11-777 Report 2: Related Work and Model Proposal by Team-Samosa

Bhavuk Sharma* **Debayan Ghosh*** **Rituraj Singh***
Saahithya Yendluri* **Wallace Dalmet***
{bhavuks, debayang, riturajs, syendlur, wdalmet}@andrew.cmu.edu

1 Related Work and Background

Our objective is to develop a Visual Question Answering system using the TVQA[1] dataset. That is, given a video V, subtitles S, a question q and 5 possible answers $\{a_i\}_0^4$, the model should pick the correct answer. In this section, we explore various datasets that are available for VQA, perform unimodal baseline analysis for comprehensive understanding of each modality, research prior work and recent techniques in the field. In contrast to the abundance of image and text based QA datasets, the datasets mentioned in the following section are similar to the TVQA dataset in terms of description, data format and purpose. On evaluating the ablation studies on prior literature, we observe unimodal baselines which show significant evidence for the importance of both visual and textual cues for a VQA task. In addition, we discuss creative techniques that can be leveraged to build a new VQA architecture.

1.1 Related Datasets

Visual QA datasets are generally composed of videos that are associated with descriptive text or dialog. Each video has a corresponding set of questions with either multiple choice, one-word or a single phrase answer. In addition, some datasets such as TVQA+ [2] provide bounding boxes for objects which can further be used to improve a VQA model.

The MovieQA [3] dataset is a large-scale QA dataset on movies. It consists of 14,944 multiple choice questions (MCQs) with five possible options. The key feature of this dataset is that it has significantly longer text as well as multiple sources of information like movie plots, subtitles, and scripts.

An important sub-task for a VQA model is to identify the moment in a video which holds the answer to a question and this can be a decisive factor for a model's performance. In DiDeMo [4] (Distinct Describable Moments), there are 10,000 videos with 40,000 pairs of expressions that refer to localized moments. The task for this dataset is to identify the moment corresponding to an input text query. The authors also present "Moment Context Network" [4] (MCN) that helps the model in differentiating between relative moments; for example, "The biker starts the race." needs the context for when the biker was at halt. In section 2, we have discussed the potential of MCN in our model design.

In the TGIF-QA[5] dataset, there are 72,000 animated GIFs and 165,000 corresponding QA pairs. In addition to the MC questions setting, the authors propose 3 new tasks. "Repetition count"; how many times a given activity is being repeated, "repeating action"; what activity is being repeated multiple times and "state transition"; change from one activity to another. Solving these 3 tasks can be incremental in determining how well the model features interpret a video.

TVQA and MovieQA datasets have a disadvantage that the videos are scripted, directed and enacted by professionals. In LifeQA [6], the authors present a new dataset with videos that capture real-life encounters. There are 275 videos with 2326 MCQs. In addition to developing a in-the-wild VQA dataset, the authors observe that the model proposed in TVQA fails to perform well on the LifeQA dataset which indicates an implicit bias in the models.

The Multimodal QA dataset [7] contains 22,918 questions that integrate information across free text, semi-structured tables and images. The dataset spans a plethora of topics like politics and science to movies, music and literature. A key feature of this dataset is that 35.7% of the questions require

[*] Everyone Contributed Equally – Alphabetical order

cross-modal reasoning, which enables us to reliably evaluate a model's ability to form inferences with cues from multiple modalities.

ActivityNet-QA [8] is based on the ActivityNet dataset [9], which consists of over 200 different classes of various activities being performed. ActivityNet contains about 140 videos per class, where classes range from household activities such as cooking and cleaning to outdoor exercises and sports. The authors of ActivityNet-QA selected a subset of 5800 videos uniformly from each of the 200 classes, and generated over 58,000 QA pairs through crowdsourcing. Since the authors use raw video data which includes audio, there has also been work on using both audio and visual features for the VQA task. [10]

The datasets described so far require manual annotations which is expensive and not scalable. The HowToVQA69M [11] is a large-scale dataset for VQA generated automatically without manual annotations consisting of 69,270,581 video clips, question and answer triplets (v, q, a). The generation process of this dataset employs a transformer based approach by applying them to speech transcripts of narrated videos from the large-scale HowTo100M [12] dataset.

1.2 Unimodal Baselines

It is imperative that we observe the influence of each modality in a cross-modal VQA design. It is possible that certain questions are better answered primarily by one modality while others may require fused cross-modal information. Analyzing unimodal baselines is a key indicator to a modality's significance in a QA task.

