공학석사 학위논문

Co-attentional Transformers for Video Story Understanding

비디오 스토리 이해를 위한 공동 주의 트랜스포머

2021년 8월

서울대학교 대학원

컴퓨터공학전공

Björn Bebensee

Co-attentional Transformers for Video Story Understanding

비디오 스토리 이해를 위한 공동 주의 트랜스포머

지도교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함

2021 년 6 월

서울대학교 대학원

컴퓨터공학전공

Björn Bebensee

Björn Bebensee의 공학석사 학위논문을 인준함

2021 년 7 월

위 원 장 _____ 김 선

부위원장 _____ 장 병 탁

위 원 _____ 황 승 원

Abstract

Co-attentional Transformers for Video Story Understanding

Björn Bebensee
Computer Science and Engineering
The Graduate School
Seoul National University

Inspired by recent trends in vision and language learning, we explore the application of co-attention mechanisms for visiolingual fusion within an application of video story understanding. Like other video question answering (QA) tasks, video story understanding requires agents to grasp complex temporal dependencies. However, as it focuses on the narrative aspect of video it also requires understanding of the interactions between different characters, as well as their actions and their motivations.

In this thesis we introduce essential concepts from natural language processing (e.g. multi-head attention) and carry out a comprehensive survey of relevant work from adjacent fields such as visual question answering, visiolingual representation learning, video representation learning and video question answering. Based on our findings we propose a novel co-attentional Transformer model to better capture long-term dependencies seen in visual stories such as dramas and measure its performance on the video story understanding task in a video question answering setting.

We evaluate our approach on the recently introduced DramaQA dataset which features character-centered video story understanding questions. Our model outperforms the baseline model by 6 percentage points in overall accuracy and at least 3.8 and up to 12.1 percentage points in accuracy on all difficulty levels and manages to beat the winner of the DramaQA challenge.

Keywords: Video story understanding, co-attention, video question answering, multi-modal learning

Student Number: 2019-21343

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Both computer vision and natural language processing have seen several breakthroughs in recent years leading to large progress in tasks combining both these modalities such as image retrieval, image captioning and visual question answering. In particular, visio-lingual representation learning has made great progress benefiting these downstream tasks (Chen et al., 2020b, Huang et al., 2020b, Li et al., 2020, Lu et al., 2019, Su et al., 2020, Tan and Bansal, 2019). We will give a broad overview over vision and language learning and a comprehensive survey of recent works addressing visual question answering, video question answering as well as visiolingual representation learning in Chapter 2.

In visual question answering (VQA) an agent is provided an image along with a natural language question about the image and should provide the correct answer (Antol et al., 2015). This kind of multimodal question answering setting requires the agent to resolve cross-modal references and selectively extract information from relevant areas of the image. Video question answering is an extension of visual question answering in the temporal domain; rather than a single image the agent should answer a question about a video which can be limited to a sequence of frames but also include subtitles or audio. As such, video question answering adds this additional dimension to an already challenging problem and has thus received considerably less attention. Not only

should models learn to utilize contextual information and references between the vision and language input but also to perform multi-step and long-term reasoning along the temporal axis.

Broadly speaking, an agent is presented with a video clip of a scene and, like in visual question answering, has to infer the correct answer to a given question in natural language but depending on the exact setting the difficulty can vary. The given video clip can consist either of a single shot or multiple shots from different angles or in different locations. While questions can be relatively simple, e.g. "What is the woman holding?", they can also be far more complex and require deeper understanding and multiple steps of reasoning, e.g. "Why is the man in the overalls angry at the cyclist?". Due to these temporal dependencies that need to be resolved and understood in order to answer more complex questions about the scene correctly, video question answering has remained a very challenging problem.

One such dataset for video question answering is the TVQA dataset which is build around short 60 to 90 seconds long video clips and questions bridging vision and language clues (Lei et al., 2018). Agents have to infer the answers by using multiple modalities (video frames, subtitle-based dialogue) as well as temporally localize the relevant part of the video. TVQA+ adds additional bounding boxes and objects annotations that link them directly to visual concepts mentioned in questions and answers (Lei et al., 2020). While the TVQA(+) dataset seems to be a popular choice for evaluation of video question answering and video understanding models (Geng et al., 2020, Kim et al., 2019, Yang et al., 2020), it does not require story-level understanding. Most questions in the dataset only require the agent to attend to a short part of the video clip (15 seconds or less) due to its particular focus on temporal localization.

In this work we instead choose to focus on story-based video understanding

for a deeper understanding of long-term dependencies and characters' actions and intentions. In order to fuse vision and language in a meaningful way, we adopt a two-stream co-attentional Transformer module inspired by recent work in visual dialog (Nguyen et al., 2020) and vision-language representation learning (Lu et al., 2019, Tan and Bansal, 2019). We evaluate our approach on the recent DramaQA dataset which aims to benchmark exactly this type of video story understanding and focuses in particular on story-level questions that are closely centered around the narrative and characters of a TV drama along with character-level annotations. Unlike in TVQA, questions focus on longer-range character interactions and aim to capture story understanding on a deeper level at both the shot and the scene level. Moreover, DramaQA allows for evaluation by difficulty level and can therefore give us a better understanding of where the strengths and weaknesses of our method lie.

The remainder of the work is structured in the following way. In Chapter 2 we introduce prerequisites necessary to understand our proposed method and perform a comprehensive survey of methods related to vision and language learning including visual question answering, video question answering and vision and language representation learning. In Chapter 3 we give a general overview over the DramaQA dataset, the key differences between video question answering and video story understanding, formulate the problem setting and introduce several baseline models that we compare our approach to. We introduce the two-stream co-attentional Transformer architecture for video story understanding and describe it in detail in Chapter 4. In Chapter 5 we experimentally evaluate our approach by comparing it to several baseline methods, the winners of the "DramaQA challenge" which was held at ECCV 2020, perform an ablation study along with several experiments to gain a deeper understanding of how our method works, and provide several qualitative examples. Finally we

give an outlook on future work in Chapter 6.

We will first review work most closely related to our approach, introduce the evaluation dataset, and finally our model architecture and experimental results.

Chapter 2

Related Work and Preliminaries

In this chapter we perform a comprehensive survey of related work in the field of vision and language learning. We look at advances in natural language processing (NLP) from recent years which have largely driven progress in research at the intersection of vision and language. Starting with attention mechanisms and the Transformer architecture we discuss related lines of research in question and answering in NLP, visual question answering and visual dialog, vision and language representation learning and more specifically adaptions of these methods to video representation learning, and finally work most closely related to ours in the field of video question answering.

2.1 Language models, Transformers and Question Answering

In order to understand many of the ideas from vision and language learning and the context in which they lie it is first necessary to understand some of the concepts that have enabled the great strides forward NLP has made in recent years.

At the core of many of the advances in NLP, vision and tasks at the intersection of the two lies the mechanism of *attention*. Given some query vector x, attention layers essentially retrieve the relevant information contained in a set

of context vectors $\{y_j\}$ (Bahdanau et al., 2015). We can for instance imagine a translation task in which we would like the resulting translation words to contain the same semantic information as their counterparts in another language, i.e. the model should pay attention to the appropriate parts of the source sentence during translation. In the case of image captioning we might be given some context $\{y_j\}$ consisting of image regions or object features and want the language generation module to retrieve different relevant contexts from these features for each word generated in the caption (Xu et al., 2015).

In the case of visual question answering (see Section 2.2) we are given a question which can be thought of as being represented by some query vector x and want the model to retrieve the appropriate context y formed by e.g. image regions or objects in the image $\{y_j\}$ (Anderson et al., 2018).

Following Bahdanau et al. (2015) attention scores are computed as a matching score e_j between the query x and the context y_j and then normalized.

$$\begin{aligned} e_j &= \text{score}(x, y_j) \\ \alpha_j &= \frac{\exp(e_j)}{\sum_k \exp(e_k)} \end{aligned} \tag{2.1}$$

Finally, the output context vector c is computed as the weighted sum of the context vectors:

$$c = \sum_j \alpha_j y_j \tag{2.2}$$

An attention layer which computes the attention of a sequence $\{x_i\}$ to itself, i.e. each query vector x_i is also in the set of context vectors $\{y_j\}$, is called a *self-attention* layer.

More formally, following Vaswani et al. (2017) we can write an attention function as a mapping of a query and a set of key-value pairs to an output, that

is, we preferentially retrieve values of matching queries and keys. This kind of scaled dot-product attention can be written as

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{(d_k)}})V \tag{2.3}$$

for a query matrix Q, key matrix K, value matrix V and d_k being the dimension of queries and keys.

Transformers (Vaswani et al., 2017), which have had massive success in NLP and since gained popularity in vision + language tasks and many other fields as well, are a network architecture built entirely on top of this notion of (self-)attention. For many sequential tasks the self-attention mechanism in Transformers have proven to be more successful than recurrent architectures since self-attention layers connect all positions with a constant number of sequentially executed operations rather than the $\mathcal{O}(n)$ steps required in recurrent neural networks thus making it much easier to learn long-term dependencies in data.

Transformers use an extension of the attention mechanism from Equation 2.3 called *multi-head attention* in which queries, keys and values are first projected to h different dimensions:

$$\begin{aligned} \text{MultiHeadAttention}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{where head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.4}$$

with parameter matrices W_i^Q, W_i^K, W_i^V. Our work applies the same kind of multi-head attention to the multimodal setting.

The effectiveness of these Transformer architectures and self-supervised pre-training methods on large corpora has been measured and demonstrated on a variety of downstream tasks. One of these tasks is question answering on the

SQuAD dataset (Rajpurkar et al., 2016) where models are given a question and have to retrieve the correct answer to the question from a given text passage. We believe a similar kind of reasoning and understanding is necessary to extract the right answers to a question from video stories.

Devlin et al. (2019) introduce several self-supervised pre-training tasks for language modeling using Transformers and demonstrate the effectiveness thereof during training on large-scale text corpora. Going beyond simple left-to-right or right-to-left modeling they introduce (i) *Masked Language Modeling* (MLM) in which input tokens are randomly masked and replaced by a `[MASK]` token or another random token as well as (ii) *Next Sentence Prediction* (NSP) in which the model predicts whether the two input sentences A, B to the model match. In 50% of cases B has been replaced by another non-matching sampled sentence B'. This task is supposed to help the model learn relationships between sentences which is especially for relevant for downstream tasks such as *question answering* (QA).

