Weight-based multi-stream model for Multi-Modal Video Question Answering

Mohith Rajesh **Sanjiv Sridhar** **Chinmay Kulkarni**

Aaditya Shah **Dr. Natarajan S**

PES University
Bengaluru, India
{mohithraj9301, sanjivsridhar01, camkoolkarni, aadishah998 }@gmail.com, natarajan@pes.edu

Abstract

There has been a tremendous success in individual domains of Computer Vision, Natural Language Processing, and Knowledge Representation. Videos are a rich source of information with the multi-modal data forms of images, audio, and optionally subtitles blended. Current research is going on in combining these individual domains which have given rise to topics such as image captioning, visual question answering, and video question answering. Video Question Answering is a model which combines research topics like object detection and recognition, temporal information processing, visual attention, and natural language processing.

In this paper, we propose a model with Attention Mechanism for Video Question Answering that assigns varying weights to the many pieces of information the video encompasses. The model combines the question with 3 streams i.e., video's frames, subtitles, and objects to get the most probable answer. The model also receives the set of answer candidates as input and predicts one of them as the most probable answer since it has been trained and tested on the TVQA dataset.

1 Introduction

We are particularly interested in how algorithms produce pertinent and convincing natural language that can describe videos and images. Such multimodal algorithms can be fine-grained and evaluated using the **Visual Question Answering (VQA)** task.

VQA systems generate responses to an image (or video) and pertinent natural language queries as input. We can assess various facets of a model's multimodal semantic understanding by asking algorithms to respond to numerous questions, from questions about object recognition, counting, motion, or appearance to more challenging questions about interactions, social relationships, or inferences concerning the cause or manner of something. Normally, correlations between motion and appearance information can provide each other with helpful attention cues.

Recent advancements in image and text question answering are significant. **Video Question Answering (Video QA)** is challenging task and is different from image QA, as the questions focus more on the videos' temporal reasoning, such as motion transition and action counting, than its spatial attributes, such as colors, sizes, shapes and spatial locations, which call for efficient temporal representation modelling. Also, the videos contain richer information to remember in terms of quantity and variety (appearance, motion, transition), thus complicating the reasoning process.

We propose a model that is based on the idea of an attention mechanism. **There are specifically three important aspects:** 1) Deferentially weighs various streams of the video; 2) Effectively utilises pre-trained models, which, in comparison to SOTA models, greatly reduces the number of trainable parameters; 3) A straightforward, understandable, and yet efficient model.

2 Literature Survey

Attention based Methodologies

(Gao et al. 2019) presented a Structured Two-stream Attention network, known as STA, to respond to a natural language question in the form of a free-form or open-ended question regarding the content of a certain video.

To better capture global contexts in video frames, complex semantics in questions, and their interactions, (Fan et al. 2019) present a unique method, an "end-to-end trainable Video Question Answering (VideoQA)" framework consisting of three core components and creating new external memory modules.

(Gao et al. 2018) proposed Motion: Appearance Co-Memory Networks for Video Question Answering, are based on Dynamic Memory Network concepts (DMN) (Kumar et al. 2016). By using two-stream models, the video is transformed into a series of motion and appearance attributes.

Using the multimodal data preprocessing techniques

One of the first works to process the video using its compressed features was done by (Kim, Ha, and Kang 2021). In order to reduce complexity, they created a novel deep neural network to provide video QA features derived from coded video bit-stream. A few anchor frames, known as "intra-coded frames" (I-frame), can be used to rebuild a series of frames after video compression.

Using more information for predicting answers is useful most of the time; (Kim, Tang, and Bansal 2020) used a mechanism of dense captioning to describe the frame more effectively and use the caption as a piece of additional information to predict the answer to the question.

Graph based techniques

Since we are dealing with spatial and temporal dependencies graphs can help establish these dependencies very well and the work by (Seo et al. 2021) presents the same. Object graphs are constructed via graph convolutional networks (GCN) to compute the relationships among objects in each visual feature.

Question/Query Processing

(Wang, Huang, and Wang 2020) aim to tackle long video question answering with their novel Matching guided Attention Mechanism (MAM) which comprises 3 modules: 1) the video and question embedding module to convert the video and question into embeddings, 2) the video content localization module to locate the parts of the video relevant to answering the question asked and 3) the answer prediction module where these selected number of frames are used to determine the answer to the question asked.