The Multitask learning [13] paper achieves an accuracy of 67.05% with video + subtitles, 44.42 % with videos only, and 64.63 % with subtitles only on the TVQA dataset. Next, the MMFT BERT [14] paper achieves an accuracy of 74.97% with video + subtitles, 53.08 % with videos only, and 70.79 % with subtitles only on the TVQA dataset. This shows that the subtitle features captures more information about the QA task in relation to the video features. However, the video features do compliment the subtitle features, leading to a better overall test accuracy for both the methods mentioned above.

Geng et al. [15] explored the relationship between different characters in the TVQA dataset to gain an understanding of video story plot. Through their Character-Aware Reasoning Network, a 66.23% accuracy was obtained by only using the subtitle modality and the accuracy was boosted by 2.62% when the visual objects were added. A further improvement was observed when character centric relationships were integrated to achieve an overall accuracy of 70.37% which demonstrated that learning the relationship of different characters in the dataset through the subtitle and the video modality is essential to obtain better results.

Winterbottom et al. [16] looked at the underlying unimodal bias in TVQA. The authors observed that the subtitle modality overshadowed the video modality in terms of information required to the answer the questions. Using BERT[17] with only subtitles, the model outperformed the TVQA baseline by a margin of 1.38% on the validation test. The authors analysed the dataset to find out samples that rely on multimodal information for answer prediction, and proposed using a subset of these samples to mitigate the bias.

In the paper, 'On the treasure of dialog' [18], Alamri et al, use KnowIT VQA [19] dataset and use dialog to summarize on a per scene and per episode basis to get a high level understanding of the video. Based on the ablation studies mentioned in the paper, using episode dialogue summary brings accuracy improvement of 5.4%. Significant increase in performance shows that dialog summaries have much required knowledge especially for scene knowledge based QA and hence can potentially be of use for our model.

Further, we study the influence of the audio modality from [20], where the authors see that addition of audio features generally improves model performance for the QA task. We observe that model performance improves by 0.13% when combined with dialog history and video features, indicating there is complementary knowledge between the video and audio signals. The paper also makes an interesting statement about how the order of the QA pairs is important since dialog history is essential for answering the current question and shuffling the QA pairs shows a reduced recall value.

1.3 Prior Work

This section discusses prior literature on VQA systems trained on TVQA and other related datasets.

In [21], the authors propose a dual memory structure that pinpoints specific temporal parts of each

Method	Valid Acc.	Test Acc.
PAMN [21]	66.38	66.77
Multi-task [13]	66.22	67.05
STAGE [2]	70.50	70.23
MSAN [22]	71.62	71.13
MMFT BERT [14]	73.55	72.89
BERT for VQA [23]	72.41	73.57
HERO [24]	72.75	74.8
iPerceive [25]	76.97	75.15
QHGN [26]	76.25	76.17

Table 1: Results on TVQA dataset

modality. The model is trained and tested on the TVQA and MovieQA datasets. They define "memory" as the average pool over a feed forward layer with video and subtitles as inputs, each having their own memory blocks. In addition, they developed a "Dynamic Modality Fusion" mechanism that uses soft-attention to adaptively fuse modalities conditioned on an input question; thereby determining the contribution of each modality. It achieves a test accuracy of 66.77 % on the TVQA dataset.

Khan et. al [14] achieve state-of-the-art performance on the TVQA dataset. The model heavily employs BERT encoders for video, subtitles, questions and answers. The generated representations are passed through a multimodal fusion transformer with multiheaded attention. They define "visual concepts" as list of labels in a video sequence that are detected using FasterRCNN which is encoded by visual BERT. The model achieves a test accuracy of 72.89 %.

In QHGN [26], the idea of embedding features in the quaternion space is introduced to represent multimodal information; video, subtitles, questions and answers. The objective is to realize multimodal and structural reasoning through a quaternion hypergraph network (QHGN). In simpler terms, a hypergraph represents the relationship between different visual objects. Clip-level objects are the vertices and the hyperedges connect if two vertices contain the same visual objects. Representing multiple modalities in the quaternion space opens an opportunity for quaternion operations such as Hamilton product for the fusion and interaction between different modalities and to extract structural reasoning. The model achieves a test accuracy of 76.17 % on the TVQA dataset surpassing MMFT.