However, further research by Liu et al. (2019) disputes this claim and shows that better performance in downstream tasks can be obtained without an additional NSP loss. Furthermore, they demonstrate that BERT is still underfit and a larger amount of training data as well as additional hyperparameter tuning yields significantly better results using the same architecture. In our work natural language understanding is key to understand both question, answers to extract semantics of the story from subtitles. For these reasons we opt to use the more robust and optimized RoBERTa (Liu et al., 2019).

2.2 Visual Question Answering

Rather than answering questions about language alone, we are interested in answering questions about visual input. Visual Question Answering (VQA) is a task in which, given an image and a natural language question about the image, the agent should provide the correct answer (Antol et al., 2015). This kind of multimodal question answering setting requires the agent to resolve cross-modal references and selectively extract information from relevant areas of the image. An example question might for instance be "What color shirt is the woman holding the baby wearing?" requiring the agent to find the correct person (i.e. the woman holding the baby) among possibly multiple persons in the image and extract the color of a specifically the shirt. The most popular such dataset is the VQA dataset (Antol et al., 2015) consisting of around 250k images, 760k questions and 10M different answers. The dataset provides a multiple-choice and an open-ended answer setting.

A simple baseline for the multiple choice setting introduced in the original VQA paper consists of an image feature extractor like VGGNet as well as an LSTM to extract question features. The extracted feature vectors are fused by element-wise multiplication and the correct answer is inferred using a multi-layer perceptron and softmax on top of the fused features. More sophisticated approaches extract object features using a feature extractor like Faster-RCNN and selectively attend to them conditioned on the question (Anderson et al., 2018). A major problem with the original VQA dataset was answer bias resulting in models which performed well based on the language modality alone and without truly understanding the visual content. For instance, 41% of questions starting with "What sport is...?" can correctly be answered with "tennis", for 39% of questions asking "How many...?" the correct answer is "2" and blindly

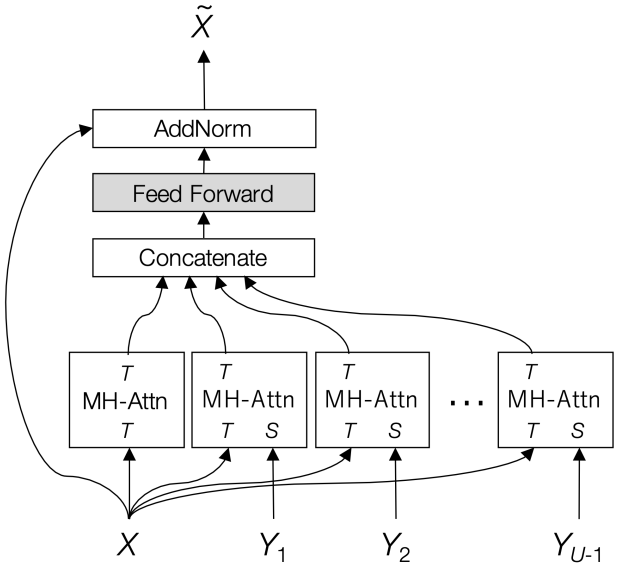

Figure 2.1 An attention module for many utilities introduced by Nguyen et al. (2020). A target utility X is attended to source utilities $Y_1, \ldots, Y_{U-1}$ to obtain an updated representation $\tilde{X}$. Figure from Nguyen et al. (2020).

answering "yes" to questions starting with "Do you see a...?" yields 87% accuracy. To this end Goyal et al. (2017) introduce VQA 2.0 which balances the original VQA dataset and provides two images which require different answers for every question thus requiring the model to actually pay attention to the visual modality. All subsequent mentions of the VQA task will refer specifically to VQA on the VQA 2.0 dataset.

VQA and the video question answering and video story understanding tasks are naturally related in that they both deal with cross-modal vision and language data and require the modal to semantically understand the contents of the visual modality to answer the natural language questions. Therefore, it is naturally important to look at existing methods in VQA to make progress in

video question answering and video story understanding. We will look at more methods (specifically cross-modal Transformers and self-supervised cross-modal pre-training schemes) for VQA in Section 2.3.

The Visual Dialog task is an extension of VQA that requires the agent to hold a dialogue about an image, meaning that in addition to the current question and the image the agent also has to extract relevant image from the dialogue history to infer the correct answer (Das et al., 2017). Nguyen et al. (2020) propose a modified multi-modal Transformer model which incorporates information from all three different modalities (image, question and history) to infer the correct answer (see Figure 2.1). The modified two-stream co-attention block that we propose for video story understanding in this work is similar in nature to the attention mechanism introduced by Nguyen et al. for many utilities.

2.3 Vision and Language Representation Learning

Similar to how Transformers can be pre-trained on language in a self-supervised manner as demonstrated by BERT to achieve better performance on downstream tasks such as question answering, it is possible to pre-train cross-modal Transformer architectures in a BERT-like manner to learn better visiolingual representations for downstream tasks such as VQA or image captioning. Although we do not perform any pre-training in this work, we will look at some of these models and their architectures which have proven to be highly effective for VQA.

We adopt the categories introduced by Luo et al. (2020) to classify the different vision and language representation learning approaches. As seen in Figure 2.2 there are three main categories (a) the single-stream share type in

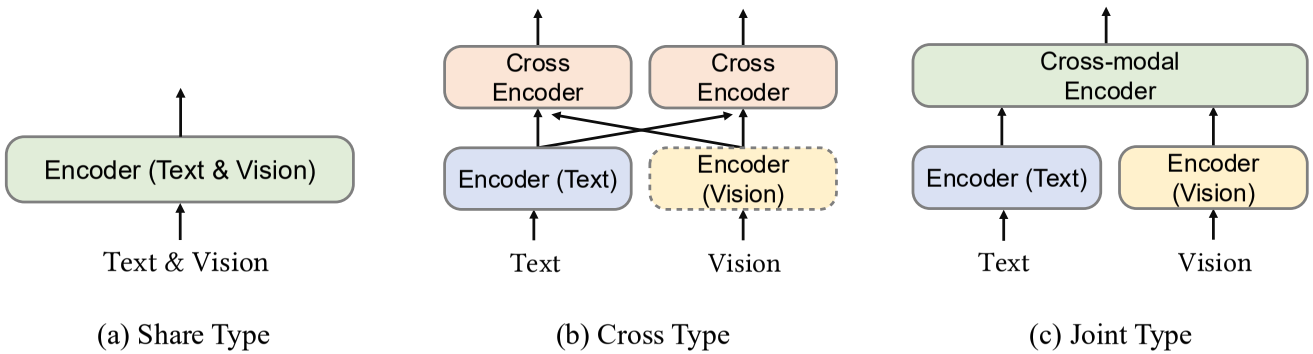

Figure 2.2 Vision and language representation learning approaches can be classified into three main types: (a) the single-stream share type, (b) the two-stream cross type and (c) the single-stream joint type. Figure from Luo et al. (2020).

which a single encoder encodes a sequence of raw text and vision signals and outputs a single sequence, (b) the two-stream cross type in which each modality is first encoded by its own specific encoder, then cross-encoded by one cross-modal encoder for each modality and finally outputting two cross-attended sequences, and (c) the single-stream joint type which uses modality-specific encoders but a single cross-modal encoder and which outputs a single cross-encoded sequence.

Single-stream share type. The main advantage of single-stream share type models is that they are relatively simple but still manage to learn useful representations through pre-training tasks. VL-BERT (Su et al., 2020) is a modification of the original BERT architecture to accommodate visual embeddings in the same Transformer encoder. Specifically, they add a form of visual feature embedding to the input features. To do this, they first extract region features from objects in the image obtained from Faster R-CNN and add `[IMG]` tokens to the token sequence to obtain a sequence like "`[CLS]` the man is standing in front of the parked car `[SEP]` `[IMG]` `[IMG]` `[IMG]` `[IMG]` `[IMG]` `[END]`". A visual feature embedding is then added to all token inputs: full image features for all language tokens and region specific features for the regions the `[IMG]` tokens represent. The model is pre-trained in a self-supervised manner on image-caption pairs using a BERT-like MLM task which is conditioned on the visual modality however. This means the model can learn visual coreferences by predicting masked words from visual context. Additionally a *Masked RoI Classification* task is introduced in which the model has to predict the object classes of masked `[IMG]` tokens using visual context. After fine-tuning the model achieves good performance on down-stream vision and language tasks such as VQA (Goyal et al., 2017), Visual Commonsense Reasoning (VCR) (Zellers et al.,

2019) and Grounding Referring Expressions (Yu et al., 2016).

Unicoder-VL (Li et al., 2020) works in essentially the same way but it adds an additional *Image-Text Matching* (ITM) pre-training task. The authors sample both positive and negative image-caption pairs and the model has to predict whether the caption describes the image. This additional task is meant to help learn a better instance-level alignment.

Two-stream cross type. Although an approach like the single-stream share type approach is conceptually appealing, it treats both modalities in the same way without sufficiently addressing their different pre-processing needs and can weaken the pre-trained BERT model used. Hence, a two-stream approach as introduced in ViLBERT (Lu et al., 2019) might be better suited to fuse information from both modalities. In ViLBERT an image represented by extracted object region features $v_0, \ldots, v_M$ is merely projected to a lower dimension whereas the sequence of word tokens is input into a Transformer encoder first (in this case a pre-trained BERT model) to obtain text embeddings $w_0, \ldots, w_N$. To attend the visual features to the language features and vice versa the authors use two co-attention Transformer blocks – one for each modality. As a result updated hidden representations $h_{v_0}, \ldots, h_{v_M}$ and $h_{w_0}, \ldots, h_{w_N}$ that are attended to the other modality respectively are obtained. To learn better representations they apply pre-training which again uses multi-modal masked modeling as well as image-caption alignment prediction. One key difference is that instead of learning to predict object classes directly, the model minimizes the KL divergence between the distribution of classes predicted for the masked object tokens and the classes predicted by the pre-trained object feature extractor. The learned hidden representations can be fine-tuned and applied to down-stream vision and language tasks like VQA, VCR, Grounding Referring Expressions,

image retrieval (given a caption) as well as caption retrieval (given an iamge).

