修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01/23/2023 (MM/DD/YYYY)

専攻名 (専門分野) Department	Computer Science and Communications Engineering	氏 名 Name	Fei Bao	指 導 教 員 Advisor	渡辺 裕 印 Seal
研究指導名 Research guidance	Research on Audiovisual Information Processing	学籍番号 Student ID number	5121FG08-1		
研究題目 Title	Research on the Video Captioning with a Late Fusion Based Multimodal Transformer Network				

1. Introduction

Video captioning has drawn great attention in recent years as the intersectional task of computer vision and natural language processing. It can generate a corresponding sentence to describe its content. Given the tremendous success of the Transformer in visual and linguistic domains, many Transformer-based models [1,2] have been proposed to tackle the task of converting vision to text by understanding various modalities within the video. However, due to the computation of the self-attention mechanism, Transformer-based models often encounter high computational complexity when dealing with the long sequence input, especially for MMT [1] which simply concatenates multiple modalities before the encoder. In addition, as for the multimodal representation in the video, visual features are usually obtained from the deep layer of a pre-trained backbone, but languages are directly embedded into the model. It means that the model thereby should put most of the computations on the linguistic representations to obtain more local context. To mitigate the above issues, we propose a Transformer-based model on the TVC dataset [1] that handles different modalities with separated encoders and fuses them at the decoder side.

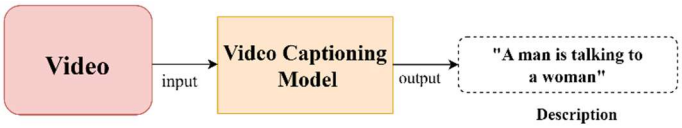

Fig. 1. Illustration of the video captioning task

2. Related works

2.1 Self-attention mechanism

Given an input sequence, self-attention can capture the internal correlation in that sequence by computing the score on each element with others. For example, the self-attention mechanism in the field of machine translation [3] can identify the corresponding German word in an English sentence. Similarly, video is also composed of sequential frames and self-attention is thus applicable to detect internal correlation among frames. The self-attention mechanism in the Transformer model is implemented using scaled dot-production, which projects the input sequence into matrix queries(Q), keys(K), and values(V). The computation of self-attention is shown as

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

Intuitively, the computational complexity is $O(n_A^2 \cdot d)$, where n denotes the length of sequence A and d is the dimension. This reveals that the cost highly depends on the length of the input sequence A.

2.2 Benchmark and baseline

Lei et al. [1] introduced the TV show Captions (TVC) which is a large-scale multimodal video captioning dataset. TVC contains 108K video clips paired with subtitles from 6 TV shows across diverse genres. Besides, each video clip is on average 13.4 seconds in length and has 2 or 4 manually annotated descriptions. In addition, they also proposed an early fusion based Multimodal Transformer (MMT) as a baseline model on the TVC dataset. To efficiently obtain the intra-modal features over two modalities and produce an accurate caption, visual features and subtitles are concatenated simply and then input into the encoder to generate the context for further process. As a result, the cost of self-attention is increased to $O((n_A + n_B)^2 \cdot d)$ due to the longer length for the integration of video sequence A and subtitle sequence B with same dimension d. Although it is a convenient way to aggregate modalities, computational complexity is also increased, and the local context in the linguistic features cannot be explored completely due to the early fusion.

3. Proposed method

We propose a novel model based on the MMT to alleviate the high computational cost of multimodal input and explore more local context in the linguistic representation. Visual and subtitle sequences being processed in the separated encoder are shown in Fig. 2. Same as MMT, video features are extracted beforehand to save the GPU memory. The visual encoder and the decoder have 2 layers while the subtitle encoder has 4 layers to acquire

more local context. Two types of sequences are summed at the fusion module in the decoder after interacting with the ground truth. Therefore, compared with the MMT, the computational cost can be reduced to $O((n_A^2 + n_B^2) \cdot d)$ since each self-attention score of two modalities is computed individually.

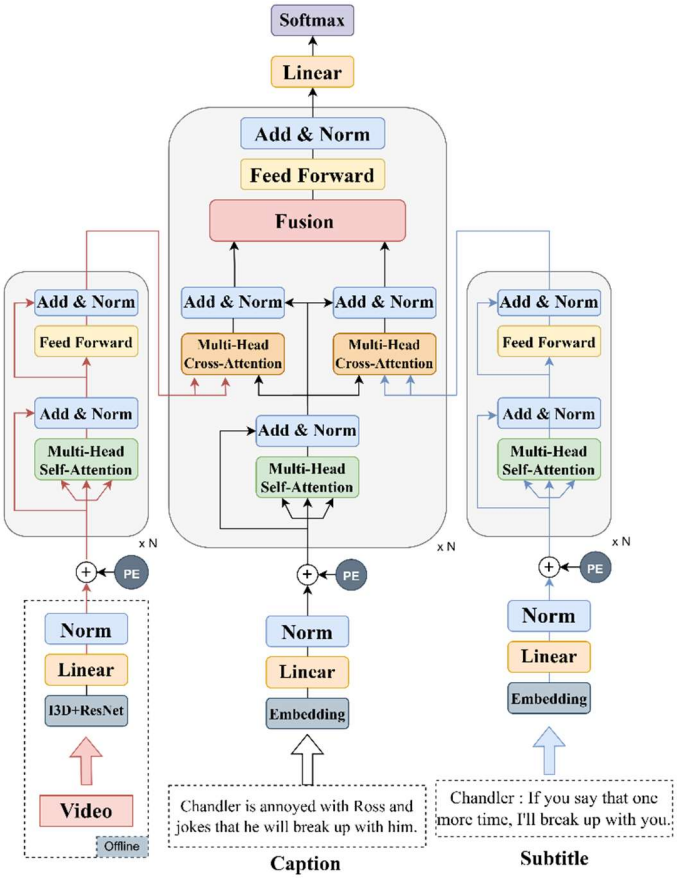

Fig. 2. Overview of the proposed method.

Moreover, we also design additional four variants to explore the effects of interacted context and global context over video and subtitle input. (1) video features and subtitles are processed in the individual encoding stream and then concatenated at the intermediate layer of the encoder. (2) two modalities are encoded in the individual stream only. (3) two modalities interact at the shallower layer and are then concatenated as variant 1 does. (4) similar to variant 2, only interacted context across two modalities are considered.

4. Experiment

We conducted several experiments to validate the effect of our model and variants. We also implanted additional experiments to investigate the effect of different numbers of subtitle layers. We found that our proposed model with 4 layers for subtitle encoder has a competitive result compared to other variants and baseline as shown in Table 1. In addition, the variant with only computing self-attention in each modality performs better than the one with computing the cross-attention of them. It confirms that there is no strong correlation between the video moment and associated subtitles. In addition, we also found that our model has the highest result on CIDEr-D when it has 4 layers in the subtitle encoder.

Table 1. Result comparison of models and the baseline.

	B@4	M	R	C
MMT	10.53	16.61	32.35	44.39
Variant #1	10.66	**16.78**	32.55	45.25
Variant #2	10.64	16.69	32.58	**45.90**
Variant #3	11.09	16.60	32.54	45.24
Variant #4	10.52	16.47	32.17	43.71
Ours (4-layer)	**11.19**	16.54	**32.67**	45.64

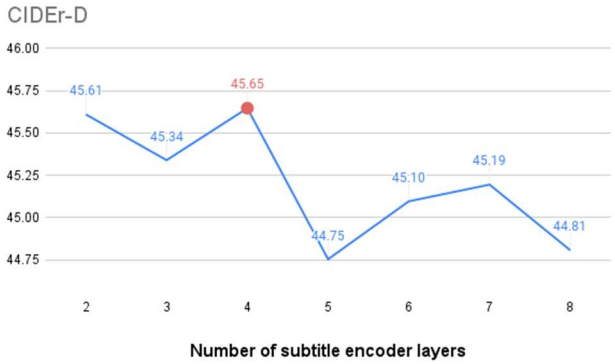

Fig. 3. The value of CIDEr-D regarding different layers in the subtitle encoder.

5. Conclusion

In this research, we investigated the effect of different multimodal fusion strategies in a Transformer network on the video captioning task. We discovered that using distinct encoders for multiple modalities and fusing them later tends to perform better and have a lower computational complexity compared with using a single encoder. Additionally, we found that focusing more on the linguistic modality leads to better results.

6. Reference

[1] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval," in *Eur. Conf. Comput. Vis.,* 2020, pp. 447–463.

[2] V. Lashin and E. Rahtu, "Multi-modal Dense Video Captioning," in *Conf. on Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 4117–4126.

[3] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser and I. Polosukhin: "Attention Is All Your Need", In *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp.6000- 6010.

Research on the Video Captioning with a Late Fusion Based Multimodal Transformer Network

A Thesis Submitted to the Department of Computer Science and Communications Engineering,
the Graduate School of Fundamental Science and Engineering of Waseda University
in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

Submission Date: January 23rd, 2023

Fei Bao
(5121FG08-1)

Advisor: Prof. Hiroshi Watanabe
Research guidance: Research on Audiovisual Information Processing

Acknowledgments

First and foremost, I would like to sincerely thank my esteemed supervisor Professor Hiroshi Watanabe for providing valuable feedback and suggestions throughout my research process. Additionally, I appreciate the experimental environment devices he provided which were greatly beneficial to my research.

Secondly, I appreciate all the members of Watanabe Lab for their insightful conversations and unforgettable time spent together in the lab.