In MSAN [22], the authors propose a Moment Proposal Network (MPN) that locates the most appropriate temporal moment from each modality (weighted) and a Heterogeneous Reasoning Network (HRN) that uses the localized moment to predict an answer using attention. Video representations are extracted through FasterRCNN trained on Visual Genome while text representations are obtained through BERT [17]. The model achieves a test accuracy of 71.13 % on TVQA.

[13] largely re-uses the TVQA model architecture. The author's propose reusing weights from intermediate layers of the Video/Subtitle QA task for Modality Alignment and Temporal Localisation. The paper then describes a Multi-Task learning based Multi-Task ratio scheduling method to learn 'easier' task earlier, which allows us to set the inductive bias at the beginning of training. This model achieves an accuracy of 67.05 % on the TVQA dataset.

Yang et al. [23] use a dual-flow BERT system for the task. The visual and subtitle features are obtained by using faster R-CNN and by tokenization respectively. These are separately concatenated with QA features and fed into two separate BERT models. These representations are summed and the prediction is computed using softmax. Different variations of the model created by altering the order of input sequences were tested on TVQA and the Pororo dataset [27]. The authors observed difference in accuracy when this order was changed. In case of TVQA, when using timestamped subtitles and video information only, all variations gave state-of-the-art performance, with the best model attaining test accuracy of 73.57%. The model performance dropped when full subtitles were used which the authors hypothesized this was occurring due to large length of subtitles and no explicit attention mapping between visual and subtitle features

'Relation-aware Hierarchical Attention framework for Visual Question Answering' [28] proposes a model on the TVQA+ [2] dataset which encodes the static and dynamic relationships and also the interactions between detected objects in the video. A encoder based on Graph Attention Network(GAT) is utilized to embed the static and semantic relationships between visual objects detected using a pre-trained R-CNN. These unimodal embeddings with QA and subtitle embeddings are fused by using hierarchical self-attention to get question-guided attention scores. This model outperforms STAGE by 2% and shows improvements because of better video understanding. Since this model based on Relation-aware Hierarchical attention(RHA) retains information about implicit relations and relative object position, it could aid us to get substantial performance improvements.

Table 1 shows the performances of various models that have been tested on the TVQA datasets. Our goal would be to use these metrics as comparison benchmarks.

1.4 Relevant Techniques

There are patterns as well as distinct approaches to how the problem of visual question answering can be solved. A common theme among these approaches is the use of visual features extraction using FasterRCNN as an example and text embeddings using BERT [17]. Depending on the model's design, these representations are fused at a later stage after passing through a sophisticated attention mechanism. Finally, a combination of fully connected layers and softmax cross-entropy are used for the final output and loss. In this section, we summarize some of the relevant techniques that are potentially helpful in building a new VQA architecture.

In [29], the authors devise "co-attention", a novel mechanism that interlinks images and questions. The image representation are responsible for tending to a given question's attention while the question representation tends to the image's attention; thereby creating an effect of "natural symmetry". This technique can be used in VQA models to interpret and attend to different regions of a frame and chunks of a question.

Gao et al. [30] and Zhuang et al. [10] use models containing memory modules, that encode input sequence information and remember relevant information that could aid answer prediction. Gao et al. use Dynamic Memory Networks with a novel dynamic fact ensemble technique. Here the authors dynamically construct and aggregate features on which attention is then computed. The work also implements a co-memory attention mechanism between motion and appearance. Gao et al. [31] have also proposed a DFAF framework that integrates cross-modal self-attention and cross-modal co-attention mechanisms to achieve effective information flows within and between the different modalities. The authors evaluate their model on the TGIF-QA dataset, outperforming existing models on all question classes. While [30] only used visual features, Zhuang et al. take a multi-channel approach, where the audio features are also fed into the network. The architecture uses of Attention Memory Units (AMUs) to fuse and then perform attention on the multimodal features. The model is evaluated on the TGIF-QA [5] and ActivityNet-QA datasets. The authors note similar performance to state-of-the-art, the model even outperforming them in some cases. However, the model struggled with counting tasks, which the authors attributed to the sampling of frames causing omission of information.

[32] proposes a jointly learnt visual and textual attention models to explore the fine-grained interaction between vision and language. The VQA specific model described in this paper referred to as r-DAN, performs visual and textual attentions using a joint memory which assembles the previous attention results and guides the next attentions. The model achieves a test accuracy of 69.0 % on the VQA dataset.