ERNIE-ViL (Yu et al., 2020) uses the same architecture but they use a scene graph parser to transform the given image caption into a scene graph first. By masking and predicting objects (i.e. "girl", "ball"), attributes (i.e. "blue", "round", "small") and relations (i.e. "on top of", "in", "holding") selected from the scene graph in the caption (rather than selecting words to mask from the caption directly) the model can learn better semantic connections. Instead of random words entire entities are masked (e.g. all of "on top of" is masked instead of a single word) leading to better results. LXMERT (Tan and Bansal, 2019) uses an architecture similar to ViLBERT but pre-trains on in-domain (VQA) data with an additional task where the model predicts whether question and image are matched.

Single-stream joint type. UNITER (Chen et al., 2020b) uses vision-specific and language-specific encoders along with a single-stream cross-modal encoder. Both the vision and the text input are first encoded using the respective encoders and then, much like in the single-stream share-type, concatenated and input to the cross-modal Transformer. While UNITER is pre-trained much like all of the above models and although the underlying ideas are similar, one notable difference is that they carefully mask words or objects so that the instance in the other modality remains intact. This is important as it helps avoid misalignment that can result from both instances being masked at the same time and the model being forced to attend to a different instance instead. UNITER is pre-trained using MLM, Image-Text Matching, a novel *Word Region Alignment* (WRA) task in which they use Optimal Transport to minimize the cost of "transporting" the contextualized image embeddings to word embeddings and vice versa, and three Masked Region Modeling (MRM) tasks: classification

of masked regions (similar to other methods above), classification with KL divergence (similar to ViLBERT), and lastly Masked Region Feature Regression in which the Transformer output of masked regions is regressed to the visual features extracted from Faster R-CNN.

Although our approach does not employ any pre-training schemes, it is most similar in its architecture to the two-stream cross type approaches like ViLBERT which encode each modality individually and then co-attend with separate cross-modal modules outputting representations for the vision modality which are attended to the language modality and vice versa. We believe it is helpful in video question answering to have features which we can score for each modality thus encouraging the model to extract useful information from both modalities.

2.4 Video Representation Learning

Video representation learning is a new but active area of research which pursues much of the same ideas as vision and language representation learning. In many cases video subtitles or transcribed audio from automatic speech recognition (ASR) are available and learned jointly with the sequence of image frames, thus providing a setting which is very similar to what we have seen thus far in Section 2.3.

VideoBERT (Sun et al., 2019) is a single-stream (share type) model using a single Transformer encoder to jointly encode ASR text from video segments and *visual words* extracted from video. While the model architecture is identical to BERT, the authors introduce a novel way of creating visual words. S3D visual features (Xie et al., 2018) are first extracted from video using a specific convolutional network for video that adds temporal convolutions and then

clustered using hierarchical k-means. Each video segment is represented by its k-means cluster centroid, meaning similar video segments will be represented by the same vector token. These centroid tokens are then added to BERT as "visual words". This allows the model to learn high-level ideas without getting distracted by e.g. textures in the video. The model is pre-trained on the same MLM task as BERT and an additional video-transcript alignment task (i.e. predict whether they match or not) on a novel dataset consisting of 966 days worth of instructional YouTube videos and their transcripts obtained via ASR. VideoBERT achieves good performance in downstream tasks like action classification and video captioning.

Uni-VL (Luo et al., 2020) takes a different single-stream joint type approach, first encoding each modality and then cross-encoding via a cross-modal Transformer. For language features pre-trained BERT is used whereas video features are extracted using S3D. Uni-VL is pre-trained on five pre-training tasks. They employ MLM and *Masked Frame Modeling* (MFM) where instead of feature reconstruction they aim to maximize mutual information between the output and the masked target using a noise contrastive estimation (NCE) loss (Gutmann and Hyvärinen, 2010, Sun et al., 2020). They adopt an MIL-NCE loss (Miech et al., 2020) to align modality-specific text encoder and video encoder outputs using negative distractor samples (i.e. negative transcripts). For video text alignment they use an additional NCE loss on the cross-modal Transformer output to learn to discriminate positive and negative video-text pairs. Lastly, they jointly train an auto-regressive Transformer decoder as well which learns to reconstruct the original transcript, thus enabling use on downstream generation tasks as well. The model is pre-trained on the HowTo100M dataset, a large-scale dataset of narrated (and transcribed) instructional videos collected from YouTube. They fine-tune and evaluate their model on text-based

video retrieval, video captioning, action segmentation and step localization as well as multimodal sentiment classification with good results, e.g. significantly outperforming VideoBERT on the video captioning task.

Notably none of the above approaches employs or experiments with a two-stream Transformer architecture for video understanding as they are largely focused on what is happening in the video modality alone for tasks like action classification, video captioning or video retrieval. Since we want to model both video and language jointly and since both modalities may be equally relevant to understand an underlying video story, we believe a two-stream approach may be better suited for video story understanding.

2.5 Video Question Answering

As an extension of VQA in the temporal domain, video question answering (video QA) adds an additional dimension to an already challenging problem and has thus far received considerably less attention than VQA. As models should both learn to resolve and utilize contextual information and references between the vision and language but also perform multi-step and long-term reasoning in the temporal axis research on video QA has been sparse.

In video question answering an agent is presented with a video clip of a scene and has to infer the correct answer to a given question in natural language. Such a scene can consist either of a single shot or multiple shots from different angles or in different locations. While questions can be relatively simple, e.g. “What is the woman holding?”, or more complex and require deeper understanding and multiple steps of reasoning, e.g. “Why is the man in the overalls angry at the cyclist?”. Due to these temporal dependencies that need to be resolved and understood in order to answer more complex questions about the scene

correctly, video question answering has remained a very challenging problem.

Pre-neural. Yang et al. (2003) introduce an early pre-neural approach and one of the very first approaches to video QA which like most machine learning research of the time relies heavily on hand-crafted video features as well as transcriptions of spoken word. In their work Yang et al. aim to retrieve news videos the user is looking for by their contents as well as to retrieve answers to question from the video transcripts. To accomplish this they use hand-crafted feature such as face features, speaker change features, and color histograms and classify videos at the shot-level into categories like interview, finance, weather and sports using HMM analysis. To answer questions they first select the video segment and then predict the correct answers from the associated transcripts obtained via speech recognition.

Encoder-decoder and RNN-based. Going beyond hand-crafted features, Zhu et al. (2017) introduce an approach based on recurrent neural networks (RNNs) with an encoder-decoder architecture. Zhu et al. extract visual features from each frame using a pre-trained convolutional neural network which are then input into the encoder RNN. Based on the output of the encoder, three decoder RNNs are pre-trained in an unsupervised way to reconstruct the present (i.e. reconstruct the current frame's features), the past (i.e. previous frames) and predict the future (i.e. future frames). Finally, the encoder is fine-tuned to answer "fill-in-the-blanks" multiple-choice questions on a dataset collected by the authors by ranking answer options constrastively with a dual-channel ranking loss. It is worth nothing however that the questions are very simple and that this type of "fill-in-the-blanks" approach to QA is prone to language biases.

Zeng et al. (2017) collect a novel dataset of ∼18k open domain videos with three to five description sentences per video. Additionally they extend multiple LSTM-based methods for VQA to the video domain. To do this they first extract spatiotemporal C3D frame features (Tran et al., 2015) which they then encode with another LSTM. As video and descriptions may not be perfectly aligned they also introduce a learning procedure which tries to mitigates the effects of misaligned clips by identifying them at training time using a ratio test.

Zhao et al. (2017) address video question answering in the open-ended setting meaning that there are no answer options and the agent has to generate the answer sequence from scratch. To this end, they propose an encoder-decoder framework which first encodes the video hierarchically using GRU and temporal and spatial attention modules. Unlike Zeng et al. (2017), Zhu et al. (2017) they first extract object region features. The encoder first computes frame-level representations by attending objects to the question via spatial attention across ROIs in a frame using the spatial attention module. Next a video-level representation is computed inferring which frames to pay attention to using a temporal attention module. The decoder network generates the answer sequence conditioned on the video context vector obtained from the encoder. To evaluate the open-ended questions they compute accuracy (i.e. does the answer match the ground-truth exactly) and WUPS which scores answers based on WordNet similarity to the ground-truth answer.

Memory-networks. Gao et al. (2018) propose a memory-based approach to video QA which utilizes cues from both appearance and vision features. Videos are segmented and for each segment visual features are extracted from the middle frame using a pre-trained ResNet-152 architecture (He et al., 2016) as appearance features and optical flow features are extracted using a pre-trained

two-stream convolutional network (Simonyan and Zisserman, 2014) as motion features. Using a "conv-deconv" network which subsequently convolutes and deconvolutes and integrates temporal context they build "facts" which are then used in their co-attention memory network. The co-attention memory network updates its appearance and motion memory in multiple cycles for multiple steps of reasoning (for each cycle attending facts to the existing memory). To update the memory motion facts are attended both to the motion memory and to the appearance memory and vice versa. Answers are computed using a linear layer on top of the concatenated appearance and motion memory vectors. Gao et al. train and evaluate their approach on the TGIF-QA dataset (Jang et al., 2017), a large-scale dataset focusing on visual reasoning on short video clips with both open-ended and multiple choice questions, outperforming previous methods by a significant margin.

Graph-based. A slightly different approach is taken by Huang et al. (2020a) who introduce a two-stream graph-based model for video QA. The language modality, that is the question, is encoded with pre-trained GloVe embeddings (Pennington et al., 2014) and a bidirectional LSTM while the video is encoded using a location-aware graph convolutional network. More specifically, they extract object features from video frames and construct a fully connected graph on the detected objects. Features in the graph encode both the bounding box location information within the frame as well as the temporal location. The video representation is computed using graph convolution. To compute interactions between the visual and the question features they introduce an interaction module which first computes the most relevant question words for visual representations using an attention mechanism and then a cross-modal representation. Finally, the answer to the question is predicted using a fully connected layer on

top of the cross-modal representation output.

