDis$).

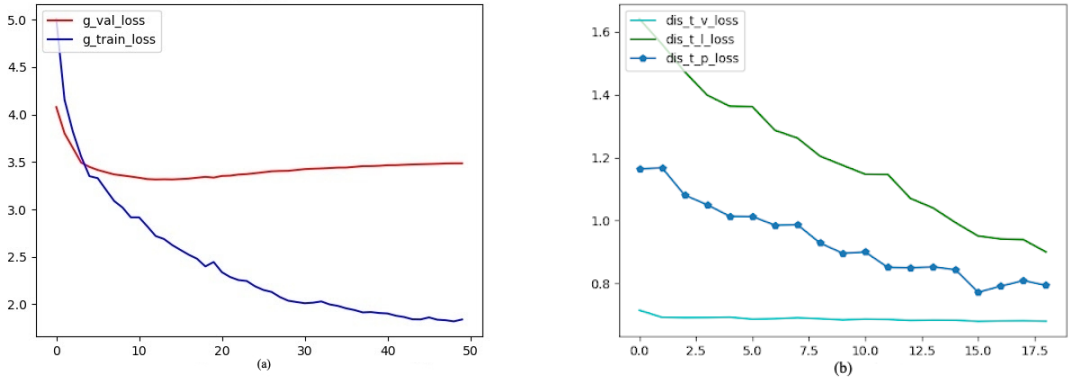

Figure 5.13. Loss plots for model, Ours ($k_{aux} = 10, Attention, HybDis$).

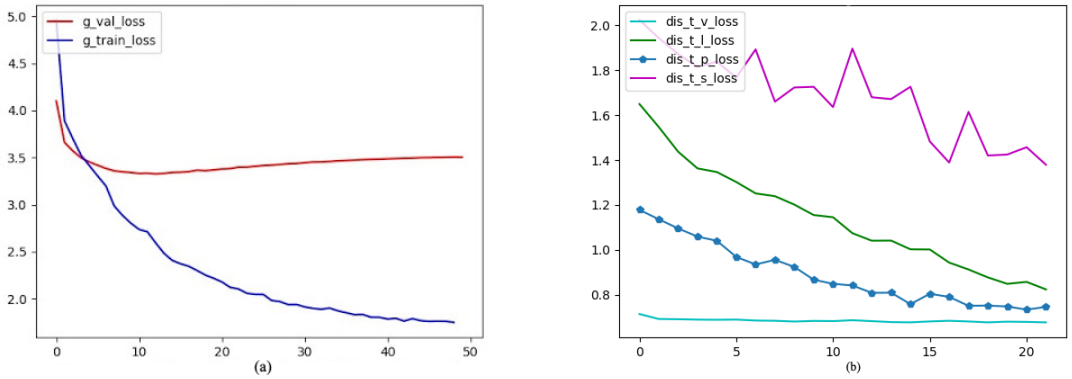

Figure 5.14. Loss plots for model, Ours ($k_{aux} = 1, Attention, HybDis + Sim$).

According to our analysis, we see that all these variants achieve better results than baseline model of AdvInf with respect to automatic metrics and they are on par or slightly worse in terms of diversity. In model $k_{aux} = 1, Concat, HybDis$), we take auxiliary phrases as a sequence. We get compatible results when compared with our basic $k_{aux} = 1, Attention, HybDis$ setup, but taking this phrase list has a bias in it. With this idea, we are deemed to accept that generated captions will be aligned with these phrase in temporal dimension. But this claim will be wrong with a generated stop-words in the captions *e.g. a, the, is*. Although scores are high, we do not accept this setup as our baseline.
Furthermore, in comparison of models $AvgPool + Concat$ and $AvgPool + Linear + Concat$, we show that using a linear layer as in $AvgPool + Linear+ Concat$ serves as setting $k_{aux} = 10$ with a trade-off and causes decrease in diversity. Although $AvgPool + Linear + Concat$ scores better in METEOR, Bleu-4 and CIDEr-D when compared with our model($k_{aux} = 1$) which uses attention, due to decrease in diversity metrics, we choose our models($k_{aux} = 1, k_{aux} = 10$) with attention as our basic approachs in comparing with baseline method, AdvInf. Our proposed hybrid discriminator with similarity sub-module, ($k_{aux} = 1, Attention, HybDis + Sim$) shows compatible results when compared with our base model, ($k_{aux} = 1, Attention, HybDis$). Further evaluations on this setup is left as a future work.