Thirdly, I am also grateful to my friends Bibo Han, Hao Li, Kun Lu, and Shengtao Li for their companionship that has made my study and life in Tokyo, especially during these unusual years.

Last but not least, I would like to express my gratitude to my parents who give me unwavering support and respect for my every decision. Without their understanding and belief in me, it would have been impossible for me to complete my study.

Abstract

Video captioning, the task of generating an automatic description for a video, is still a challenging job for machines. It requires the ability to capture the major event and comprehend the dynamic content within the video. Given the tremendous success of the Transformer in computer vision and natural language processing, many Transformer-based models have been proposed to tackle the task of converting vision to text by understanding various modalities within the video. However, due to the computation of the self-attention mechanism, Transformer-based models often encounter a high computational complexity when dealing with long sequence input. In addition, they also struggle to capture the inter-modal feature from early fused multimodal representations and different modalities are not able to be processed accordingly based on their differences.

To address these issues, we present a novel late fusion based multimodal Transformer network. Our proposed model balances the computational cost and accuracy by processing each modality individually and adding extra layers for the linguistic representation. Additionally, we also design extra four variants to explore the impact of inter-modal and intra-modal features. Eventually, our model achieves 45.65 CIDEr-D and 32.67 ROUGE-L on the TVC dataset, demonstrating its effectiveness.

Keywords: video captioning, multimodal learning, computer vision, video description

List of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Tables

Chapter 4

Chapter 1 Introduction

1.1 General Introduction

1.1.1 Background of Video Captioning

Despite humans being proficient at describing the visual content of a particular video through their visual perception and natural language, it remains a challenge for computers to do with a same accuracy. As the video is a combination of numerous different elements, such as salient objects, motions, backgrounds, audio, etc., machines or computers must be able to distinguish the most crucial information and provide a grammatical and understandable language sentence. While deep learning technology has achieved great success in the field of Computer Vision (CV) and Natural Language Processing (NLP), video captioning also comes to view as the intersection of these two fields for tackling the problem. Video captioning requires comprehending the content of a video and providing a corresponding linguistic description by recognizing visual features, much like image captioning does. Image captioning involves recognizing visual features and describing the content of an image through language. However, the complexity of video description is greatly increased by the need to comprehend dynamic content and track its key components along the sequence of frames.

To bridge the gap between visual media and spoken language, research in the field of video captioning commenced by manually extracting features from video and combining subjects, verbs, and objects (SVO) to accurately describe human activities [1]. After entering the era of deep learning, the task of video description typically consists of two parts. Initially, a deep neural network such as the convolutional neural network (CNN) that is pretrained on a large-scale of the dataset, is employed to acquire spatial and temporal features automatically from the video, then recurrent neural networks (RNN), LSTM or Transformer are engaged to learn the potential correlation

between extracted video features and caption [2,3]. In addition, not only the occurred objects and motions in the video but audio [4] and subtitle [5] information are also taken into the consideration to comprehend the content as video is a combination of multiple modalities.

With the help of video captioning, many tasks including searching for video content and human-robot interaction [6] could be achieved in the future. In addition, video captioning technology could assist millions of people who are blind or visually impaired. [7].

1.1.2 Single Sentence Oriented and Multiple Sentences Oriented

Typically, video captioning can be categorized into two types, depending on the output – single-sentence-based and multiple-sentence-based.

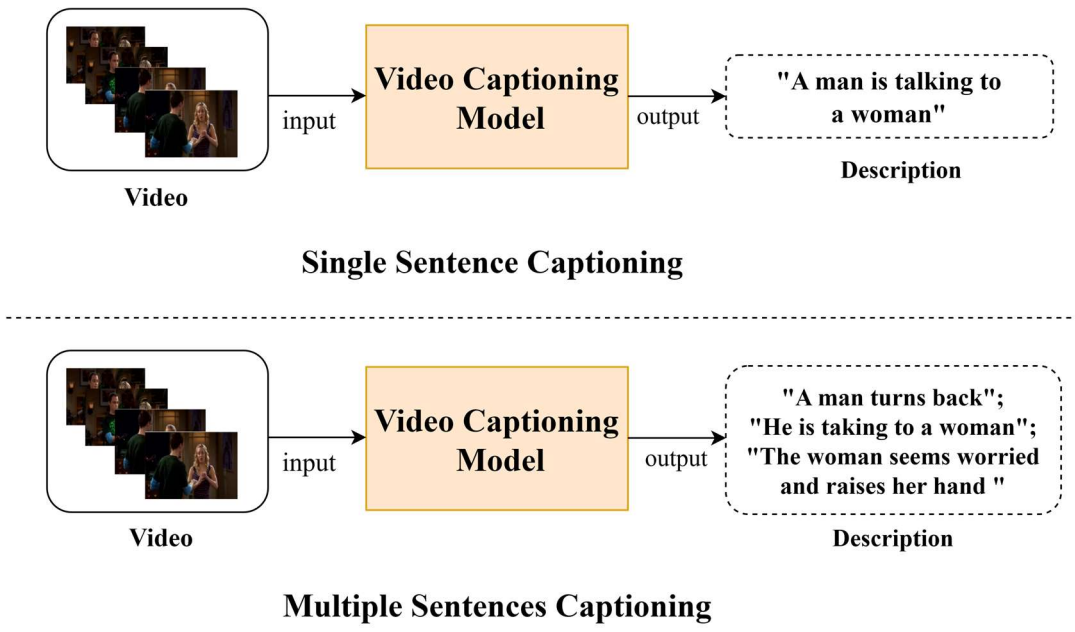

Fig. 1. Illustration of single sentence captioning and multiple sentences captioning for a video of TVC [5].

Single-sentence-based video captioning sometimes can also be known as video summarization[8]. The entire content of the video can be summarized in a single line. However, it is not always guaranteed to provide an accurate summary when there are multiple events occurred in the video, which may lose crucial information, especially

if the video is lengthy.

Multiple sentences generation for a video is also termed dense video captioning. It is suggested to alleviate this issue because it is obvious that only one sentence will not be enough to adequately describe a lengthy video. When there are several events in the video, the dense video captioning model needs to first identify and localize each one before generating a caption for each event of different lengths which may even be overlapped [9].

1.1.3 General architecture

The majority of video captioning techniques [2,3,4,10] adopt the encoder-decoder paradigm and are divided into a two-stage process, feature extraction and caption generation, to address a vision-language issue [33]. In most cases, a deep neural network backbone that has been pretrained on extremely large datasets detects and extracts significant appearance and motion features from the video in the encoding stage. Certain items in the video can be identified by features such as object edges, corners, colour, and texture. The caption generation decoding stage then receives the learned representations and outputs the appropriate description sentence. Due to the fact that training and inferencing backbone extractor and description generator at the same time requires huge GPU memory, most works in this field focus on the sentence generation part and so does our research. Video features are extracted beforehand and then utilized at the decoding stage.

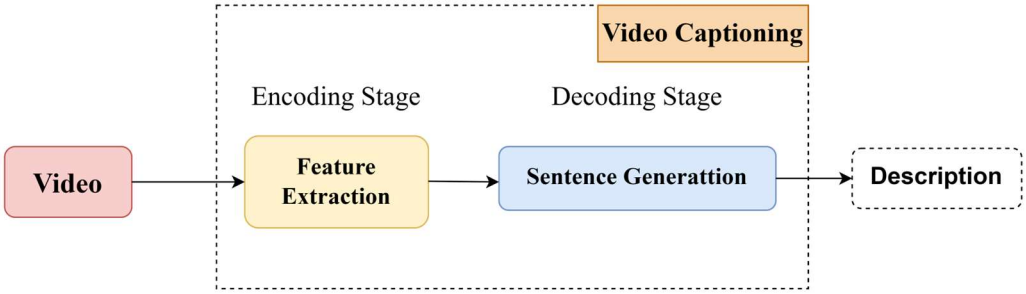

Fig. 2. The workflow of general video captioning approach.

1.2 Problem Statement

1.2.1 Computational cost

With the rapid advancement of technology, the amount of data is getting increased exponentially. This, in turn, requires more computational cost to process the data, which is a problem when met with limited hardware capabilities. How to faster deal with large amounts of data under the current hardware limitations has become the most important problem. The field of video captioning also faces the problem of handling large amounts of data efficiently. Compared to the image caption, a video usually contains more frames and modalities, which means that the model needs more computational cost to deal with extra video representations obtained from the encoding stage. The complexity of attention-based model even reaches quadradic when applied to the larger input sequence due to the computational method of dot-product attention. More details about cost reduction will be introduced in Section 2.1.3.

1.2.2 Drawback of early fusion

A video generally comprises multiple modalities such as images, audio, and subtitles, therefore, being able to effectively learn from the correlation between different modalities might help us better comprehend the video content. Although previous works including Lashin et al. [4] and Lei et al. [5] noticed the importance of multimodal information, they only simply concatenate image features with text before the encoder. This naive approach means that the context of the concatenated sequence cannot be learned explicitly because the image's appearance features are typically

extracted from the deep layers of the backbone, while text features are not. This makes it difficult to effectively understand the relationships between the different modalities in the video. Therefore, different modalities should be carried out in different encoders and the text stream requires more encoder layers to be processed.

1.3 Thesis Outline

The outline of this thesis is followed as below:

Chapter 1: We describe the background of video captioning in the field of the vision-language process and how captions are generated. We also introduce its two categories in terms of the generated results and general architecture adopted by other works. In addition, two critical problems faced by current video captioning models are illustrated and motivated us to propose our model.