TVQA dataset. Lei et al. (2018) introduce the TVQA dataset consisting of ∼150k QA pairs on ∼22k video clips from six popular TV shows. The clips are relatively short at 60 to 90 seconds and the questions bridge vision and language clues. To answer correctly agents have to infer the answers by using information from multiple modalities, namely video frames and dialogue subtitles as well as temporally localize the relevant moment within the video. To this end Lei et al. also introduce a baseline "Multistream" model in the original TVQA paper. For visual features they extract objects and attributes in the image using Faster R-CNN pre-trained on Visual Genome (Krishna et al., 2017) and full frame features using a pre-trained ResNet model. To encode the resulting visual sequences they employ a bidirectional LSTM. The question, answers and subtitles are first encoded using a pre-trained GloVe embedding and then similarly input to a bidirectional LSTM each. In order to model the different modalities jointly they employ a context matching module which, given a query vector, produce a context-aware query output (i.e. video-aware-question and video-aware answer representations). The obtained context-matched features are input into a linear layer with softmax to obtain answer scores. The final scores are are obtained by summing the scores of each context-matching stream. Specifically their model uses a question-video-answer stream which matches both question and answers to the video as well as a question-subtitle-answer stream which matches them to the subtitles. In experiments they find that their baseline model reasons over the video clip reasonably well.

However, it is worth nothing that this model does not compute any interactions between video and subtitles although they may be necessary to answer some of the questions. For instance some things may be set in jest or angrily

which might not be clear from the transcribed dialogue alone but might require inference over the characters' facial expressions as well.

(Yang et al., 2020) take a different approach to video question answering by leveraging a pre-trained BERT model. Instead of using video input directly, Yang et al. use a two-stream BERT model for video QA which encodes the semantic content of a video scene as the visual concept labels of the detected objects given by Faster R-CNN. Using these language labels of objects (i.e. "blonde hair") instead of the visual features allows them to use the embeddings produced by BERT, which have proven impressively effective in the language domain, for visuals as well. In their experiments their model performs significantly better than the TVQA baseline but it's important to keep in mind that the model is not truly reasoning over video frames but merely over the object labels produced by Faster R-CNN and thus lacks actual understanding of what is happening visually.

(Geng et al., 2020) use a similar approach but instead first infer character names for face bounding boxes through multi-instance co-occurrence matching. Next, they infer relations between these characters and objects detected in the image (e.g. "¡Lily, hold, flower¿") and embed them using a word embedding layer. Finally, they take a Transformer-based encoder-decoder approach to encode question, answer, subtitles and visual relation features and compute scores with a linear layer on the output of the Transformer decoder. While this approach manages to take into account which characters are present in the video and how they relate to one another, it also does not use visual information beyond the extracted object relations.

TVQA+ adds additional bounding boxes and objects annotations that link them directly to visual concepts mentioned in questions and answers (Lei et al., 2020). While the TVQA(+) dataset seems to be a popular choice for evaluation

of video question answering and video understanding models (Geng et al., 2020, Kim et al., 2019, Yang et al., 2020), it does not require story-level understanding. Most questions in the dataset only require the agent to attend to a short part of the video clip (15 seconds or less) due to its particular focus on temporal localization.

DramaQA dataset As the TVQA(+) dataset largely focuses on questions which require shorter video segments that need to be localized in the video clip (cf. temporal localization in TVQA+), Choi et al. (2021) introduce the DramaQA dataset for video story understanding. The DramaQA dataset aims to benchmark understanding of story-level questions rather than those shorter dependencies seen in the TVQA dataset. In particular, it is based on the Korean TV show "Another Miss Oh", consisting of 23 928 video clips, of which 803 are scene-level and 23 125 are shot-level clips, and spanning 18 episodes in total. The questions are multiple-choice with five possible answers to choose from and can be categorized into four difficulty levels. Depending on the level of difficulty higher-level understanding of the scene may be necessary to answer correctly. For a more detailed look at the DramaQA dataset see Chapter 3.

As we aim to primarily address video question answering in the context of video story understanding in this work we believe the DramaQA dataset is best suited and we conduct our experiments on the DramaQA dataset (see Chapter 5). Choi et al. also introduce a novel context matching model (Choi et al., 2021) which they evaluate on this dataset and which will serve as a baseline for our work. A more detailed explanation of this approach as well as further baselines provided in the DramaQA paper follow in Chapter 3.

Chapter 3

Dataset and Problem Formulation

In this chapter we give a detailed overview over the DramaQA dataset (Choi et al., 2021), why we have chosen it for our experiments of our video story understanding model and several baseline methods introduced in the DramaQA paper.

3.1 Dataset

In this work we focus on the more narrow problem of video question answering in the video story understanding setting. In this setting an agent is presented with a video clip that tells a story and has to answer questions about the story as a whole. Instead of only focusing on motion or appearance alone like it is the case in the short video clips of the TGIF-QA dataset (Jang et al., 2017), agents are expected to resolve references between all modalities and perform multi-step long-term reasoning to understand characters' actions and intentions.

While many works in video question answering focus on what is observed in the visual modality alone, the story understanding setting typically provides transcriptions or subtitles of spoken dialogues which is often crucial to understand interactions between characters. Moreover, it is naturally necessary to understand who the different characters in the story are.

A natural choice for video stories are TV shows and dramas as they closely

model real-world interactions and scenarios. As previously mentioned in Section 2.5, one such dataset for video question answering on TV show plots is the TVQA(+) dataset which is build around short 60 to 90 seconds long video clips and questions bridging vision and language clues (Lei et al., 2018, 2020). Agents have to infer the answers by using all available modalities as well as temporally localize the relevant part in the video clip. Additionally, in the extended TVQA+ version of the dataset, additional bounding boxes and objects annotations that link them directly to visual concepts mentioned in questions and answers are provided. While it is a popular dataset for video question answering, the TVQA(+) is largely focused on questions that revolve around particular actions. In particular, most questions in the dataset only require the agent to attend to a short part of the video clip of 15 seconds or less due to its specific focus on temporal localization. We argue story understanding focuses on more abstract ideas and while key parts or interactions in the story are relevant to the understanding of the story as a whole an understanding of shorter video sections does not require full story-level understanding.

To this end, Choi et al. (2021) introduced the DramaQA dataset (see also Section 2.5) which aims to benchmark exactly this type of video story understanding and focuses in particular on story-level questions rather than the shorter dependencies seen in the TVQA dataset. The DramaQA dataset is based on the Korean TV show "Another Miss Oh", consisting of 23 928 video clips, of which 803 are scene-level and 23 125 are shot-level clips, and spanning 18 episodes and 20.5 hours of video in total. The dataset additionally provides character-level annotations for each of the 217,308 frames.

The total number of questions contained is 17 983. All questions are multiple-choice with five possible answers to choose from and can be categorized into four logical complexity levels. Depending on the level of logical complexity of

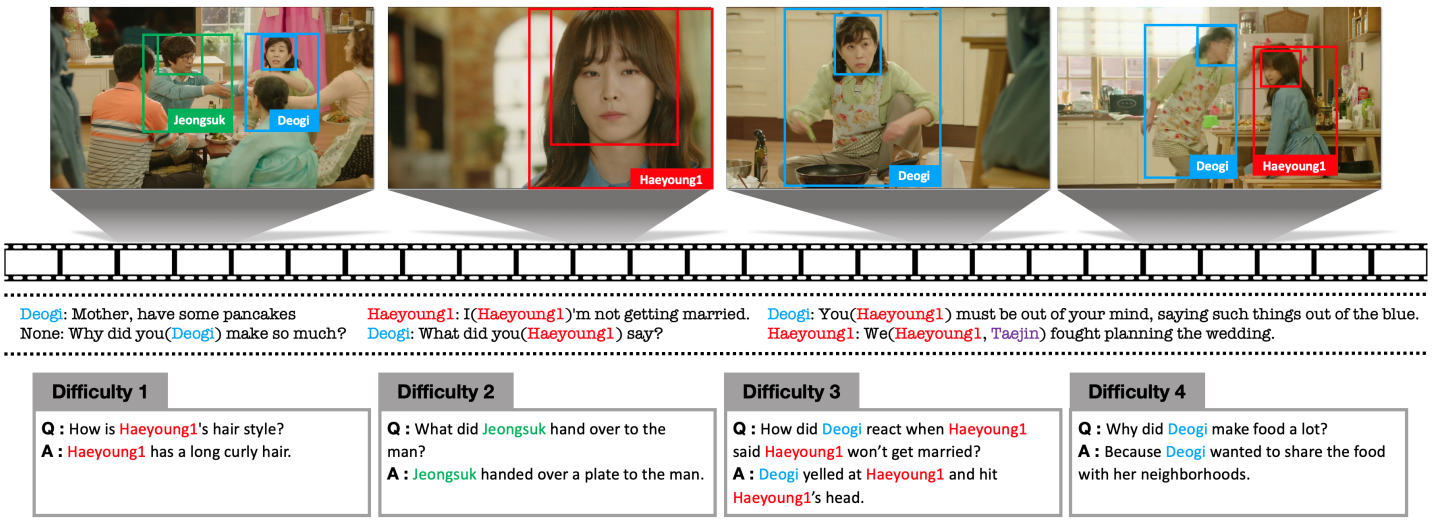

Figure 3.1 Example of the video question answering task for video story understanding on the DramaQA dataset for all difficulty levels. Difficulties range from 1 (easiest) to 4 (hardest) with easier ones being focused single observations and harder questions requiring more long-term reasoning and multimodal knowledge. Character names (highlighted) are co-referenced between script, question, answers and bounding boxes. Figure from the original DramaQA paper (Choi et al., 2021).

the question higher-level understanding of the scene may be necessary to answer correctly. The logical complexity ranges from level 1 which only requires a single supporting fact to answer the question (i.e. "Who is holding the phone?"), level 2 which requires multiple such supporting facts to level 3 and 4 which require to reason across the temporal axis and to understand causal relationships between multiple supporting facts. Hereinafter, we will refer to logical complexity levels simply as *difficulty levels*. An example of questions of each difficulty level can be seen in Figure 3.1. Additionally, questions can be categorized by the *memory capacity* needed in order to answer the question and the dataset provides both shot-level and scene-level questions. Along with the image frames the dataset provides the coreference-resolved dialogue scripts for each video clip. Additionally, it provides bounding boxes for characters appearing in each frames with visual metadata annotations containing name, behavior, emotion i.e. "$\{\text{Doegi}, \text{standing up}, \text{surprise}\}$".

As we focus primarily on video story understanding with an emphasis on deeper understanding of long-term dependencies in video and characters' actions and intentions in this work, we choose to evaluate our approach on the DramaQA dataset which was specifically introduced for this type of video question answering. Furthermore, DramaQA allows for evaluation by difficulty level and can therefore give us a better understanding of where the strengths and weaknesses of our method lie.