[33] builds upon previous SOTA VQA architectures, by developing 13 attention mechanisms and introducing a simplified classifier to the model. The main contribution of this paper was describing the steps through which one might combine and tune different attention models. This model achieves an accuracy of 64.78 % on the VQA 2.0 dataset.

In 'On the hidden treasure of dialog' [18], the authors use KnowIT dataset [19] and in addition to the plot summary and video, use raw dialog per scene, dialog summary per scene and dialog summary for a whole episode in addition to the video and audio. Each of the above input sources is followed by a transformer that attempts to answer questions based on each stream. Modality fusion is done using a multi-head self-attention block. This paper shows the importance of dialog history and shows possibilities of improvement on the TVQA model baseline since it worked so well with the KnowIT dataset.

In 'Object-Centric Representation Learning for Visual Question Answering' [34], the authors build a relational graph of objects in the videos, whose features and interactions are dynamically and conditionally inferred. The object's spatial and temporal lives are summarized into résumés. The model uses question-based temporal attention and Graph Convolutional Networks (GCN) to correlate between object neighbours and then uses reasoning engines like MACNet for predictions. The paper also uses contextual objects (using ResNet [35]) to

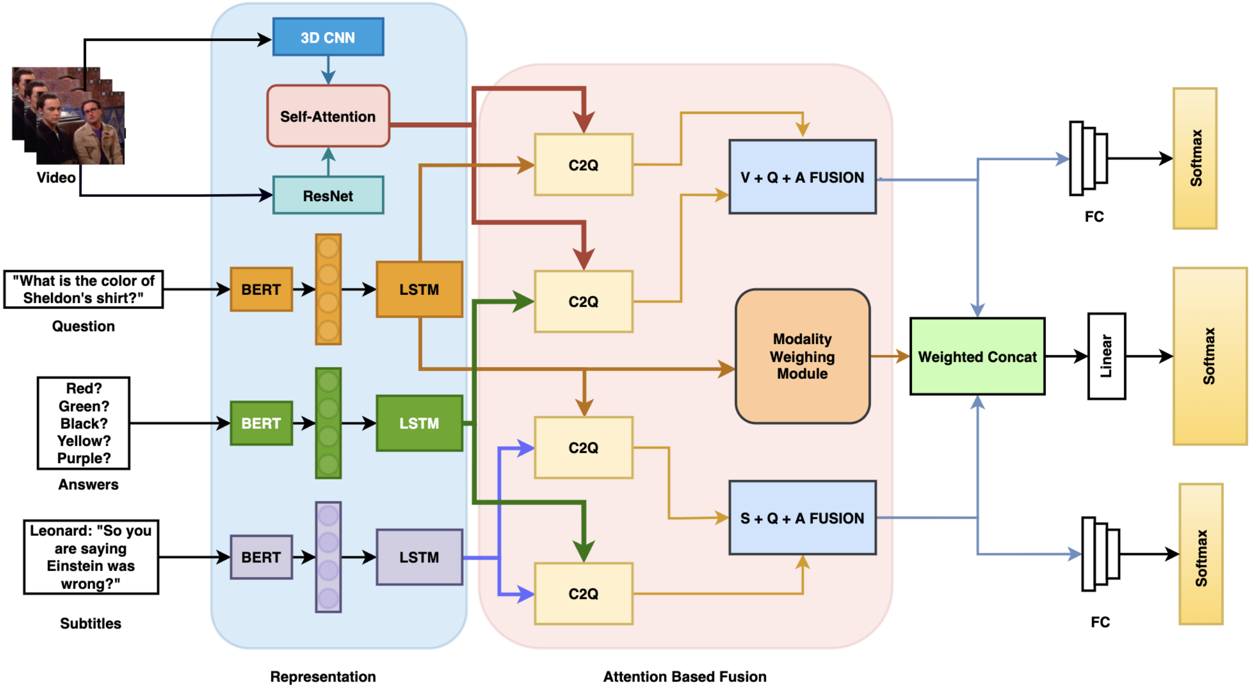

Figure 1: Proposed Model for VQA

account for any imperfections in the RCNN based detector. Contextual objects and object résumés are something that can be explored further if needed in our model to answer relational questions between objects.