Chapter 2: We introduce several related methodologies of video captioning and discuss their advantages and disadvantages, ranging from the LSTM-based to the Transformer-based with multimodal involved. Furthermore, we also introduce widely used relevant datasets and evaluation metrics.

Chapter 3: We demonstrate the framework of our proposed method with the explicit introduction. To better understand the application of the attention mechanism in video captioning, we provide a detailed explanation on the training and inference process of the model. Furthermore, we also present additional 4 variants to investigate the impact of inter-modal and intra-modal features within multiple modalities in the TVC [5] dataset.

Chapter 4: We briefly introduce the experiment environment of the proposed model. Based on the results of our proposed model and the other four models, we compare them with the baseline model and provide a discussion. Besides, we also implement extra ablation experiments to explore how different numbers of linguistic encoder

layers and interaction between different modalities affect the result.

Chapter 5: We conclude this thesis and give a discussion on the improvement in the future.

Chapter 2 Related Works

2.1 Previous methods in video captioning

2.1.1 Long Short-term Memory based

Generally, a video can be seen as a sequence of images, where the order of frames plays a crucial role in comprehending the meaning or context of the video. Likewise, the textual description also contains sequential information, where the meaning will be changed or unmeaningful if the order of words in the sentence is mixed up. Long Short-term Memory [11] has the capacity of resolving the sequential task and long dependency between two elements, therefore, Venugopalan et al. [3] proposed a novel model S2VT using a stack of two LSTMs to tackle the sequence-to-sequence task.

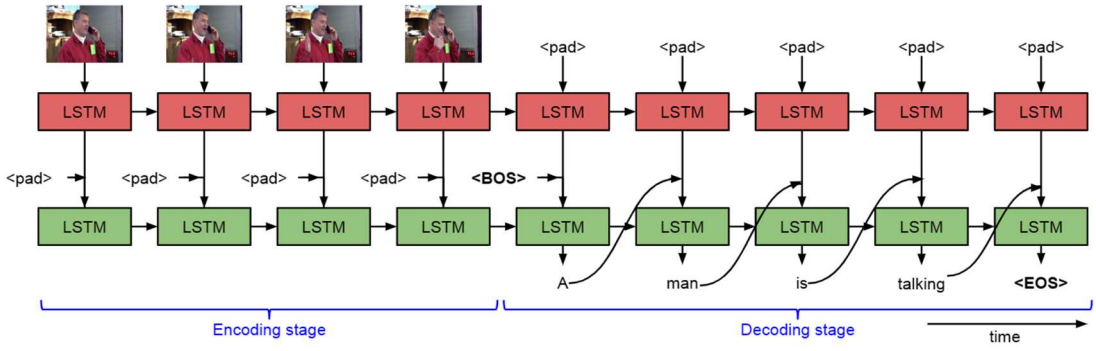

Fig. 3. Framework of S2VT [3].

They first extract visual appearance features from RGB images of the video with a pretrained AlexNet [12] and also the 16-layerVGG [13]. Meanwhile, a CNN [14] initiated with weights trained on the UCF101 [15] video dataset is used on the optical flow images to obtain the temporal information. Then the sequence of feature is fed into the first layer (coloured red) in the encoding stage to model the visual frames sequence. The output hidden state from each LSTM module of the first layer is input into first layer's next LSTM module and the second layer's LSTM module with null padded input words involved. At here, they use the second layer to model the words order. If the model exhausts the input frames and the second layer is fed the beginning-of-

sentence (<BOS>), it will start to generate the word one by one maximizing the log-likelihood of the predicted output sentence with learned context from the encoding stage. Eventually, they apply a softmax function on the words' probability distribution and choose an appropriate word with a highest weighted sum of the score by the flow and RGB networks. During the inference phase, the model will not terminate emitting word token until generating an end-of-sentence tag (<EOS>).

With the help of the stack of two LSTMs, the model can map feature of input video to the specific token in the caption and thus comprehend the content and context. Additionally, the sequential information can also be understood by inputting the optical flow of video.

Although the S2VT is able to handle the variant length of input and learn the temporal representation of the video, it still has a limited performance while obtains long-term dependencies as each unit only looks the previous ones during the processing. Moreover, LSTM is computationally expensive and cannot handle multiple inputs in parallel.

2.1.2 Transformer based

Transformer has capability of handling long-range dependencies as the self-attention mechanism enables the model to scan all of the input tokens simultaneously, which is impossible for LSTM models to do. Additionally, Transformer models require less memory than LSTMs and can analyse many input and output sequences concurrently, making them more effective and quicker at inference and training.

The most crucial component in the Transformer is the self-attention mechanism. Given an input sequence, self-attention can capture the internal correlation in that sequence by computing the score on each element with others. For example, using self-attention mechanism in machine translation[16] is able to identify the corresponding German word in a English sentence. Similarly, video is composed of a sequential frames

and self-attention is also applicable to detect internal corelation among frames.

Assume we have a video feature $V = \{f_1, f_2, f_3, \ldots, f_n\}$ where f_x denotes the feature vector contained in the frame x and total amount of frames is n. The computation of self-attention score for f_1 in a video is demonstrated in the Fig. 4. We first project each feature vector into three different vector query (q), key (k) and value (v) by multiplying individual weighted matrix. Then we take a dot-product of the q with k of the respective vector and is divided by a scaling factor to stabilize the gradient change. It subsequently is applied with a softmax function and dot-products with v and added together to get the self-attention score for current vector f.

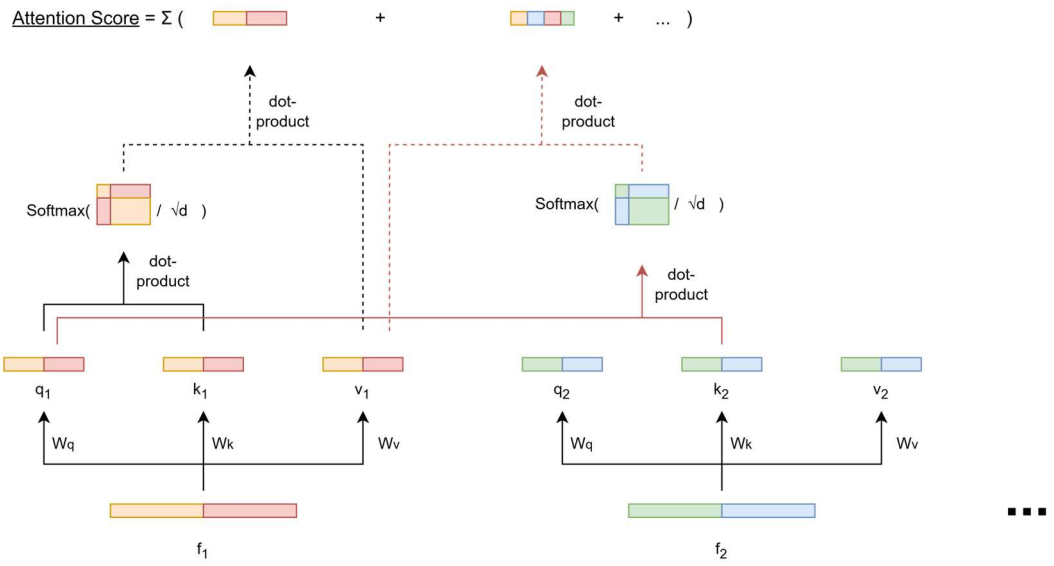

Fig. 4. Illustration of how the vector f1's self-attention score is calculated.

Therefore, if we compute the self-attention for a video, the matrix of outputs can be shown as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

Intuitively, we can find that the computational complexity highly depends on the length of input sequence and the cost is $O(n^2 \cdot d)$ based on (1) and Fig. 4, where n denotes the sequence length and d is the dimension.

With the help of attention mechanism, Transformer-based models typically perform well on the sequence-to-sequence task, including the video captioning. Based on this characteristic, Zhou et al. [17] propose the first RNN-free and Transformer-based model to do dense video captioning. The model is composed of a vanilla transformer encoder-decoder [16] and an extra proposal decoder Temporal Convolutional Networks (TCN) that localizes the events from a video. The encoder takes pre-processed visual representations from pretrained ResNet200 [18] and BN-Inception [19] and further encodes all context information. Then the obtained context is transformed to TCN to output event proposal. The output from visual encoder and proposal decoder is processed into captioning decoder to explore internal corelation between the video segment and its corresponding description sentence. However, the model only puts the concentrations on the visual modality, and it can be further improved with considering other modalities in the video.

2.1.3 Multimodal based

When humans learn a new thing, visual perception, auditory perception and olfactory perception are essential cues to cognize it. Likewise, machine can make use of multiple types of information format (modality) of the video involving visual and sound to understand the content. Different modalities are mapped into a same dimensional space and interacted to improve the performance. Thus, it is a significant assignment concerning how to effectively fuse various modalities. In general, fusion strategy can be divided into three types in accordance with the position it fused – early fusion, intermediate fusion and late fusion.

- **Early fusion:** Multiple modalities fused at the shallower layers or before the model can be seen as early fusion. This is a simple and convenient approach to integrate multimodal information into a one-stream network and acquire the global context over all modalities. However, local context in each modality may not be explored explicitly and it is not able to carry out certain one

modality specifically for extra requirements if early fused multiple modalities.

- **Intermediate fusion:** It is also termed middle fusion when modalities are integrated in the middle of the network. Each modality is processed at each stream in advance and later fused before making final decision or prediction. Similar to early fusion, intermediate fusion also cannot capture the complex interactions and relationships between different modalities.