3 Dataset

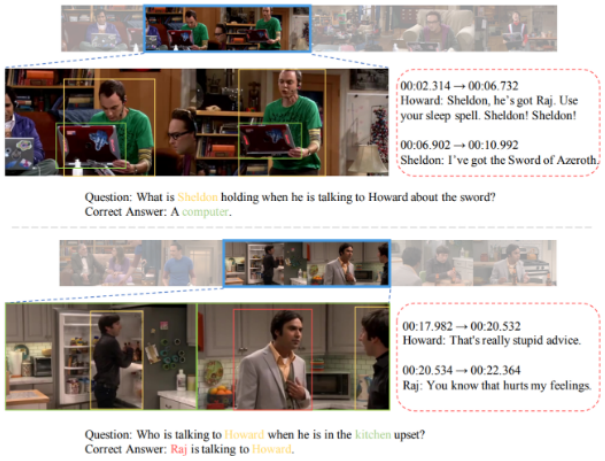

Figure 1: Examples from the TVQA dataset

We are utilizing the **TVQA** dataset (Lei et al. 2019), which is based on 6 well-known Television shows from 3 different genres: crime, sitcoms, and medical dramas. 152.5K human-written QA pairs are gathered from this data. The dataset's multimodal compositionality is its main focus. In particular, a novel large-scale dataset is gathered based on real-world video content with rich dynamics and realistic social interactions.

The dataset offers four significant advantages. First of all, it contains 21,793 video clips from 925 episodes and is natural and extensive. Each show has 7.3 seasons on average, allowing for extensive character interactions and changing relationships. Seven questions are associated to each video clip, and each question includes five answers (1 correct). Second, since video clips are usually between 60 and 90 seconds in length, there are often more social interactions and activities present, which makes it harder to grasp the video. Third, each QA video clip comes with the dialogue (character name + subtitle). For many of the questions gathered, it is essential to accurately understand the relationships between the dialogue, video frames, and question-answer pairs. Fourth, algorithms are needed to localize pertinent moments since questions are compositional.

Three tasks with significant annotation are made possible by the dataset: "Question-driven moment localization", "QA on the grounded clip", and "QA on the complete video clip".

4 Preprocessing

Subtitle Parsing

The dataset contains subtitles along with video frame data. This is another mode of useful information that can be used to train the model better.

In the subtitles, the speaker's name is mentioned in the bracket at the start and their dialog is mentioned after. These are parsed out and converted to a sentence. This process is repeated for the entire subtitle file and the sentences generated are concatenated to form a paragraph that will be used by the model.

If the subtitle ended with a full stop (.), then the generated sentence is of the format: *Speaker* **said**, *dialog*, if subtitle ended with question mark (?) then it is *Speaker* **asked**, *dialog*. When the speaker was not specifically stated, the word *Someone* was used in place of the speaker's name.

An example conversion would be as follows:

- Subtitles

- (Beckett:)Who are you?
- (Javier:)It's Alison.
- I see.

- Converted paragraph

Beckett asked, Who are you? Javier said, It's Alison. Someone said, I see.

5 Methodology

The size of the dataset, which includes subtitles and video clips with an average frame count of 300, makes it necessary for the model's trainable parameters to be numerous. This has been countered by using pre-trained models like **CLIP-ViT-L-14** for frame encoding and the **DeBERTa-v2-xlarge** language model to obtain the word embedding for each word in the subtitle, question, and answer. This not only helps to create a lighter model but also encodes rich information of frames and words because they learned to encode features during the pretraining phase.

For different information of video different pre-trained models are used : 1) CLIP ViT for frames; 2) DeBERTa for subtitles; 3) **GloVe** for objects detected in Videos. Encodings of each of the above mentioned pre-trained models forms three streams and they each contribute in model prediction.