5.4. Attention Weight Visualization

In this part, we evaluate attention weights learned by our model, Ours ($k_{aux} = 1, Attention, HybDis$) . In this setup, our proposed generator exploits phrases extracted from the single most visually similar image to the middle frame of a given video event and uses an attention

mechanism to focus on different auxiliary phrases at different times of decoding. We show our visualized weights for five different events in a given video. In figures between Figure-5.15. and 5.19.; y-axis is for auxiliary phrases and x-axis is for generated captions. ”no-bias” in figures stands for no bias option as detailed in Section-4..

When auxiliary phrases for all figures are evaluated, we show that we fetch semantically meaningful words which can contribute caption generation process. As in Figure-5.15., our generated caption has *riding, down a street*, and our fetched phrases from image captions of CC dataset are correspondingly, *traveling, dirt road*. *no-bias* option [62], is shown in every figure and means not using any of the auxiliary noun and verb phrases while generating captions. In second figure, Figure-5.16., we show our auxiliary phrases are *portrait, man, violin, playing*. When we compare our fetched phrases, we show that they are aligned with generated caption. In this example model used especially phrases *violin* and *playing* with high attention weights while generating words *playing the violin*.

In Figure-5.17., phrases as *motor boat travels*, *rivers* are conceptual and visually in alliance with event represented in video. We see that while generating caption water model had positive feedback from phrase *river*. In 5.18., model fetches similar words when considered with output sentence. As we see, while generating word *eye*, model strongly attended to phrase *eyes* from phrases. And at last example, we see phrase as *ball and football player* which are truly aligned with action in the event.

As a result from all these weight figures, we show our proposed model successfully fetches captions that are accordance with target captions, and phrases extracted from these captions can be used in video caption generation process successfully.

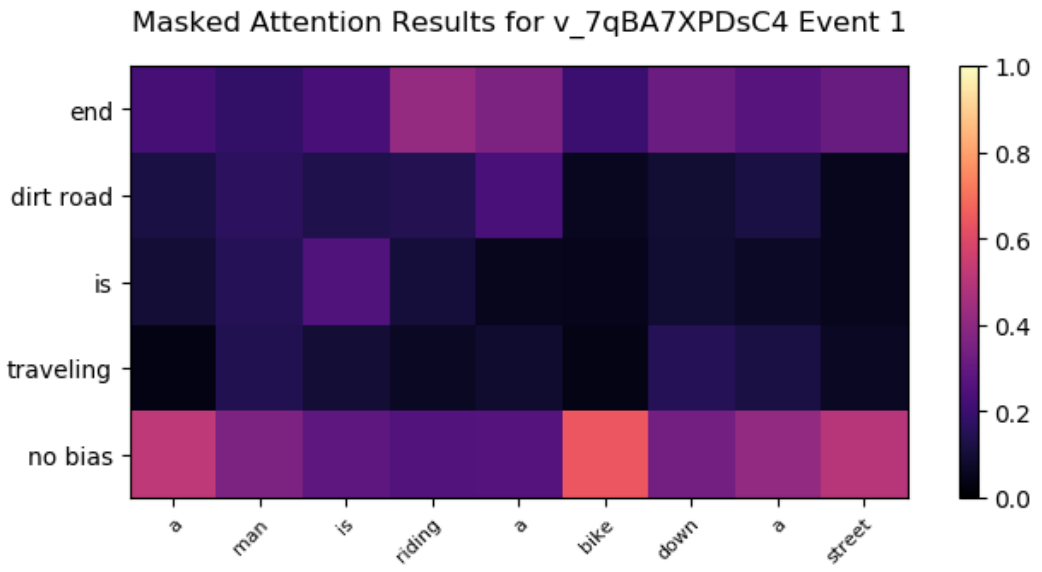

Figure 5.15. Attention visualization for an event in video with id *7qBA7XPDsC4* from ActivityNet [13]. Auxiliary phrases are *end, dirt road, is, travelling*. We use no-bias option with line "no-bias" as shown.