- **Late fusion:** Multimodal representation is combined at the deeper or last layer of a deep neural network for late fusion strategy. The internal context or each modality or the interacted corelation across different modality are fully explored. Nevertheless, from the perspective of information integrity, the global context is ignored, and model cannot have a comprehensive cognition on all modalities.

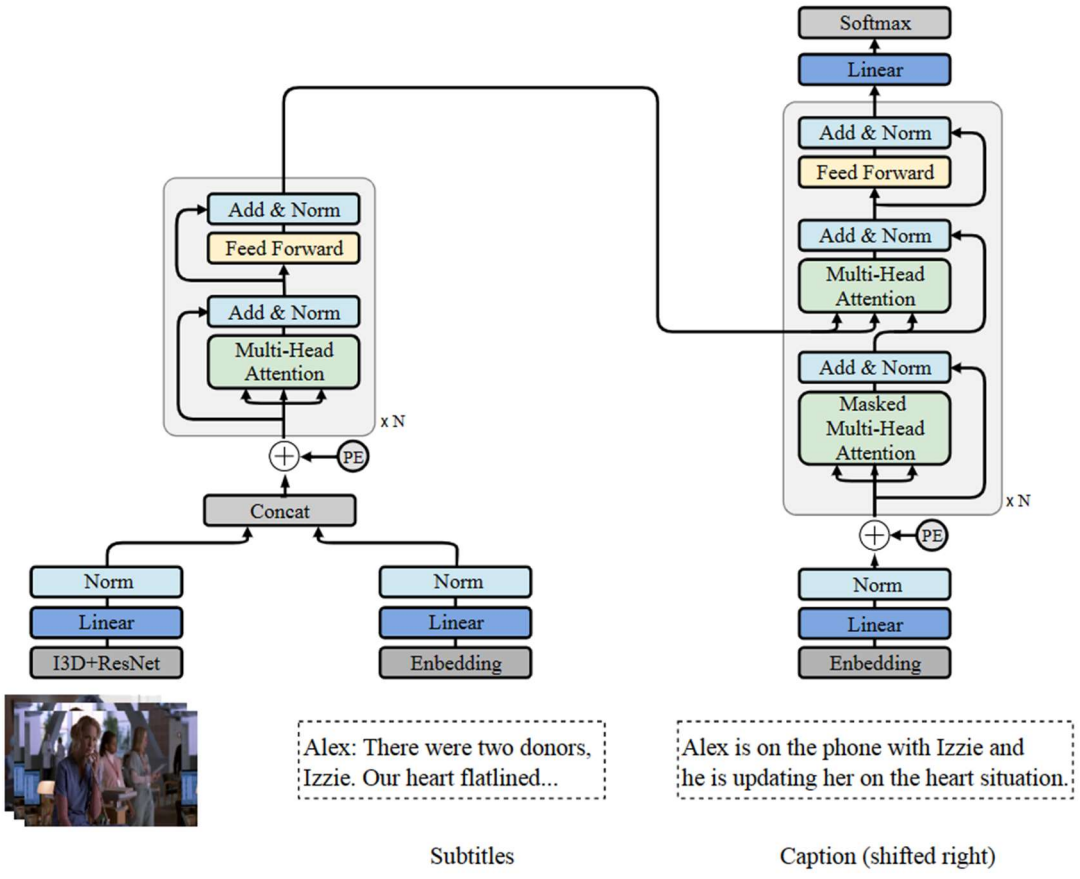

Fig. 5. Overview of the Multimodal Transformer model [5].

As video is a natural multimodal combination involving images, audio and text, multimodal learning came to the view of video captioning research. Lei et al. [5] propose a Multimodal Transformer (MMT) to tackle the captioning task on the TVC dataset. They adopt a vanilla Transformer encoder-decoder architecture to deal with the task as Fig.5 demonstrated. The appearance feature of frame and motion feature of video is extracted by I3D [20] and ResNet-152 [18], respectively. They concatenate and normalize these two features beforehand and concatenate again with embedded word tokens before inputting into the encoder. However, the concatenation of multimodal representation makes the input longer and self-attention computation in the encoder will suffer a high computational complexity. Additionally, it only captures the limited local context in each individual modality.

Therefore, inspired by the multimodal learning and characteristic of the Transformer network, we made a series of improvements to the MMT by focusing on reducing the computational cost and acquiring more contextual understanding for the subtitle in TVC dataset.

2.2 Benchmark Datasets

Apart from the model's architecture, dataset is another one of the most important components in the field of artificial intelligence. A constructed model needs to be trained and validated on a large scale of dataset and then used to handle relevant tasks in real life. That means the quality and structure of dataset determines the performance of the model. To better understand the application in the field of video captioning and how our model works, it is highly necessary to present relevant commonly used datasets – MSVD [21], MSR-VTT [22] and TVC.

2.2.1 MSVD

The Microsoft Research Video Description Corpus (short for MSVD) is introduced

by David et al. [21] to narrow the gap between vision and language. This video consists of 1970 video segments derived from YouTube videos and 70,028 English descriptions. Each video segment is on the length of 4 -10 seconds and paired with 40 human-annotated English sentences provided by Amazon Mechanical Turk (AMT) workers. The video segments cover various categories but each one only has one main action or events to avoid unambiguous description. Besides, the audio, subtitles or other text in the video is muted to prevent annotators from choosing biasing lexical words in the description.

2.2.2 MSR-VTT

Xu et al. [22] argues that the primary problems existed in most current benchmarks are specific fine-grained domains with limited sized of videos and simple descriptions [22]. Accordingly, they present a new large-scale video dataset MSR-VTT (MSR-Video to Text) that has a comprehensive list of 20 categories videos ranging from music to ads. The whole dataset is composed of 7,180 videos collected from top 150 video search results on a commercial video search engine based on 257 representative queries. Each video clip has average 20 seconds and 20 natural sentences manually annotated by AMT workers. They also give a guidance on the split ratio for training and validating the video captioning models that are 6513, 2990, 497 clips in the training, testing and validation set, respectively. Compared with MSVD, MSR-VTT has a larger number of videos and more complicate sentences.

2.2.3 TVC

Lei et al. also introduces a novel multimodal video captioning dataset TVC (standing for “TV show Caption”) to associate the linguistic representation with visual content. They choose the TV shows as the resource of the dataset since the drama shows usually contain more intricate interactions between actors and dynamic contents [5]. The dataset is composed of 108K video moments over 6 diverse genes and paired with 262K descriptions.

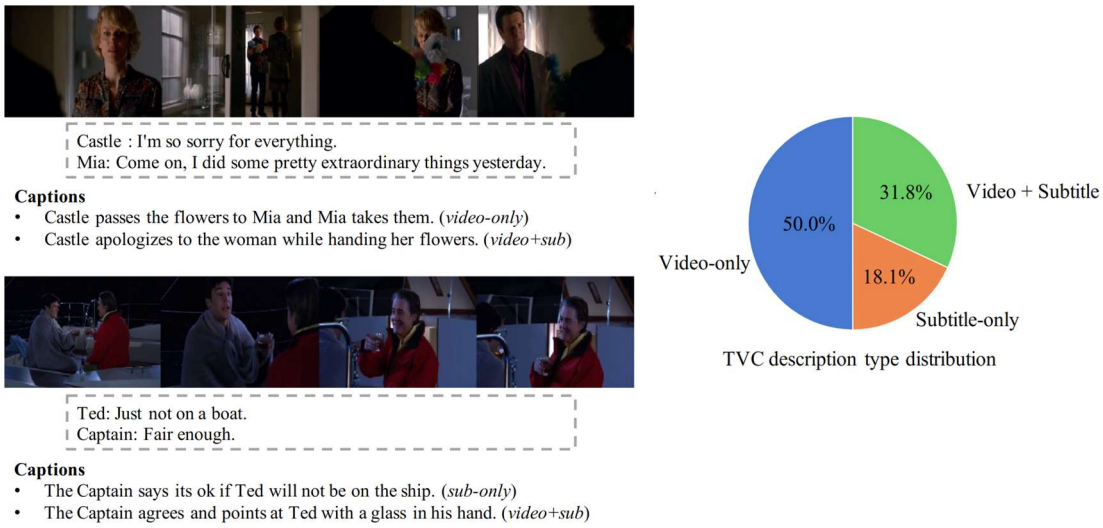

Fig. 6. Two samples of TVC caption descriptions and description type distribution [5].

One sample has three components – video moment, subtitle and captions. The video moment has an average length of 9 seconds and is associated with subtitles. There are 2 video descriptions for each video in training set and 4 descriptions for the video in the validation and test set. Each caption is followed by a phrase which refers that it is annotated by only video content, subtitle or both. The description type distribution in TVC dataset is showed in Fig. 6. There is a half description annotated based on video only and around one thirds annotated based on video and subtitle.

2.3 Evaluation Metrics

Using metric algorithms is an objective and fair way to evaluate the performance of the model. It can evaluate the model's ability to implement the correct operation in terms of different aspects. The four metrics including BLEU [23], METEOR [24], ROUGE-L [25] and CIDEr-D [26] are used to claim if the generated sentence is related to the ground-truth, which refers the generation is appropriate if two sentences are similar.

2.3.1 BLEU

It is expensive to measure the similarity between human outputs and outputs generated by machines relying on human's judgement. Thus Papineni et al. introduce the first automatic machine translation evaluation named Bilingual Evaluation Understudy (BLEU). It evaluates the quality of generated prediction by computing the ratio of occurrence of exactly matched n-gram words as well as their order in any ground-truth sentence. It also has a penalty function to avoid too short generation. BLEU-4 is precision oriented and the most frequently utilized by focusing the 4-gram words.