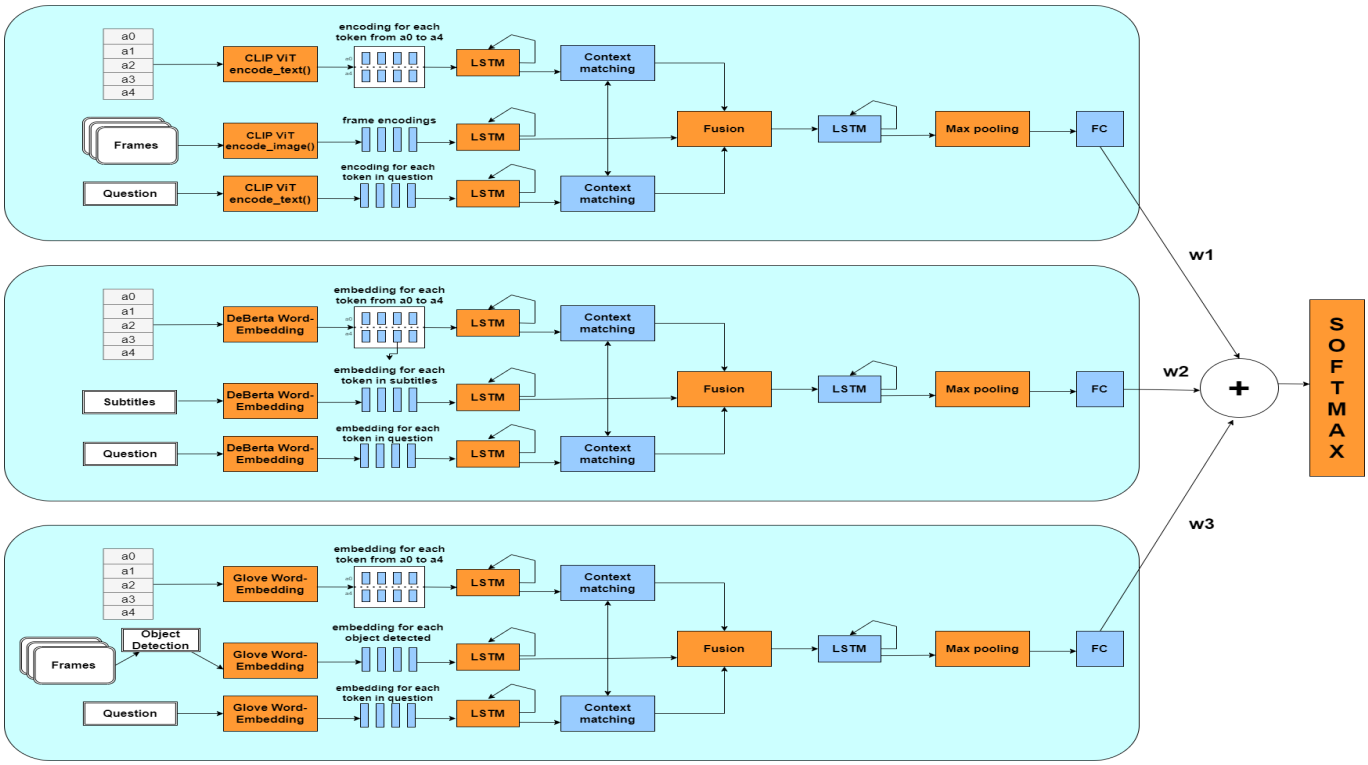

Figure 2: Illustration of our multi-stream model for Multi-Modal Video QA.

CLIP-ViT-L-14

CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on several image and text pairs. Without actively optimizing for the task, it can be instructed in natural language to predict the most pertinent text excerpt given an image. This zero-shot capability assisted in solving numerous significant challenges in computer vision. We have leveraged this capability in our approach to solve Video Question Answering.

Pre-trained CLIP model provides two methods: one to encode image features and another to encode text features; it is pre-trained to minimize the Contrastive Loss of Image-Text pairs. In other words, the text encoding and image encoding should be analogous if the text accurately describes the image, and divergent if it doesn't. This objective learned by the model is particularly important for us for the joint reasoning of video frames and text.

DeBERTa-v2-xlarge

Language models facilitate transfer learning, which is a particularly effective method of learning. Transfer learning facilitates the ability to train a model on a large amount of data, frequently unsupervised, and then use that pre-trained model to effectively learn downstream tasks. We have used DeBERTa language model to obtain the word embeddings of subtitles.

GloVe

We first detect the objects in the video then use the GloVe model to get the embedding of the each unique object that is identified in the video. We prefer to use GloVe over language model like BERT or DeBERTa due to the contextual nature of the BERT embeddings hindering the model by sequentially processing the intrinsically unordered objects detected.