3.2 Problem formulation

Next, we will give a formal definition of the video question answering setting. We formulate the problem of video QA for understanding of visual stories in manner similar to Choi et al. (2021). That is, for a given question sequence Q and given

the reference video clip's sequence of transcribed subtitles S along with a visual feature stream V and visual metadata M, we want to infer the correct answer sequence $A_i \in \{A_1, \ldots, A_5\}$. Question, answer options, and subtitles are given as a raw text sequence. The visual feature stream consists of a series of region features extracted from the characters' full body bounding box for each frame in the reference video.

3.3 Baseline models

Choi et al. also introduced several baseline methods in the original DramaQA paper. A first very simple baseline is given by encoding both the question and answer using pre-trained GloVe embeddings, then taking the average across the sequence and simply computing the dot product similarity between them as an answer score. However, as this method ignores the subtitles and video completely, answer accuracy is barely better than random.

A slightly better baseline is a simple multilayer perceptron model. Question-answer pairs and subtitles are embedded using pre-trained GloVe embeddings and then encoded individually using bidirectional LSTMs. Visual features are encoded similarly. Finally, all of the context streams are mean-pooled across time and the resulting vectors are concatenated and scores using a multilayer perceptron. This model performs reasonably for lower difficulty questions but poorly for higher difficulty questions.

Finally, Choi et al. introduce a more complex, multi-stream context matching model which is based on the same context matching module that is also used by Lei et al. (2018) and focuses on character-guided representations. In the visual modality, character bounding box features are extracted from video frames using ResNet-18 and bounding box annotations consisting of behavior

and emotion are converted to word embeddings and along with a one-hot vector denoting the character concatenated to the visual features. The resulting visual stream is encoded via bidirectional LSTM to obtain a "low-level story representation". Similarly, question-answer pairs and subtitles are embedded via pre-trained GloVe embeddings and then encoded via bidirectional LSTM. In the case of subtitles an additional one-hot encoded vector annotating the speaker of the script for each word in the sequence is concatenated to the embeddings before the sequence is input into the LSTM.

In contrast to the model introduced by Lei et al., they use an additional context matching module with character queries in order to obtain character-guided higher-level representations of the story. Specifically, they learn an embedding of main characters and construct queries for the context matching as the sum of character representations occurring in the question-answer pair or in the video depending on the stream. By context-matching the low-level representations to the queries they obtain character-guided representations E_S and E_V for the subtitle and the visual feature stream. Finally, each of the low-level and high-level representations for subtitle and visual feature streams is context-matched to the QA pair and a score is computed for each of the streams using a linear layer and softmax. The total score is the sum of individual stream-scores.

In evaluations the final multi-level multi-stream model performs better than the adopted multi-stream TVQA model which also uses the additional character and visual metadata annotations.

We will introduce our model for video story understanding on the DramaQA dataset in Chapter 4 and evaluate it and compare the results to the above described baselines in Chapter 5.

Chapter 4

Proposed Method

Our method takes inspiration from the aforementioned recent advances in vision and language fusion (see Section 2.3) especially with applications to visual question answering and visual dialog. Recent work by Nguyen et al. (Nguyen et al., 2020) on the visual dialog task introduced a new type of co-attention layer for three or more input modalities (that is, the image, question and dialog history) with fewer trainable weights, that can be stacked in order to better integrate dependencies between many different utilities. While ViLBERT's co-attention layer (Lu et al., 2019) only takes two modalities as input, it co-attends them in a similar fashion but uses more powerful language representations from BERT as well as a multimodal pre-training scheme that enables the model to learn better visiolingual representations.

We argue that video question answering and video story understanding require agents to fuse vision and language features to infer the correct answer to a question in the same fashion and can benefit from a similar co-attention layer as well. We will adopt such co-attention layer for the video story understanding setting. Naturally, better comprehension of questions and what is being asked for and answer options will lead to better results as well. Beyond that, specifically video story understanding requires a nuanced understanding of language in order to fully incorporate information from dialogues between characters. To

this end, we base our model on a more powerful pre-trained language model which can grasp said nuances on a much finer level.

For an overview over our proposed architecture see Figure 4.1. We introduce a two-stream co-attentional Transformer model which first separately encodes each modality using a modality specific encoder, thus addressing the specific pre-processing needs of each modality, and then co-attends the two modalities using L co-attentional Transformer layers each consisting of two co-attentional Transformer modules (see Figure 4.2). In these modules the visual feature stream is attended to language features and vice versa the language feature stream is attended to the visual features. The output of this two-stream co-attentional Transformer are accordingly updated representations that incorporate information from the other modality. To obtain a single context vector per stream we perform max pooling along the sequence dimension (rather than the feature dimension). Finally, we obtain one score per stream using a single linear layer which takes the modality's context vector as input. To compute the total answer score the vision and language scores are summed up.

4.1 Model architecture

Following Choi et al. (2021) we use question features, subtitle features, answer features, meta features (behavior and emotion) as well as visual features. Visual features are extracted from the annotations of character bounding boxes given in the dataset using a pre-trained image feature extractor such as ResNet (He et al., 2016) yielding D_V-dimensional image representations. Questions, answers and subtitles are raw text input as provided by the DramaQA dataset.

Meta features are annotations provided for each character bounding box describing an action (i.e. "drink", "eat", "dance") as well as an emotion (i.e.

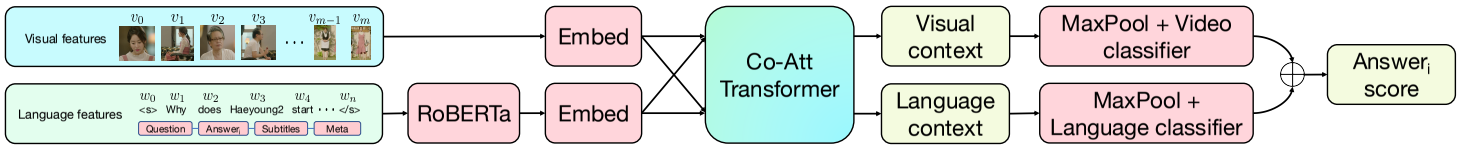

Figure 4.1 Our two-stream transformer model manages to learn cross-modal interactions through co-attention. The final answer score takes into account both the visual context attended to the language representations as well as the language representations attended to the visual bounding boxes.

"fear", "happiness", "neutral"). In order for the language model to infer "who is doing what and feeling how" from the meta features more easily we experiment with transforming them to sentences by including the character name with both the action and the emotion. We construct these sentences either as simple word sequences, i.e. "Doegi standing up sadness.", or full sentences, i.e. "Doegi is standing up and feeling sadness". We will compare both approaches in Chapter 5.

As language plays a key role in understanding dialogue in video stories as well as question and answering, we use a strong pre-trained Transformer model to encode the language modality. In particular, to encode the given language token sequences and to obtain language representations we use a pre-trained RoBERTa model (Liu et al., 2019), a variant of the widely successful BERT model that achieves significantly better performance with the same architecture but tuned hyperparameters and more training data. We leverage the powerful language representations from the RoBERTa model to learn better visio-lingual representations for video story understanding and to better reason across e.g. longer subtitle sequences.

In order to predict the correct answer we will compute scores for each answer individually and then select the highest scoring answer. Therefore, we will have a distinct language feature stream $l_{i_1}, \ldots, l_{i_N}$ for each of the five provided answer options along with the visual feature stream $v_1, \ldots, v_M$.

Language encoder. For each language feature stream we first fuse all textual inputs as follows; for the i-th answer option we obtain the concatenation of question Q, subtitle sequence S, sentence-encoded metadata features M and

the answer A_i resulting in

$$l_i = [Q; S; M; A_i]. \tag{4.1}$$

During concatenation, we separate the tokenized sequences using RoBERTa's "start-of-sentence and "end-of-sentence" tokens as follows:

$$\langle s \rangle\ Q\ \langle /s \rangle \quad \langle s \rangle\ S\ \langle /s \rangle \quad \langle s \rangle\ M\ \langle /s \rangle \quad \langle s \rangle\ A_i\ \langle /s \rangle \tag{4.2}$$

Although RoBERTa has not been pre-trained for more than two sentence types, this separation will aid the model in differentiating between the different input types. Moreover, we add a stream of segment IDs $s_n \in \{0, \dots 4\}$ for $n \in N$, where N is the language sequence length, signaling what type of input each of the parts of the language stream belongs to (i.e. 0 for question tokens, 1 for subtitle tokens, and so on). Fusing all textual inputs gives us a total of 5 language token sequences $l = \{l_1, \dots, l_5\}$ containing the respective question-answer pair along with the subtitles and meta features. Given the token sequence l_i of length N, we can now use RoBERTa to obtain language representations $\ell_i \in \mathbb{R}^{N \times d_L}$ for all i where d_L is the hidden size of the text representation.

$$\ell_i = \text{RoBERTa}(l_i) \tag{4.3}$$

Next, we use a linear layer projection to obtain representations in a joint visio-lingual embedding space of dimension d

$$L_i = \text{Linear}(\ell_i) \tag{4.4}$$

with $L_i \in \mathbb{R}^{N \times d}$ and "Linear" denoting a fully connected linear layer.

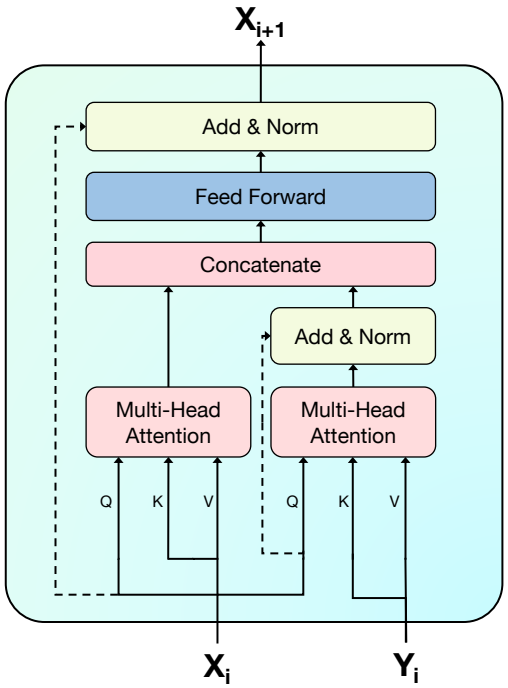

Figure 4.2 A single co-attention block attends a target utility X to a source utilitiy Y. We use two of these blocks per layer, attending both the visual representations to the language representations and vice versa. Given input representations $L_i^{(k-1)}, V_i^{(k-1)}$ for answer option i, we obtains new utilities $L_i^{(k)}, V_i^{(k)}$ that are attended to one another.