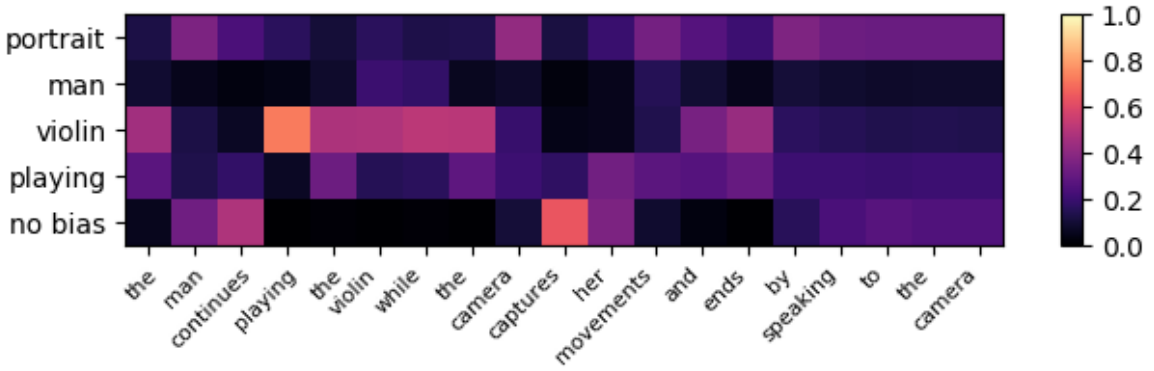

Figure 5.16. Attention visualization for an event in video with id *fJNauQt9Di0* from ActivityNet [13]. Auxiliary phrases are *portrait, man, violin, playing*. We use no-bias option with line "no-bias" as shown.

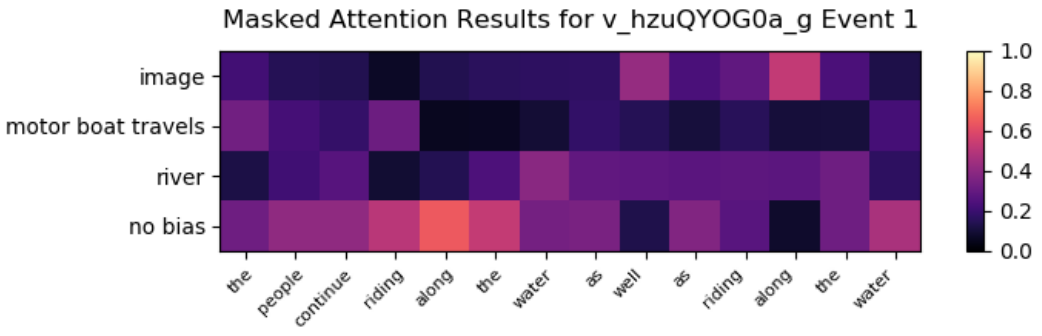

Figure 5.17. Attention visualization for an event in video with id *hzuQYOG0ag* from ActivityNet [13]. Auxiliary phrases are *image, motor boat travels, river*. We use no-bias option with line "no-bias" as shown.

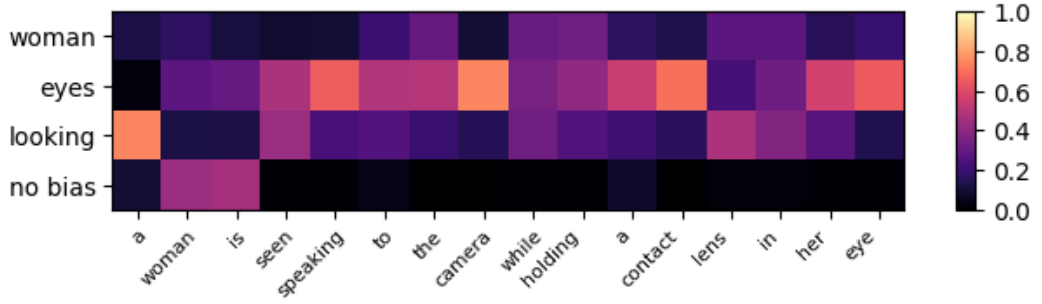

Figure 5.18. Attention visualization for an event in video with id *sAAARH12tdc* from ActivityNet [13]. Auxiliary phrases are *woman, eyes, looking*. We use no-bias option with line "no-bias" as shown.