The 3 pre-trained models stated above are utilised in three different streams:

- **Stream 1**: The Video Frames, question, set of answer candidates are encoded using the CLIP ViT. It's worth noting that the encodings for question, set of answer candidates are created for each word rather than one encoding for question or one encoding for each answer candidate.
- **Stream 2**: The first layer output of the DeBERTa Language Model is used to acquire the embedding of each token in the subtitles, question, and answer candidates.
- **Stream 3**: Using OpenCV, the objects in the video are detected; these text-based detected objects and question-and-answer candidates are encoded using GloVe. Again, this encoding is done for each word, as opposed to the complete entity.

Joint Modeling of Context and Query

To represent the contextual inputs (subtitle, video frames) and query (question-answer pair), we employ a **LSTM**

and "context-matching module", the latter is based on earlier works' (Vaswani et al. 2017) **context-query attention layer**. It accepts both context vectors and query vectors as inputs, and then generates a collection of context-aware query vectors based on how similar each context-query pair is.

Considering the subtitles feature stream (Fig 2) as an example, maxlen for subtiles is defined as 512 tokens, 40 for question and 30 for each answer candidate, after passing the embedding obtained using DeBERTa to the first LSTM (Fig 2) we obtain subtiles embedding, ${H_{sub}}^{512\times300}$ (superscript represent the dimension excluding the batch-size) which is used as context input, question embedding, ${H_q}^{40\times300}$ and answer embedding, $\{{H_{a_i}}^{30\times300}\}_{i=0}^{4}$, where the question and answer embedding are used as queries.

We pass the context-query pairs ($H_{sub} - H_q$ and $\{H_{sub} - H_{a_i}\}_{i=0}^{4}$) into the context-matching module, we obtain a subtitles-aware-question representation, ${G_{sub,q}}^{512\times300}$, and subtitles-aware-answer representation, $\{{G_{sub,a_i}}^{512\times300}\}_{i=0}^{4}$, which are then fused with subtitles context using Fusion Layer (eq 1).

$$\{\mathrm{M}_{sub,a_i} = [H_{sub}; G_{sub,q}; G_{sub,a_i}; H_{sub} \odot G_{sub,q}; H_{sub} \odot G_{sub,a_i}]\}_{i=0}^{4} \quad (1)$$

Where $\odot$ is the element-wise product. The fused feature, $\{{M_{sub,a_i}}^{512\times1500}\}_{i=0}^{4}$ is fed into another BiLSTM, the vector, $\{{U_{sub,a_i}}^{600}\}_{i=0}^{4}$ is obtained by max pooling the output of LSTM at each timestamp, ${U_{sub,a}}^{5\times600}$ is obtained by concatenating the ${U_{sub,a_i}}^{600}$ for each i=0 to i=4 which is followed by a pair of fully connected layers with 500 and 1 hidden units, both with dropout 0.5 and ReLU activation. Thus transforming $U_{sub,a}$ from $R^{5\times600}$ to $R^{5\times500}$ to $R^{5\times1}$, as finally it is squeezed to obtain the 5-dimensional output which represents a vote for each answer.

Similarly other two streams shown in (Fig 2) also go through the same process, the maximum video frames is set to 300 and maximum objects detected is set to 20 and finally weighted average of R^5 vector from each Stream is computed and passed to softmax layer to obtain the answer probabilities.

6 Experiments and Results

The dataset has been converted into TensorFlow dataset objects which allows getting the slices of an dataset in the form of objects directly from the disk. The model is trained with AdamW optimizer (lr=3e-4, weight_decay=1e-5), to minimise the categorical cross-entropy loss.

The weights W1, W2 and W3 mentioned in (Fig 2) is set to 0.2, 0.6, 0.2 respectively, since the model with subtitle stream alone performed much better than the model with only Objects or Frames streams, the weight for subtitles stream is more than 0.5. The model achieved the validation accuracy of **68.07%**.

The table below illustrates a comparison between the proposed model and other works(Lei et al. 2019; Le et al. 2021; Li et al. 2020; Zellers et al. 2021):

Model	No. of parameters	Val Acc
TVQA(S+V+Q)	-	66.46%
Our Model	17.5M	68.07%
HCRN	44M	71.40%
HERO	119M	73.60%
MERLOT	223M	78.70%

By assigning varying weights to various video streams and effectively utilising the pre-trained model for each stream, it can be observed that the proposed model outperformed the model that was proposed along with the TVQA dataset (Lei et al. 2019).