Vision encoder. To obtain visual representations, we extract image region features of the main characters for each frame from the annotated bounding boxes using a pre-trained image feature extractor yielding feature representations of d_V for each bounding box. We concatenate all of these bounding boxes in the order of appearance, that is in the frame order, giving us a sequence of visual features $v \in \mathbb{R}^{M \times d_V}$. For these visual features we also use a linear layer to project these dimensions from the image feature dimension d_V to the joint visiolingual embedding space of dimension d yielding representations $V \in \mathbb{R}^{M \times d}$,

$$V = \text{Linear}(v). \tag{4.5}$$

Co-attentional Transformer. In order to capture both dependencies between the two modalities and within the temporal axis, we use a co-attentional Transformer (see Figure 4.2) to obtain vision and language context representations that have been fused with the other modality. Given a tuple of vision and language representations in the joint embedding space $(L_i^{(k)}, V_i^{(k)}) \in \mathbb{R}^{N \times d} \times \mathbb{R}^{M \times d}$, for the k-th layer with $k \leq K$, we obtain the attended representations using a single co-attention layer consisting of two co-attention blocks as follows:

$$L_i^{(k+1)} = \text{CoAtt}(L_i^{(k)}, V_i^{(k)}) \tag{4.6}$$

$$V_i^{(k+1)} = \text{CoAtt}(V^{(k)_i}, L_i^{(k)}) \tag{4.7}$$

Naturally, it is possible to stack K of these co-attention layers to obtain representations that are more deeply fused. However, we find that $K = 1$, i.e. a single co-attention layer, works best in our experiments (see Chapter 5). We repeat this co-attention procedure for each language stream (i.e. for each answer

option).

Finally, we take the maximum along the sequence for both the language-attended visual and the vision-attended language stream and use a linear classifier to obtain a vision and language answer score for each answer.

$$S_V(i) = \text{Linear}(\max_{i \in M} V_i^{(K)}) \tag{4.8}$$

$$S_L(i) = \text{Linear}(\max_{i \in N} L_i^{(K)}) \tag{4.9}$$

The final scores are computed by simply taking the sum of the vision and language answer scores for each answer option:

$$\text{Score}(i) = S_V(i) + S_L(i) \tag{4.10}$$

We select the answer with the highest score as our prediction of the correct answer index j, i.e.

$$j = \arg\max_i \text{Score}(i). \tag{4.11}$$

We evaluate the performance of this approach in the next chapter.

Chapter 5

Experiments

In this chapter we perform various quantitative and qualitative analyses to evaluate the performance of our proposed method. We evaluate on the DramaQA dataset for video story understanding and compare our model to baselines from the original DramaQA paper as well as the winners of the DramaQA challenge held at ECCV 2020[1]. For qualitative analyses we provide example outputs for questions of varying difficulty. Furthermore, we perform several ablation studies evaluating performance with a subset of features, different ways of encoding the meta features, different number of co-attention layers as well as with a single-stream (joint type) cross-modal Transformer encoder similar to those seen in prior work discussed in Chapter 2 (Chen et al., 2020b, Luo et al., 2020, Sun et al., 2020).

5.1 Implementation details

For our experiments we extract 2048-dimensional region features from the bounding box annotations using a ResNet-152 model that has been pretrained on ImageNet, meaning we set $d_V = 2048$. For the linguistic stream we use RoBERTa$_{\texttt{BASE}}$ consisting of 12 Transformer encoder layers with a hidden size

[1]For more information about the DramaQA challenge at ECCV 2020 visit https://dramaqa.snu.ac.kr/Challenge/2020 (Archived: https://web.archive.org/web/20210529142058/https://dramaqa.snu.ac.kr/Challenge/2020)

of $d_L = 768$ and 12 attention heads. We believe that the RoBERTa$_{\texttt{LARGE}}$ model may lead to better results but have chosen the `BASE` model for our experiments due to resource constraints. We use the pre-trained version of the dataset provided by Huggingface[2] which has been pre-trained on the union of BookCorpus (Zhu et al., 2015), Wikipedia[3], CC-News[4], OpenWebText[5] and Stories (Trinh and Le, 2018) datasets, a total of 160GB of text data.

We use a joint embedding space of dimension $d = 300$. We use a single co-attentional transformer layer with 6 attention heads. For the maximum length of the video input sequence we choose $M = 300$, for the maximum number of tokens in the text input sequence $N = 300$. We train the model using two Titan Xp GPUs with a batch size of 6 for a total of 5 epochs. We use the Adam optimizer with a learning rate of 10^{-4} and a weight decay of 10^{-5}. To train the model we use a softmax and a cross-entropy loss on the predicted answer scores.

5.2 Quantitative results

We compare our co-attentional Transformer model against the baselines reported by Choi et al. (2021) on the test set. Namely we compare our model with the multi-stream model from the TVQA paper (Lei et al., 2018) which has been adopted to use the additional annotations in the DramaQA dataset by Choi et al., multi-stream and multi-level context matching model which

[2] `https://huggingface.co/roberta-base` (Archived: `https://web.archive.org/web/20210529144331/https://huggingface.co/roberta-base`)

[3] `https://en.wikipedia.org/wiki/English_Wikipedia` (Archived: `https://web.archive.org/web/20210529144853/https://en.wikipedia.org/wiki/English_Wikipedia`)

[4] `https://commoncrawl.org/2016/10/news-dataset-available/` (Archived: `https://web.archive.org/web/20210525052535/http://commoncrawl.org/2016/10/news-dataset-available/`)

[5] `https://github.com/jcpeterson/openwebtext` (Archived: `https://web.archive.org/web/20210529145507/https://github.com/jcpeterson/openwebtext`)

Model	Level 1	Level 2	Level 3	Level 4	**Overall**	**Difficulty Avg.**
Dot product (Q+A) (Choi et al., 2021)	0.3064	0.2720	0.2616	0.2225	0.2827	0.2656
MLP (Choi et al., 2021)	0.5724	0.4912	0.4132	0.3985	0.5129	0.4688
TVQA (Lei et al., 2018)	0.7480	0.7257	0.5330	0.5575	0.6945	0.6410
Multi-stream Context Matching (Choi et al., 2021)	0.7596	0.7465	0.5736	0.5663	0.7114	0.6615
Ours	**0.8064**	**0.7843**	**0.6846**	**0.6870**	**0.7724**	**0.7406**

Table 5.1 Evaluation results on the DramaQA test set by question logic level, overall and average across the difficulty levels. Higher levels require more complex reasoning.

	Overall	Difficulty Avg.
MCM (Choi et al., 2021)	0.7201	0.6421
Discriminative	0.7207	0.6654
Question + Answer	0.5630	0.5566
+ Subtitles + Video + Meta	**0.7743**	**0.7360**

Table 5.2 Evaluation results on the DramaQA validation set. We compare our model to an additional discriminative baseline which adopts the discriminative decoder from the visual dialog paper (Das et al., 2017) on top of the MCM baseline as well as a simple RoBERTa question + answer baseline.

uses character-guided representations (here "Multi-Stream Context Matching" or MCM for short) and two more simple baselines, i.e. a "Dot product" baseline which simply computes scores as the dot product similarity of mean-averaged question and answer word embeddings as well as a "MLP" baseline which encoded language embeddings and visual features using LSTM modules and computes scores using an MLP on top. For detailed explanations of the baseline models see Section 3.3. Results of our evaluation can be seen in Table 5.1. Our model outperforms all of the baselines. We improve upon the Multi-Stream Context Matching baseline on all difficulty levels; at least by about 3.8 percentage points and at most by 12.1 percentage points. Overall on the entire test set we can see a large improvement of ∼6 percentage points and an ∼8 percentage point improvement in difficulty average.

Across the difficulty levels we see the largest improvement over levels 3 and 4 where causal and long-term reasoning is necessary to infer the correct answer. We attribute this to the stronger fusion of vision and language resulting from the co-attentional Transformer model, thus allowing the classifier to take advantage of more complex cross-modal clues. Additional results obtained on the validation data set can be seen in Table 5.2.

Team name	Level 1	Level 2	Level 3	Level 4	**Overall**	**Difficulty Avg.**
GGANG	**0.81**	**0.79**	0.64	**0.70**	**0.77**	0.73
Sudoku	0.78	0.74	0.68	0.67	0.75	0.72
HARD KAERI	0.76	0.73	0.56	0.59	0.71	0.66
Ours	**0.8064**	0.7843	**0.6846**	0.6870	**0.7724**	**0.7406**

Table 5.3 Comparison of our model's results with the winners of the DramaQA challenge 2020 held at ECCV 2020. Results are evaluated on the DramaQA test set. All results of winners are rounded to two decimal places as they are reported on the scoreboard. The winning criteria was difficulty average.

DramaQA challenge. We also compare our model to the three winners of the DramaQA challenge at ECCV 2020. Evaluation results on the scoreboard of the DramaQA challenge are rounded to two decimal places so we report them similarly here. Although the overall evaluation results on the full dataset are very similar to the first place winner, our model outperforms the winning model on difficulty average, the winning criteria of the challenge.

5.3 Qualitative results

We perform an analysis of qualitative results do gain a deeper understanding of how our model operates. The first two qualitative examples for inference on the validation set can be seen in Figure 5.1. The video clip at hand is a relatively long, scene-level clip and questions have difficulty levels 3 and 4. We can see that for both of the questions long-term reasoning is necessary.

The first question requires close attention to the video clip with additional clues being given in the subtitles as Haeyoung1 does simply not respond to the angry comments made by Deogi. Our model correctly infers that Haeyoung1 simply leaves the room while ignoring Deogi's outburst. The second question focuses more strongly on the subtitles. Our model incorrectly predicts that Deogi is mad at Haeyoung1 for leaving the laundry. While it can be seen from the subtitles that she does indeed leave the laundry this is not the reason Deogi is mad but rather a consequence of it. As the predicted answer is related to something mentioned in the subtitles it seems our model has been confused and could not correctly infer the reason Deogi is angry.