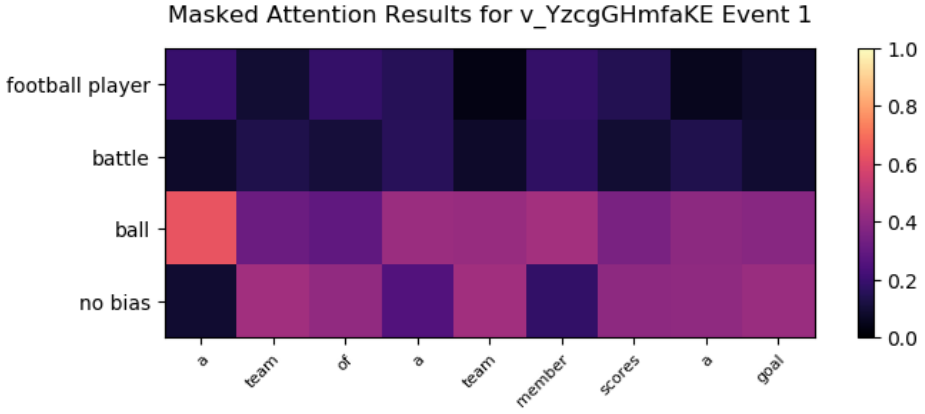

Figure 5.19. Attention visualization for an event in video with id *YzcgGHmfaKE* from ActivityNet [13]. Auxiliary phrases are *football player, battle, ball*. We use no-bias option with line "no-bias" as shown.

6. CONCLUSION

6.1. Conclusion

In this thesis, we propose a new dense video captioning model that allows auxiliary image captions to be used in generating natural descriptions for a given video. Our proposed auxiliary data enhancement pipeline enables extracting meaningful phrases as auxiliary information to be used in caption generation process. This pipeline retrieves a set of images having a content similar to that of the input video and collects a group of noun and verb phrased from the captions of the retrieved images. Then, it utilizes these phrases as additional context information and integrates it with the video features via an attention module within the decoder.

We visualize attention weights for auxiliary phrases and show that our method fetches contextually similar noun and verb phrases that can be used in generation process. Furthermore, we show that with an attention mechanism to incorporate these weighted phrases, we leverage using these phrases effectively. Also, we learn a no-bias option which corresponds to not using any auxiliary data if any of the phrases are irrelevant or not needed in generation pipeline.

Experiments on the ActivityNet dataset demonstrate that the proposed model gives more accurate and more diverse video descriptions than a baseline model. As a result, our method outperforms a baseline model when compared with automated metrics and on par with diversity metrics.

We show variants for our base model. These models represents different fusion techniques to combine textual information from image captions to video captioning process. Even in these setups, we get better results than a baseline model.

With our new proposal over a similarity sub-module in hybrid discriminator, we suggest penalizing the model if generated captions are not similar to fetched closest caption. We show that our results are compatible with our base proposal.

Our pipeline is quite generic, and proposed pipeline and attention mechanism can be used in other dense video captioning models. With the generic approach within this proposal, any image captioning dataset can be used as auxiliary data source instead of Conceptual Captions dataset.

6.2. Future Work

For future work, it would be interesting to adapt the proposed framework under zero-, or few-shot learning scenario and to explore the use of other kinds of auxiliary data extracted from different modalities.

We plan utilizing similarity discriminator with captions other than only the closest one. A new approach to enhance captions for this discriminator will be pursued. To disable noise caused by bias of using to much auxiliary phrases, a pipeline other than only extracting noun or verb phrases can be tried. We use only textual information from image captions dataset to enhance caption generation process. Along with this textual information, image features of closest ones can be used as an additional input.

We use whole Conceptual Captions dataset in image retrieval phase. This method can be used more effectively. A research over generating chunks of images from this dataset based on events or object labels will be conducted.

Using a video captioning dataset as auxiliary data source can be effective, too. With our pipeline, we can fetch similar events from auxiliary videos and use their captions in video captioning process.

REFERENCES

[1] Jae Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. Adversarial inference for multi-sentence video description. In *CVPR*, pages 6591–6601. **2019**.