2.3.2 METEOR

Since BLEU only focus on the precision of generated words occurred in references which is limited for similarity judgement, Metric for Evaluation of Translation with Explicit Ordering (METEOR) is proposed by Banerjee et al. to address BLEU's weakness with considering the precision, recall and penalty function to prevent from extreme situation. Besides the exact lexical match of words, METEOR also judges the sentences on stemming and synonyms. The METEOR score is computed by the harmonic mean of unigram precision and recall with most of weights on recall. The result is presented as:

$$\text{Score} = \frac{10PR}{R + 9P} * (1 - Penalty) \tag{2}$$

where *P* and *R* refers unigram precision and recall, respectively. The penalty function is computed as illustrated in (3)

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched}\right)^3 \tag{3}$$

where *#chunks* is the number of consecutive words occurred in both generation and ground-truth, *#unigrams_matched* is the number of matched unigram of generation in references. Consequently, the generation will get lower score on penalty if #chunks is

lower and we say it is more natural and similar.

2.3.3 ROUGE-L

ROUGE [25] is an another commonly used evaluation metric in the field of video captioning and its full name is Recall-Oriented Understudy for Gisting Evaluation. It calculates the score by comparing the model-generated sentence with human-written sentences to claim the similarity, which is pretty like the BLEU metric. But BLEU places more emphasis on precision and ROUGE is based on the recall value, in another words, ROUGE computes how many numbers of n-gram in ground-truth is generated in machine's output sentence. ROUGE-L is a variant of ROUGE, and it measures the similarity of the longest matching sequence of words using Longest Common Subsequence (LCS) [25]. Compared with n-gram-based BLEU and METEOR, ROUGE-L only concentrates on the matching of longest consecutive string and does not need to define the number of *n* beforehand.

2.3.4 CIDEr

CIDEr is specifically designed for image caption problems. This metric treats each sentence as a "document" and represents it in the form of a Term Frequency Inverse Document Frequency (TF-IDF) vector [26]. It calculates the cosine similarity between reference captions and the model-generated caption by calculating the TF-IDF weight of each n-gram, to measure the consistency of captions. Consequently, it is a weighted evaluation metric and can focus on the key point in the sentence, in another word, it measures how natural the sentence is in relation to the human's expression.

Chapter 3 Proposed Method and Variants

3.1 Proposed method

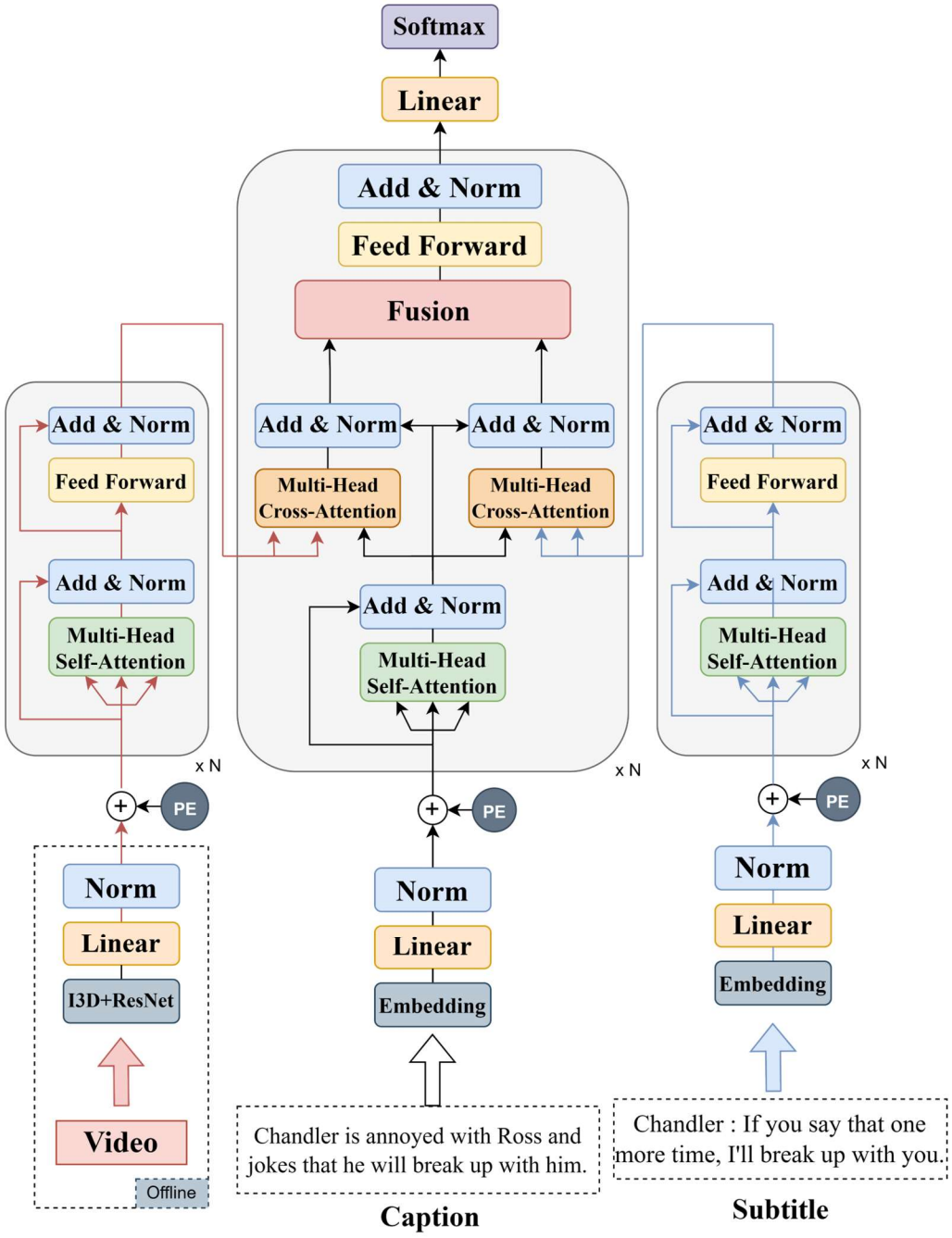

Fig. 7. Overview of our proposed method *2streams_dec*.

Motivated by the baseline on TVC dataset – MMT, we investigate the Transformer network involving multimodal learning and propose a new method to tackle the video

caption generation task. Our model first encodes the video feature and subtitle in different encoders with different layers, then the model perceives the internal interacted information of video-caption and subtitle-caption. Two sequences are integrated by summation in fusion block and subsequently transported into next decoder layer.

The overview of our model is shown in Fig. 7, we build up two individual encoders to encode the video feature and subtitle. We found that the video feature is usually obtained from the deep layer of a pretrained deep neural network and then embedded into the network, but the text is directly embedded into it after tokenization. That means the extracted video feature has more rich information than the text. If they are simply concatenated before the model and then go through the same encoder as the MMT, the model will be hard to detect the local context in the subtitle sequence and not be able to keep balance on exploring contextual information between the video and subtitle. Hence, we decide to split one encoder into two different encoders and handle the different type of input. At here, video encoder is composed of a stack of 2 identical layers and the subtitle encoder has 4 identical layers in order to achieve more internal contextual understanding.

In another hand, as we discuss in section 2.1.2 and 2.1.3, the self-attention mechanism will suffer a higher computational cost for a concatenated sequence since it highly depends on the length of input sequence. Given an input sequence A and another input sequence B which has the same dimension d and length n, the cost will turn to $O((n_A + n_B)^2 \cdot d)$ if two sequence are concatenated before the encoder as MMT does. However, two sequences are input into different encoders and the self-attention calculation is computed on the only one sequence. Eventually, the overall computational complexity is reduced to $O((n_A^2 + n_B^2) \cdot d)$ in our model.

During the training phase, appearance feature and motion feature of the video is pre-processed while captions and subtitles are tokenized into tokens with <BOS> and <EOS> tag attached at the beginning and end, respectively. Video features and text

tokens are mapped into a same dimensional embedding space and added the position to each token using sine and cosine functions. For two encoders, video feature and subtitle tokens are handled separately to achieve internal context representation. For the decoder, since the future token is predicted by knowing its previous tokens, the future tokens in input sequence is masked at each step in order to prevent the model of knowing the predicted tokens. The cross-attentions of video-caption and subtitle-caption are obtained by interacting with outputs from two encoders. Then two cross-attentions are combined and further transported to next decoder layer. The whole model is trained using Maximum Likelihood Estimation (MLE). Given the video *V* and subtitle *S*, we denote the model's generation as *caption* that is composed of a series of words *w*, hence the model is to maximum each word's log likelihood to generate a precise and natural sentence as shown in (4)

$$Caption = argmax_{\theta} \sum_{t=1}^{n} logP(w_t \mid V, S\ ;\ \theta) \qquad (4)$$

During the inference phase, after acquiring the encoder's output, the decoder starts to decode the <BOS> tag and predict next word on the basis of the encoded context. Then the generated word is shifted into the decoder and predict next word, it will stop generating till outputting the <EOS> tag.

3.2 Model Variants

We also proposed extra models to investigate the effect of inter-modal and the intra-modal features over video and subtitles.