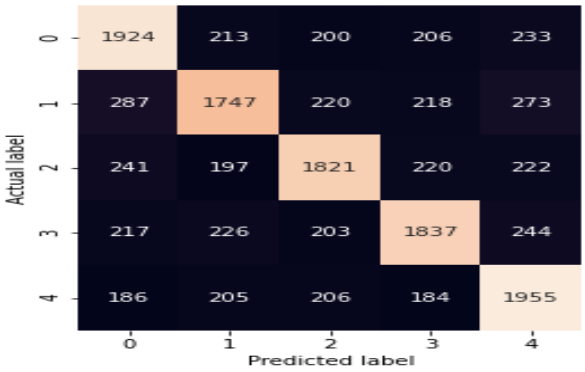

Figure 3: Confusion Matrix

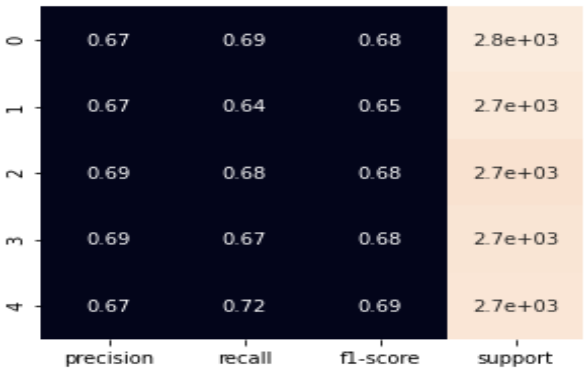

Figure 4: Classification Report

7 Conclusion and Future Work

Even if the model results are satisfactory, there are a few benefits to this model: 1) The number of trainable parameters (**17.5M**) are substantially lower than that of SOTA models; 2) The model's architecture is simpler, making it simple to add any additional video information as a new stream.

In the proposed experiment setting, the weights W1, W2, and W3 mentioned in (Fig 3) are fixed; however, they can be a trainable parameter that could enhance the results. There is a lot of research being done on improving the performance of the language and vision models, thus just replacing the pre-trained models utilized in the proposed method with the new SOTA pre-trained models can boost performance.

References

Fan, C.; Zhang, X.; Zhang, S.; Wang, W.; Zhang, C.; and Huang, H. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 1999–2007.

Gao, J.; Ge, R.; Chen, K.; and Nevatia, R. 2018. Motion-appearance co-memory networks for video question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 6576–6585.

Gao, L.; Zeng, P.; Song, J.; Li, Y.; Liu, W.; Mei, T.; and Shen, H. 2019. Structured two-stream attention network for video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01):6391–6398.

Kim, N.; Ha, S. J.; and Kang, J. 2021. Video question answering using language-guided deep compressed-domain video feature. *Proceedings of the IEEE/CVF International Conference on Computer Vision* 1708–1717.

Kim, H.; Tang, Z.; and Bansal, M. 2020. Dense-caption matching and frame-selection gating for temporal localization in videoqa. *arXiv preprint arXiv:2005.06409*.

Kumar, A.; Irsoy, O.; Ondruska, P.; Iyyer, M.; Bradbury, J.; Gulrajani, I.; Zhong, V.; Paulus, R.; and Socher, R. 2016. Ask me anything: dynamic memory networks for natural language processing. *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)* 1378–1387.

Le, T. M.; Le, V.; Venkatesh, S.; and Tran, T. 2021. Hierarchical conditional relation networks for multimodal video question answering. *IJCV*.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2019. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696v2*.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. Hero: Hierarchical encoder for video+language omni-representation pre-training. In *EMNLP*.

Seo, A.; Kang, G.-C.; Park, J.; and Zhang, B.-T. 2021. Attend what you need: Motion-appearance synergistic networks for video question answering. *arXiv preprint arXiv:2106.10446*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30.

Wang, W.; Huang, Y.; and Wang, L. 2020. Long video question answering: A matching-guided attention model. *Pattern Recognition* 102:107248.

Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J. S.; Cao, J.; Farhadi, A.; and Choi, Y. 2021. Merlot: Multimodal neural script knowledge models. In *NeurIPS*.