[2] Jun Xu, Tao Mei, Ting Yao, and Yong 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*, pages 5288–5296. **2016**.

[3] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In *CVPR*, pages 4213–4222. **2018**.

[4] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *ICML*, pages 957–966. **2015**.

[5] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137. **2015**.

[6] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, **2010**.

[7] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, pages 4594–4602. **2016**.

[8] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *CVPR*, pages 6588–6597. **2019**.

[9] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565. **2018**.

[10] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*. **2017**.

[11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555. **2018**.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. **2016**.

[13] Fabian Caba, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970. **2015**.

[14] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37, **2019**.

[15] Spandana Gella, Mike Lewis, and Marcus Rohrbach. A dataset for telling the stories of social media videos. In *EMNLP*, pages 968–974. **2018**.

[16] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*. **2018**.

[17] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, pages 4584–4593. **2016**.

[18] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, pages 8739–8748. **2018**.

[19] Jingwen Wang, Wenhao Jiang, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, pages 7190–7198. **2018**.

[20] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, pages 7492–7500. **2018**.

[21] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *CVPR*, pages 6790–6800. **2018**.

[22] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *ECCV*, pages 468–483. **2018**.

[23] Huijuan Xu, Boyang Li, Vasili Ramanishka, Leonid Sigal, and Kate Saenko. Joint event detection and description in continuous video streams. In *WACV*, pages 396–405. **2019**.

[24] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271*, **2020**.

[25] Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. Dense procedure captioning in narrated instructional videos. In *ACL*, pages 6382–6391. **2019**.

[26] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *CVPR Workshops*, pages 958–959. **2020**.

[27] Sheng Li, Zhiqiang Tao, Kang Li, and Yun Fu. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4):297–312, **2019**.

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680. **2014**.

[29] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, **2017**.

[30] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134. **2017**.

[31] Zheng Xu, Michael Wilber, Chen Fang, Aaron Hertzmann, and Hailin Jin. Learning from multi-domain artistic images for arbitrary style transfer. *arXiv preprint arXiv:1805.09987*, **2018**.

[32] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, pages 533–536, **1986**.

[33] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, **1997**.

[34] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, **1994**.

[35] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*. **2013**.

[36] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36, **2019**.

[37] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659. **2016**.

[38] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, **2004**.

[39] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, **2005**.

[40] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision*, pages 404–420. Springer, **2000**.

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, **2014**.

[42] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, **2016**.

[43] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383. **2017**.

[44] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4894–4902. **2017**.

[45] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Comput. Surv.*, 52(6), **2019**.

[46] Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond Mooney, Kate Saenko, and Sergio Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*. **2013**.

[47] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, pages 2712–2719. **2013**.

[48] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *ICCV*, pages 433–440. **2013**.

[49] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *ICCV*, pages 4534–4542. **2015**.

[50] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *CVPR*, pages 4507–4515. **2015**.

[51] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *CVPR*, pages 1029–1038. **2016**.

[52] Rakshith Shetty and Jorma Laaksonen. Frame-and segment-level features and candidate pool evaluation for video caption generation. In *ACM-MM*. **2016**.

[53] Mihai Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Spatio-temporal attention models for grounded video captioning. In *ACCV*, pages 104–119. **2016**.

[54] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, pages 6504–6512. **2017**.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008. **2017**.

[56] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2005.05402*, **2020**.

[57] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *ICCV*, pages 4135–4144. **2017**.

[58] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*, pages 2970–2979. **2017**.

[59] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *ICCV*, pages 3362–3371. **2017**.

[60] Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *AAAI*. **2018**.

[61] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. **2017**.

[62] Golan Pundak, Tara N Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. Deep context: end-to-end contextual speech recognition. In *2018 IEEE spoken language technology workshop (SLT)*, pages 418–425. **2018**.

[63] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, **2016**.

[64] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Workshop on statistical machine translation*, pages 376–380. **2014**.

[65] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. **2002**.

[66] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575. **2015**.

[67] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, **2017**.

[68] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*. **2009**.

[69] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. **2016**.

[70] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99. **2015**.

[71] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, **2017**.

[72] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086. **2018**.

[73] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, **2014**.

[74] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *ACL*. **2015**.