3.2.1 2to1stream_selfAttn

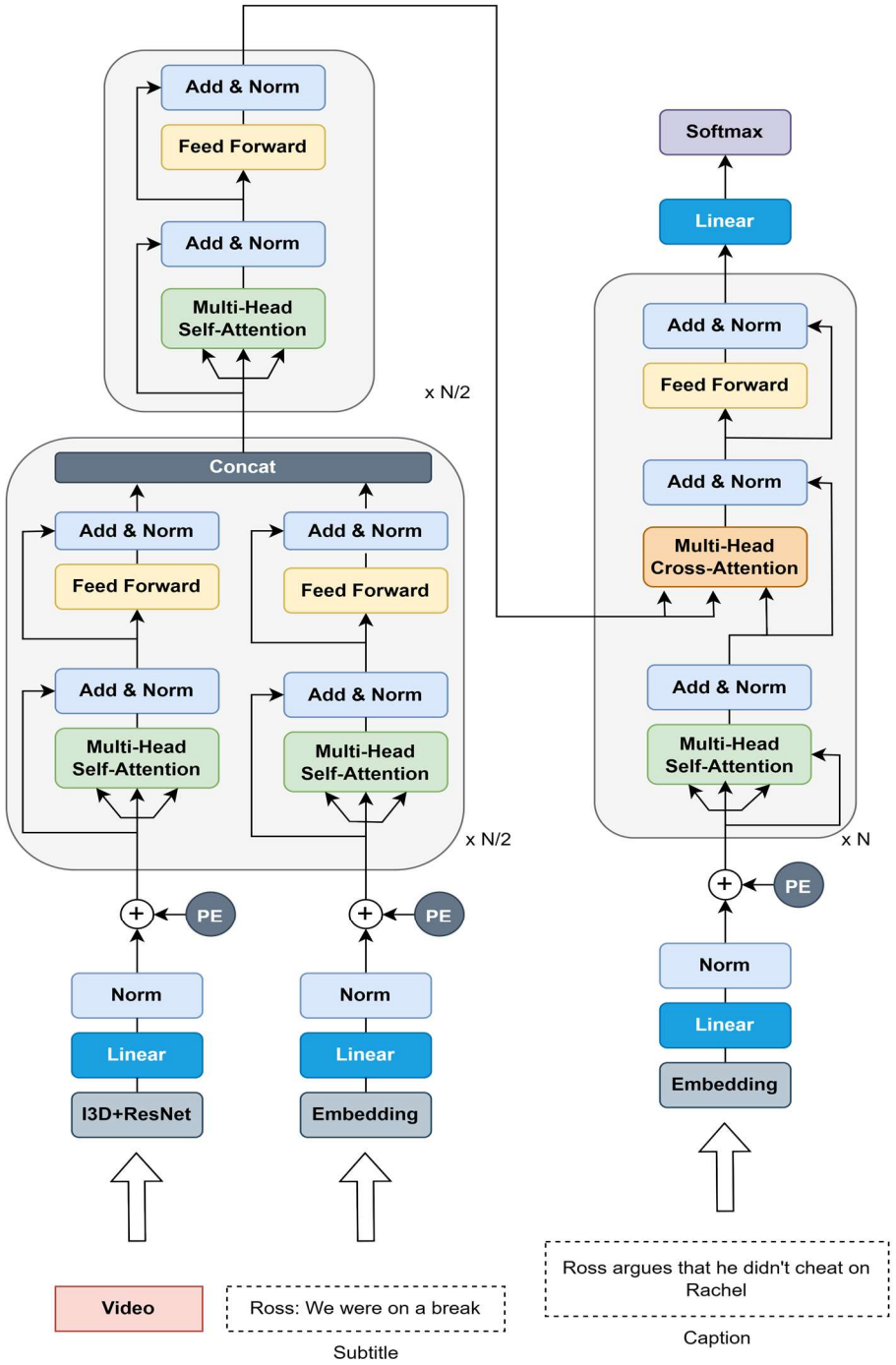

Fig. 8. Overview of variant #1 *2to1stream_selfAttn*.

We design a variant model to investigate the consequence of global context over visual modality and subtitle. Two modalities are processed in the individual stream to explore the local context and then concatenated at the intermediate layer of the encoder.

It is further computed by the self-attention to acquire the global context. This model shares same hyper-parameters with the proposed model. It has one layer for former part and latter part in encoder and two layers in decoder. Only self-attention is computed in the network. This model can be seen as intermediate fusion based multimodal transformer.

3.2.2 2streams_selfAttn

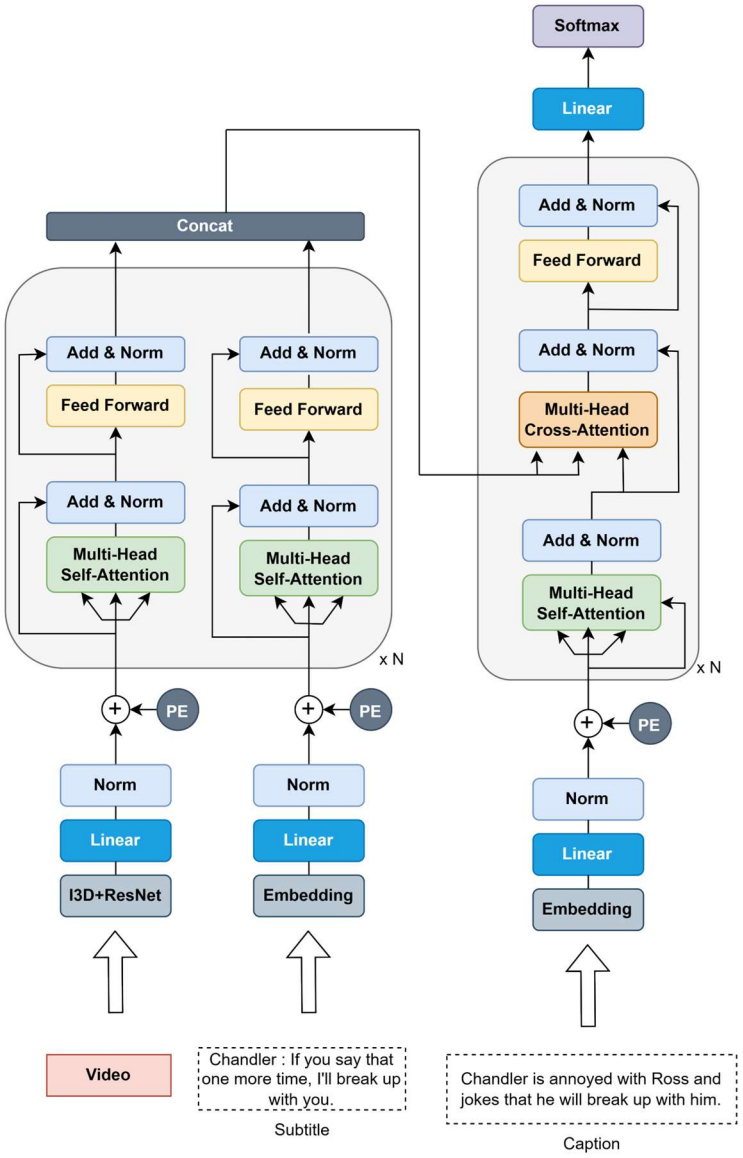

Fig. 9. Overview of variant #2 *2streams_selfAttn*.

This model is similar to *2to1stream_selfAttn* but we discard the latter part of its encoder. Only the local context in each modality is considered and they are

concatenated at the end of encoder. Compared with *2to1stream_selfAttn*, we design this model for exploring the effect of the global context on the performance. This model has two layers for encoder and decoder, and it can be termed intermediate fusion based or special late fusion based.

3.2.3 2to1stream_crossAttn

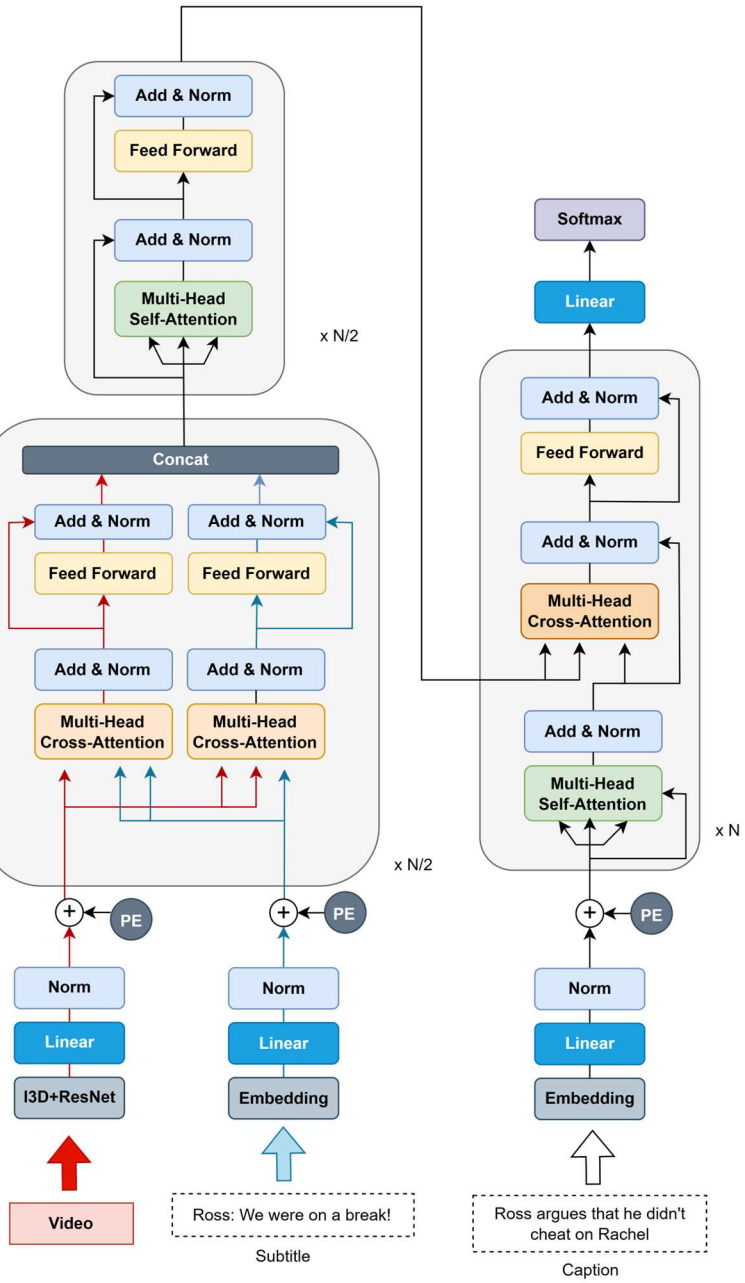

Fig. 10. Overview of variant #3 *2to1stream_crossAttn*.