Another set of inference examples is shown in Figure 5.2. The first question asks for the reason behind Deogi telling Haeyoung1 to wear a mask and requires the model to pay attention particularly to the subtitles. Our proposed model answers the question correctly and does not get distracted by Haeyoung1's

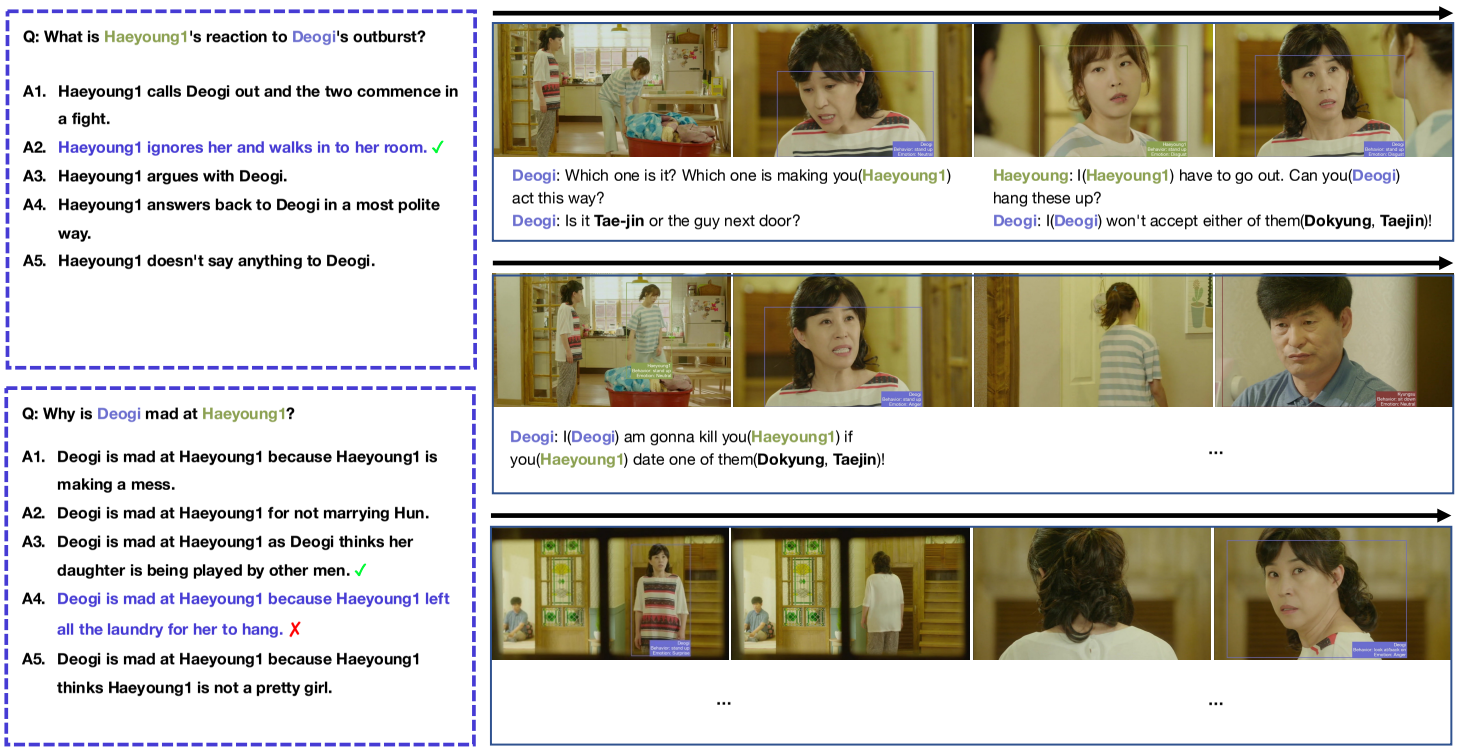

Figure 5.1 Inference example on the validation set. A subset of frames in the video along with character bounding boxes and annotations can be seen on the right. Subtitles are below the video frames. Predicted answers are highlighted in blue. Correct answers are marked with ✓ whereas incorrect predictions are marked with ✗.

Figure 5.2 Inference example on the validation set. A subset of frames in the video along with character bounding boxes and annotations can be seen on the right. Subtitles are below the video frames. Predicted answers are highlighted in blue. Correct answers are marked with ✓ whereas incorrect predictions are marked with ✗.

Figure 5.3 Inference example on the validation set. A subset of frames in the video along with character bounding boxes and annotations can be seen on the right. Subtitles are below the video frames. Predicted answers are highlighted in blue. Correct answers are marked with ✓ whereas incorrect predictions are marked with ✗.

mention of using make up. The second question asks for what Deogi is doing in the video clip. This question also requires information from the subtitles while the fact that she is in the kitchen and holding a pan in some frames needs to be ignored. Our model correctly predicts the answer (i.e. she is asking a question) from the given inputs.

The last example can be seen in Figure 5.3 and shows an example of a form of bias present in the DramaQA dataset. The question asks for the name of the character who performs a simple action (i.e. putting the glasses back on). As the answer to this is not mentioned in the subtitles, the model should pay attention to the video to infer the correct answer. We find that our model correctly infers the person performing the action. However, it is worth noting that this particular example exhibits some bias and the answer can be easily inferred from the meta features due to the fact that only Haeyoung1 and Sungjin appear in the video frames. It is then easy to deduce that the one performing the action is in fact Sungjin as none of the other answer options contain a person that appears in the frames.

5.4 Ablation study and additional experiments

We perform an ablation studies as well as several experiments to see which components and modalities contribute most to the scores achieved by our model.

Ablation study. In the ablation study we only train the model using a subset of the features provided in the DramaQA dataset. Results of the ablation study can be seen in Table 5.4. All models that do not include video but only the language modality (i.e. Question + Answer + Meta) do not use our co-attentional transformer module but instead simply compute the score directly on the language representations obtained from RoBERTa. Interestingly it can

	Overall	Difficulty Avg.
MCM (Choi et al., 2021)	0.7201	0.6421
Discriminative	0.7207	0.6654
Question + Answer	0.5630	0.5566
+ Subtitles	0.6410	0.6398
+ Meta	0.7403	0.6980
+ Meta + Video	0.7462	0.7036
+ Meta + Video + Subtitles	**0.7743**	**0.7360**

Table 5.4 Ablation study on the DramaQA validation set.

	Overall	Difficulty Avg.
$K = 1$	**0.7743**	**0.7360**
$K = 2$	0.7699	0.7300
$K = 3$	0.7673	0.7204
$K = 1$ (RoBERTa frozen†)	0.7512	0.7054
$K = 2$ (RoBERTa frozen†)	0.7521	0.7097
$K = 3$ (RoBERTa frozen†)	0.7535	0.7085

Table 5.5 Experiments with different numbers K of co-attentional Transformer layers on the DramaQA validation set. Rows marked with † indicate that the first 10 layers of RoBERTa have been frozen and only the last two layers are being fine-tuned. In all other rows RoBERTa is being fine-tuned fully.

be seen that the model using only the question, the answer as well as the meta features already outperforms the baseline model which explicitly uses all modalities to compute answer scores. This can be attributed to the fact that there may be some amount of bias in the questions with regard to emotions and actions by the characters that are given in the meta features which the powerful language model is then able to fully take advantage of. We can also see that meta features are indeed more useful than the subtitles of the dialogue which can be attributed to the same as well as the fact that subtitles are likely to be more noisy and lengthy. We note that all modalities contribute to the total score obtained by our model. However, despite the fact that the model is able to answer more complex questions it is clear that the model is still not able to fully utilize dependencies between the modalities and future research is needed to develop models that can fully take advantage of interactions between different modalities.

Number of co-attention layers. We experiment with a varying number K of co-attention layers in Figure 5.5 to find the optimal number of layers. We evaluate with $K \in \{1, \ldots, 3\}$ both while fully fine-tuning RoBERTa and only fine-tuning the last two layers of RoBERTa and freezing the first 10 layers. We find that regardless of the number of layers, fine-tuning the entirety of RoBERTa works better than not fine-tuning RoBERTa. Additionally we find that performance of our model is highest with $K = 1$. The rest of the experiments in this sections are performed with $K = 1$ and fine-tune all layers of RoBERTa.

Meta feature encoding. We evaluate different ways of encoding the meta features which are provided in the DramaQA dataset along with the visual

	Overall	Difficulty Avg.
Meta feature sentences	**0.7743**	**0.7360**
Meta feature words	0.7691	0.7314
Numerical meta features	0.7623	0.7312

Table 5.6 Experiments with various textual and numerical encodings of meta features on the DramaQA validation set.

	Overall	Difficulty Avg.
Single-stream	0.7597	0.7156
Ours, two-stream	**0.7743**	**0.7360**

Table 5.7 Comparison of our two-stream co-attention Transformer approach to a simple single-stream Transformer on the DramaQA validation set.

bounding boxes in language. Meta features are annotations describing an action (i.e. "drink", "eat", "dance") as well as an emotion (i.e. "fear", "happiness", "neutral"). To better infer "who is doing what and feeling how" we transform them to sentences by including the character name with both the action and the emotion. We compare simple word sequences consisting of the simple concatenation of name, behavior and emotion, i.e. "Doegi standing up sadness.", transforming them into full sentences, i.e. "Doegi is standing up and feeling sadness." as well as encoding them numerically, i.e. "7 12 23" to probe to what extent the language model priors help in leveraging the meta features. Results can be seen in Table 5.6. We observe that in overall accuracy meta features as full sentences perform only slightly better than encoding them simply as words. Only encoding them numerically performs slightly worse. In terms of difficulty average all methods perform similarly.

Single-stream transformer. To gain a deeper understanding of the performance of our two-stream co-attentional Transformer approach we compare

it with a single-stream (joint type) cross-modal Transformer encoder similar to those seen in prior work (Chen et al., 2020b, Luo et al., 2020, Sun et al., 2020). Specifically, we encode both visual and language stream in the same way as described in Chapter 4 but replace our co-attention module with a regular Transformer. Scores are also computed similarly on the resulting single output stream. Results can be seen in Table 5.7. We observe that our two-stream model outperforms the single-stream approach by around 1 percentage point overall and 1.56 percentage points in difficulty average demonstrating the effectiveness of emphasizing vision and language scores equally.