As there are two modalities – video and subtitle in the TVC dataset, we force the

video feature and subtitle tokens know each other to acquire the interaction context between them. Same to *2to1stream_selfAttn*, we also concatenate two modalities at the intermediate layer and further compute the self-attention on the concatenated sequence to acquire the global context.

3.2.4 2streams_crossAttn

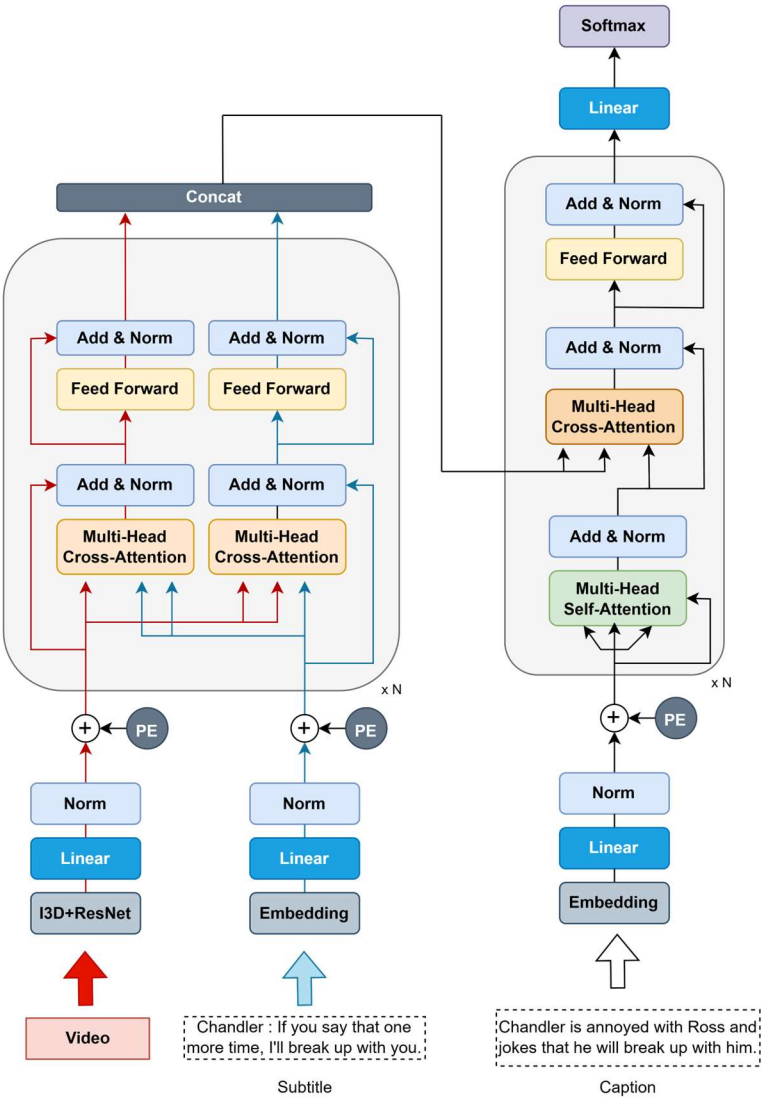

Fig. 11. Overview of variant #4 *2streams_crossAttn*.

We also remove the global context computation component as did for *2streams_selfAttn*. The internal interacted context is explored explicitly and further concatenated at the end of encoder. It shares same hyper-parameters with other variants.

Chapter 4 Experiments

4.1 Experiment details

To have a fair and equitable comparison with the baseline MMT, we implement exactly same process to the video features and texts as MMT does [5]. At here, the 2048D appearance features are extracted by ResNet-152 pretrained on ImageNet [27] at FPS3 and then max-pooled every 1.5seconds to obtain the clip-level feature. Likewise, we extract 1024D motion features by I3D that is pretrained on Kinetics-600 [28] for action recognition. We then concatenate two types of video features and apply the L2-normalization. As for the subtitle and caption, we utilize the GloVe [29] to tokenize the sentence into tokens and embed them into 300D vectors. Finally, all features are projected into the same embedding space using linear layers and layernorm [30] layers. These operations are all implemented and stored on the device beforehand. They are taken as the input into the model during the training and inference phase.

We implement our experiments involving ablation experiments and reproduce the MMT on the Nvidia GeForce GTX Titan X with 11GB memory and . Codes are written in PyTorch 1.1.0 and executed under the OS of Ubuntu 20.04.

4.2 Experiment result

Table 1. Result comparison of the proposed method and baseline MMT on *val* set.

	B@4 ↑	**METEOR ↑**	**ROUGE-L ↑**	**CIDEr-D ↑**
MMT (Paper)	10.53	**16.61**	32.35	44.39
MMT (Reproduce)	10.56	16.53	32.27	44.18
Proposed method (2streams_dec)	**11.19**	16.54	**32.67**	**45.64**

Results of MMT and our proposed model is revealed in Table 1, the “paper” in the

second line means the values of four metrics are from original paper [5]. We reproduce the MMT and find the results are slightly reduced, except the BLEU-4 metric, when comparing to the results of paper. We think that this is due to the differences in hardware and experimental environment.

As we can see from the Table. 1, our proposed method outperforms the baseline model across BLEU-4, ROUGE-L and CIDEr-D, especially the value of CIDEr-D is increased by 1.46 comparing to the reproduce one. This indicates that our model can generate more natural output than the baseline. Although our model's result on the METEOR is slightly lower than the baseline, we think it is still acceptable as the METEOR metric only proves that the generated output from our model has fewer words that occurred in the ground-truth.

Sheldon : I made more accurate diagrams of the expansion of the early universe on the nursery wall with the contents of my diaper.
Amy : Are you getting sick?

Ground-truth:

- Amy opens the door and Sheldon closes it. (*video-only*)
- Amy takes her bag off and Sheldon complains about the lecture. (*video-subtitle*)

Baseline: sheldon opens the door and enters the room.
Ours: sheldon walks into the living room and closes the door.

Fig. 12. A sample of the result comparison.

A sample result is presented in Fig. 12. Sentences in the dashed box are subtitles in that video moment. With reference to the video and ground-truth, we can see that the baseline gives a wrong judgement on who "opens the door", whereas our model is able

to be aware of the person who “closed the door”.

4.3 Ablation experiment

As we claim that the subtitle encoder requires extra layers to handle the linguistic presentation, we conduct a controlled experiment to determine the optimal number of subtitle encoder layers that has best results. Moreover, several experiments are implemented to research the impact of global context and interacted attention across video stream and subtitle stream.

4.3.1 Subtitle encoder layers

We test our proposed model with different number of subtitle encoder layers and other hypha-parameters are fixed. The line charts of each metric result are shown as below:

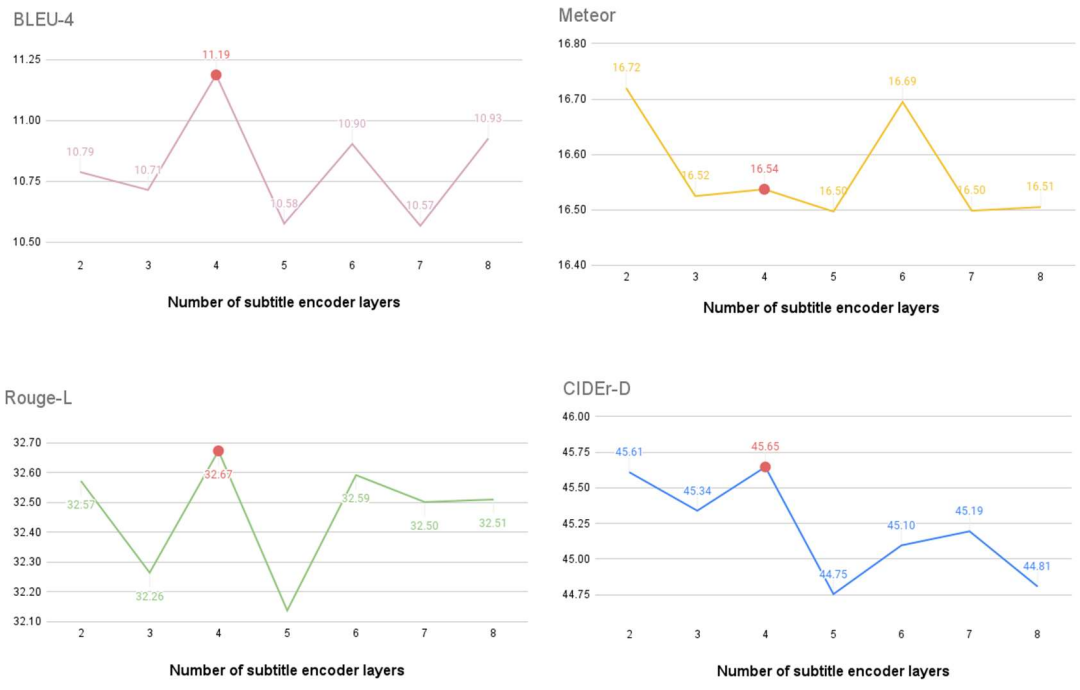

Fig. 13. Line charts of each metric regarding the number of subtitle encoder layers.

We found that the model with 4 layers in subtitle encoder has the highest results across BLEU-4, ROUGE-L and CIDEr-D in Fig. 13. It means that our model can generate a more natural, correct and longer sentence than the baseline. Moreover, the

value of these three metrics is getting decreased along the y-axis after the amount of subtitle encoder layers is 4. Furthermore, our model with 4 layers still has a competitive result as it is the third highest value on the METEOR metric.

Additionally, when the amount of subtitle encoder layers is 6, our model has a competitive result across BLEU-4, METEOR and ROUGE-L. Even so, it fails to achieve a higher value on CIDEr-D which is the most important metric to evaluate the performance. Therefore, we choose 4 as the number of subtitle encoder layers for our proposed model.

4.3.2 Results of variants

We also tested four variants to investigate the effects of global context and cross-attention on two modalities. All results are shown as below:

Table 2. Results of all variants and proposed model with 2 or 4 subtitle encoder layers.

	B@4 ↑	**METEOR ↑**	**ROUGE-L ↑**	**CIDEr-D ↑**
MMT (paper)	10.53	16.61	32.35	44.39
MMT (reproduce)	10.56	16.53	32.27	44.18
Variant#1 (2to1stream_selfAttn)	10.66	**16.78**	32.55	45.25
Variant#2 (2streams_selfAttn)	10.64	16.69	<u>32.58</u>	**45.90**
Variant#3 (2to1stream_crossAttn)	<u>11.09</u>	16.60	32.54	45.24
Variant#4 (2streams_crossAttn)	10.52	16.47	32.17	43.71
Proposed model (2 layers)	10.78	<u>16.72</u>	32.57	45.61
Proposed model (4 layers)	**11.19**	16.54	**32.67**	<u>45.64</u>

Results of four variants and proposed model are displayed in Table 2. The number in bold denotes the highest value in the column and the number with underscore denotes the second highest value in the column. In addition, the phrase in the parenthesis of last two lines denotes the number of subtitle encoder layers in the proposed model.

We found that our proposed model with 4 layers for subtitle encoder still has a competitive result comparing other variants and baseline from Table 2. In addition, the variant with only computing self-attention in each modality performs better than the one with computing the cross-attention. It confirms that there is no strong correlation between the video moment and associated subtitles. We think there are two major reasons: (1) the subtitle is the line from actors, and it usually has a context information

with previous or future video segment. The referred person or event in the subtitle is not guaranteed to appear in the current associated video moment. (2) as Fig. 6 illustrated, there is only one thirds captions in the dataset are annotated based on the video and subtitle. Thus, the model is not able to learn the corelation between video and subtitle during the training. Furthermore, we also found that there is almost no difference between the model that considers the global context and the one that does not. Especially when comparing the result of the models computing self-attention, the increased values are lower than 0.1 if the global context is considered.

Chapter 5 Conclusion and future works

5.1 Conclusion

In this thesis we investigated the effect of different fusion strategies of a multimodal Transformer in the field of video captioning. Start from the background of video captioning, we introduced the difference of single sentence captioning and dense video captioning. We also provided a general framework to display how the model generates the prediction. Furthermore, two crucial issues and several related works on video captioning were discussed, which motivates us to propose our method.

Two problems that urgently need to be solved are to alleviate the high computational complexity for longer input sequence during computing the self-attention and specifically encode the linguistic sequence individually to obtain more interior context. Hence, we proposed our method to tackle two problems by improving the MMT. We also designed additional four variants to see the impact of global context and interacted representation between video and subtitle.

With respect to the result of our proposed method, we found that it has a better performance if handling two modalities individually than handling in a same stream. Especially, the model achieves the best result across most of metrics when it has 4 layers in the subtitle encoder. Regarding the results of four variants, we also found that model with exploring local context only is better than the one that does not. We think the reason is that there is no strong corelation between the current video segment and associated subtitles. In addition, there is an insufficient amount of ground-truth annotated according to video and subtitle in the TVC dataset, which prevents the model from being able to learn the interior context.

5.2 Future works

Even though our model performs better than the baseline on the TVC dataset, there is still room for improvement, and it is worth continuing to explore ways to enhance the model through further research.

Firstly, due to the limited computational power of graphic card and video memory, we cannot finetune the backbone that extracts the video features and train our Transformer network simultaneously. The backbone can capture the corresponding feature and generator can output more precise description if applicable.

Secondly, since the huge success has been achieved by image transformer [31] and video transformer [32] in the field of computer vision, it is possible to break a video into multiple tubes as Vivit [32] does, subsequently it can encode the video tubes instead of extracted feature to attain more context.

Thirdly, the larger vocabulary size a tokenizer has, the ranger variety of words the model can understand. Therefore, if we could change a better tokenizer that has a larger vocabulary size, the result of video captioning method can be potentially improved.

Lastly, we only tested the summation for the fusion module in the decoder, however, not all features from the video segment and subtitle are useful and meaningful to generate description. Hence, a more natural output can be generated if there is an adaptive fusion strategy which can “teach” the model when and how to combine all modalities

Bibliography

[1] A. Kojima, T. Tamura, and K. Fukunaga, “Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions,” *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 171–184, Nov. 2002.

[2] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, “Jointly Modeling Embedding and Translation to Bridge Video and Language,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4594–4602.

[3] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to sequence-video to text,” in *Proc. IEEE international conference on computer vision*, 2015, pp. 4534–4542.

[4] V. Lashin and E. Rahtu, “Multi-modal Dense Video Captioning,” in *Conf. on Comput. Vis. Pattern Recognit. Workshops,* 2020, pp. 4117–4126.

[5] J. Lei, L. Yu, T. L. Berg, and M. Bansal, “TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval,” in *Eur. Conf. Comput. Vis.,* 2020, pp. 447–463.

[6] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, “HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips,” In *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2630–2640.

[7] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville and B. Schiele “Movie Description,” *Int. Journal Comput. Vis..*, vol. 123, no. 1, pp. 94–120, May 2017.

[8] R. Radarapu, A. S. S. Gopal, M. NH, and A. K. M., “Video summarization and captioning using dynamic mode decomposition for surveillance,” *Int. Journal Inf. Tech.*, vol. 13, no. 5, pp. 1927–1936, Oct. 2021.

[9] V. Iashin and E. Rahtu, “A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer,” *arXiv preprint arXiv:2005.08271v2,* May 2020.

[10] M. Chen, Y. Li, Z. Zhang, and S. Huang, “TVT: Two-View Transformer Network for Video Captioning,” in *Pro. Asian Conf. Machine Learning*, 2018, pp. 847–862.

[11] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Comm. of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[13] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *CoRR*, 2014.

[14] G. Gkioxari and J. Malik, “Finding action tubes,” in *2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 759–768.

[15] K. Soomro, A. Zamir, and M. Shah, “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild,” *arXiv preprint arXiv:1212.0402*, Dec. 2012

[16] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser and I. Polosukhin: “Attention Is All Your Need”, In *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp.6000- 6010.

[17] L. Zhou, Y. Zhou, J. Corso, R. Socher, and C. Xiong, “End-to-End Dense Video Captioning with Masked Transformer,” in *Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8739–8748.

[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Pro. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[19] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Int. Conf. Machine learning*, 2015, pp. 448–456.

[20] J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset.” In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2017, pp. 6299-6308.

[21] D. Chen and W. Dolan, “Collecting Highly Parallel Data for Paraphrase Evaluation,” in *Pro. Annual Meeting Association Computational Linguistics: Human Language Technologies*, 2011, pp. 190–200.

[22] J. Xu, T. Mei, T. Yao, and Y. Rui, “MSR-VTT: A Large Video Description Dataset for Bridging Video and Language,” In *Pro. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5288-5296.

[23] P. Kishore, S. Roukos, T. Ward, and W. Zhu. "Bleu: a method for automatic evaluation of machine translation." In *Proc. of the 40th annual meeting of the Association for Computational Linguistics*, 2022, pp. 311-318.

[24] S. Banerjee and A. Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[25] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Text Summarization Branches Out*, 2004, pp. 74–81.

[26] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based Image Description Evaluation.” In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566-4575.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2009, pp. 248–255.

[28] K. Will, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman and A. Zisserman "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950*, May 2017

[29] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” in *Proc. of the 2014 Conf. Empirical Methods in Natural Lang. Processing*, 2014, pp. 1532–1543.

[30] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer Normalization,” *arXiv preprint arXiv:1607.06450*, Jul. 2016

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhuai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby “An image is worth 16x16 words: Transformers for image recognition at scale,” *ArXiv Prepr. ArXiv201011929*, Oct. 2020.

[32] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, “Vivit: A video vision transformer,” in *Proc. IEEE Int. Conf. Comput. Vis,* 2021, pp. 6836–6846.

[33] S. Chen, T. Yao and Y.G. Jiang, “Deep Learning for Video Captioning: A Review”, in *Int. Joint Conf. Artif. Intell.*, 2019, pp. 6283-6290.