Chapter 6

Conclusion and Future Work

In this work we have performed a comprehensive survey of works lying at the intersection of vision and language learning, dealing with visual question answering, video question answering, vision and language representation learning, and video representation learning. Based on our findings, we have introduced a novel two-stream co-attentional Transformer architecture for story-based video understanding that successfully learns long-term dependencies present in video stories as well as cross-modal relationships. We have evaluated our architecture in a video question answering setting with character-centered annotations and questions on the DramaQA dataset. Our model outperforms the Multi-Stream Context Matching baseline model on every difficulty level by at least 3.8 and up to 12.1 percentage points on higher difficulty levels that require more complex reasoning. Moreover, our method beats the winners of the DramaQA challenge held at ECCV 2020. To gain a deeper understanding of how our method works, we have presented several qualitative examples along with an ablation study and additional experiments demonstrating the efficacy of our architectural choices.

6.1 Future work

Possible future directions for this work include the use of BERT-like pre-training strategies such as seen in e.g. ViLBERT (Lu et al., 2019), ERNIE-ViL (Yu et al.,

2020), LXMERT (Tan and Bansal, 2019), and UNITER (Chen et al., 2020b) for single-image vision and language learning as well as e.g. UniVL (Luo et al., 2020), VideoBERT (Sun et al., 2019), and CBT (Sun et al., 2020) for video representation learning. We believe such a pre-training strategy may help to obtain better fused representations before applying them to downstream tasks such as video question answering and video story understanding. Similarly, it may be useful to employ proxy tasks such as masked character prediction or masked object prediction to better align visual and lingual representations.

Similarly to the approach taken by ERNIE-ViL (Yu et al., 2020) it may be possible to construct and predict scene graphs that describe what is happening in the visual domain from video descriptions; this can either be on a fine-grained level or on a more coarse level depending on the video descriptions provided with the latter being more appropriate for the video story understanding setting. By masking and predicting entities or relations in the resulting scene graph models may be able to better align vision and language representations.

Recent work (Chen et al., 2020a, Sun et al., 2020) has also explored the use of contrastive learning for video and multimodal learning and it may be viable in the context of video story understanding and video question answering as well.

Lastly, we believe a more structured approach using spatiotemporal scene graphs to represent who is in the scene, what their intentions are and who and what objects they interact with throughout the scene may aid in video story understanding as well. We believe this may especially be useful in settings with less training data (e.g. the DramaQA, TVQA datasets) as graphs are able to make use of the inherent structures in video data such as characters remaining consistent across frames, but also across modalities where a name in language refers to the same character seen in the frame. However, in order to

apply graph representations to such multimodal settings, some questions will have to be resolved: 1) What is the best way to encode language features (e.g. from subtitles) in graph form? 2) How can we use graphs to fuse the different modalities? 3) How can we do inference on this graph to obtain the correct answer? We hope to address these questions in future work.

Bibliography

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. URL https://ieeexplore.ieee.org/document/8578734. Cited on pages 6 and 9.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. URL https://arxiv.org/abs/1505.00468. Cited on pages 1 and 9.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1409.0473. Cited on page 6.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020a. URL https://arxiv.org/abs/2002.05709. Cited on page 54.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt

Representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 104–120. Springer, 2020b. URL https://arxiv.org/abs/1909.11740. Cited on pages 1, 15, 39, 52, and 54.

Seong-Ho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Min Su Lee, and Byoung-Tak Zhang. DramaQA: Character-centered video story understanding with hierarchical QA. In *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 1166–1174, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16203. Cited on pages vii, 24, 25, 26, 27, 28, 32, 40, 41, 42, and 49.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017. URL https://arxiv.org/abs/1611.08669. Cited on pages v, 11, and 42.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1423. Cited on page 8.

Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages

6576–6585, 2018. URL https://arxiv.org/abs/1803.10906. Cited on page 20.

Shijie Geng, Ji Zhang, Zuohui Fu, Peng Gao, Hang Zhang, and Gerard de Melo. Character matters: Video story understanding with character-aware relations. *arXiv preprint arXiv:2005.08646*, 2020. URL https://arxiv.org/abs/2005.08646. Cited on pages 2, 23, and 24.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://arxiv.org/abs/1612.00837. Cited on pages 10 and 13.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. URL http://proceedings.mlr.press/v9/gutmann10a.html. Cited on page 17.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. URL https://arxiv.org/abs/1512.03385. Cited on pages 20 and 32.

Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*

(AAAI), pages 11021–11028, 2020a. URL https://aaai.org/ojs/index.php/AAAI/article/view/6737. Cited on page 21.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020b. URL https://arxiv.org/abs/2004.00849. Cited on page 1.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766, 2017. URL https://arxiv.org/abs/1704.04497. Cited on pages 21 and 25.

Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Gaining extra supervision via multi-task learning for multi-modal video question answering. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. URL https://ieeexplore.ieee.org/document/8852087. Cited on pages 2 and 24.

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. URL https://dl.acm.org/doi/10.1007/s11263-016-0981-7. Cited on page 22.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1369–

1379, 2018. URL https://www.aclweb.org/anthology/D18-1167/. Cited on pages 2, 22, 26, 29, 40, and 41.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8211–8225, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.730. Cited on pages 2, 23, and 26.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pages 11336–11344, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6795. Cited on pages 1 and 14.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692. Cited on pages 8 and 34.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NIPS)*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf. Cited on pages 1, 3, 14, 31, and 53.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li,

Xilin Chen, and Ming Zhou. UniVL: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. URL https://arxiv.org/abs/2002.06353. Cited on pages vii, 11, 12, 17, 39, 52, and 54.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9879–9889, 2020. URL https://arxiv.org/abs/1912.06430. Cited on page 17.

Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. In *16th European Conference on Computer Vision (ECCV)*, pages 223–240, 2020. URL https://arxiv.org/abs/1911.11390. Cited on pages vii, 3, 10, 11, and 31.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL https://www.aclweb.org/anthology/D14-1162/. Cited on page 21.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D16-1264. Cited on page 8.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, page 568–576, Cambridge, MA, USA, 2014. MIT Press. URL https://papers.nips.cc/paper/2014/file/00ec53c4682d36f5c4359f4ae7bd7ba1-Paper.pdf. Cited on page 21.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygXPaEYvH. Cited on pages 1 and 13.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. URL https://arxiv.org/abs/1904.01766. Cited on pages 16 and 54.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer, 2020. URL https://openreview.net/forum?id=rJgRMkrtDr. Cited on pages 17, 39, 52, and 54.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, November 2019. Association for Computa-

tional Linguistics. URL https://www.aclweb.org/anthology/D19-1514. Cited on pages 1, 3, 15, and 54.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. URL https://ieeexplore.ieee.org/document/7410867. Cited on page 20.

Trieu H Trinh and Quoc V Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018. URL https://arxiv.org/abs/1806.02847. Cited on page 40.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Cited on pages 6 and 7.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. URL https://link.springer.com/chapter/10.1007/978-3-030-01267-0_19. Cited on page 16.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference*

on Machine Learning, pages 2048–2057. PMLR, 2015. URL https://dl.acm.org/doi/10.5555/3045118.3045336. Cited on page 6.

Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. Videoqa: question answering on news video. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, pages 632–641, 2003. URL https://dl.acm.org/doi/10.1145/957013.957146. Cited on page 19.

Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. BERT representations for video question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1556–1565, 2020. URL https://ieeexplore.ieee.org/abstract/document/9093596. Cited on pages 2, 23, and 24.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020. URL https://arxiv.org/abs/2006.16934. Cited on pages 15, 53, and 54.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, pages 69–85. Springer, 2016. URL https://link.springer.com/chapter/10.1007/978-3-319-46475-6_5. Cited on page 14.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6720–6731, 2019. URL https://ieeexplore.ieee.org/document/8953217. Cited on page 13.

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, page 4334–4340, 2017. URL https://dl.acm.org/doi/abs/10.5555/3298023.3298196. Cited on pages 19 and 20.

Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3518–3524, 2017. URL https://www.ijcai.org/proceedings/2017/492. Cited on page 20.

Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421, 2017. URL https://dl.acm.org/doi/10.1007/s11263-017-1033-7. Cited on pages 19 and 20.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, page 19–27, USA, 2015. IEEE Computer Society. ISBN 9781467383912. URL https://ieeexplore.ieee.org/document/7410368. Cited on page 40.

국문초록

비전과 언어 학습의 최근 동향에 영감을 받아, 우리는 비디오 스토리 이해의 응용 프로그램 내에서 시각 언어 융합을 위한 co-attention mechanism 적용을 탐구한다. 다른 비디오 질문 답변(QA) 작업과 마찬가지로, 비디오 스토리 이해는 에이전트가 복잡하게 얽혀 있는 시간에 따른 의미 의존성을 파악해야 한다. 그러나, 비디오의 서술적 측면에 초점을 맞추면서, 에이전트는 또한 다른 등장인물들 사이의 상호작용과 그들의 행동과 동기를 이해하여야만 한다. 본 논문에서는 자연어 처리에서 필수적인 개념(예: multi-head attention 다중 헤드 주의)을 소개하고 시각적 질문 답변(visual question answering), 시각적 표현 학습(visiolingual representation learning), 비디오 표현 학습 및 비디오 질문 답변(video representation learning and video question answering)과 같은 인접 분야의 관련 작업에 대한 포괄적인 조사를 수행한다. 우리의 연구 결과를 바탕으로 우리는 드라마와 같은 시각적 스토리에서 보이는 시간에 흐름에 따른 장기적인 의미 의존성을 더 잘 포착하고 비디오 질문 답변 설정에서 비디오 스토리 이해 작업에 대한 성능을 측정하기 위한 새로운 공동 주의 트랜스포머 모델(novel co-attentional Transformer model)을 제안한다. 우리는 최근에 소개된 인물 중심의 비디오 스토리 이해 질문을 특징으로 갖는 Drama QA 데이터 세트에 우리의 새로운 모델을 적용해 평가한다. 우리 모델(∼77% 정확도)은 기본 모델(∼71% 정확도)대비 전체적으로 6% 이상 정확도가 높았으며, 모든 어려운 난이도 문제에서도 모든 다른 모델보다 최소 3.8%, 최대 12.1% 정확도 이상을 능가하였다. 이는 기존의 Drama QA challenge에 제출되었던 모든 우수한 모델의 성능을 능가함을 확인했다.

주요어: 비디오 스토리 이해, 공동 주의, 비디오 질문 답변, 다중모델 학습법
학번: 2019-21343