2 Model Proposal

In this section, we propose our own architecture for visual question answering with inspiration drawn from the designs discussed in previous sections. Our goal for this design is to incorporate cross modal information with a mechanism to determine the influence of a particular modality. For example, "What is the color of Sheldon's shirt?" requires greater "attention" to the video encoding pipeline. However, questions such as "What did Sheldon say when Penny fell?" requires two sub-tasks that need to be solved; localizing the moment when the event of "falling" occurred and capturing the subtitle where Sheldon speaks. Naturally, this requires cross-modal fusion of the representations of the video and corresponding subtitle.

2.1 Overall Model Structure

Figure 1. shows the overall pipeline of our proposed model with more details of our network discussed in this section. Firstly, multi-source information from different modalities (video, subtitle, question and answers) will be feature extracted by pre-trained models. Secondly, we will employ attention based approaches to learn the intra- and inter-modality relations between the modalities and find aggregated features through fusion. Finally, a learned weighting mechanism will adjust the weight with which each modality is fused with its output passed through softmax for getting the final prediction.

2.2 Feature Representation

Video Representation: We plan to use ResNet-152 pre-trained on ImageNet to extract the spatial features and 3D convolution technique to capture the temporal features. An alternative for representing visual features is Faster-RCNN technique. However, Faster-RCNN primarily works well for object detection tasks whereas ResNet is known for its feature extraction capabilities and capturing patterns in an image. 3D convolution works well on video data having a third time dimension when compared to an image and will help to capture the lower level features in temporal domain. SlowFast networks [36] are also used to extract the video motion features, however, they require more computation and may end up learning similar features which can limit the performance of the model. Once we get the 3D CNN and

ResNet feature representation, we will apply self attention mechanism between them to learn the intra-modality relations between objects across various temporal video frames.

Text Representation: The sentences from the subtitles and QA pairs will be fine-tuned with the BERT[17] base model to obtain fixed d-dimensional word embeddings from its second-to-last layer. The extracted word embeddings will be put into a BiLSTM network, to capture the long term dependencies in the sentences. And the last hidden state of the BiLSTM will give the question, subtitles and candidate answers representations. Word2Vec and Glove word embeddings can also be used for text representation, however, they are context independent whereas BERT is well known for capturing context between sentences. GRU is an alternative for LSTM as it is less complex and faster, however, LSTM remembers longer sequences which is more suited for our task.

2.3 Attention Based Fusion

2.3.1 Context to Query (C2Q) module

Through the C2Q modules, we aim to find the (a) Subtitle-Question (b) Subtitle-Answer (c) Video-Question (d) Video-Answer Attended Embeddings.

Here, the "Attention Query" are the LSTM-Encoded Subtitle/Video Features, and the "Attention Key/Value" are the LSTM-Encoded Question/Answer Features. The context-to-query attention weights signifies which word in query (Question, Answer) is most relevant to each word in context (Subtitle, Video).

In theory, Self-Attention works better in terms of capturing dependencies and is more efficient than the traditional LSTM-Encoded Attention vectors. We would like to start off with the LSTM-Encoded Attention Vectors, and then extend it to self-attention based encoders, once our pipeline is up and running.

We could go further and use the Multi-Head Attention method which uses multiple iterations of computation to capture relevant information from different child spaces, which may lead to an overall better representation.

2.3.2 Modality Weighing Module

The input of this module are the LSTM-Encoded question embeddings. The goal of this module is to compute the Modality Contribution Factor α_s and α_v, where $\alpha_s = 1 - \alpha_v$. These weights will then be passed to the Weighted Concat block where the Subtitle and Video Embeddings will be fused.

2.3.3 V + Q + A Fusion Module

This block concatenates the (a) Video-Question and (b) Video-Answer attended Embeddings to form a V+Q+A Fusion Embeddings.

2.3.4 S + Q + A Fusion Module

This block concatenates the (a) Subtitle-Question and (b) Subtitle-Answer attended Embeddings to form a S+Q+A Fusion Embeddings.

2.4 Output and Loss Function

We pass the weighted concatenation of the two Fusion modules through a feedforward neural network. The output is computed using a softmax layer, which assigns probabilities to the five answer choices.

In our model we propose to use three losses: a loss for when using only video modality (L_v), when only using subtitles (L_s), and finally when using fused modalities (L_f). All of the above losses are cross-entropy losses. Finally, a weighted combination of the above three loses is taken to get a final prediction loss.

$$Loss = w_v L_v + w_s L_s + w_f L_f \tag{1}$$

where w_v, w_s, w_f are the weights for video only, subtitle only, fused modality based losses respectively.

An alternative to the cross-entropy loss that can be used is the LP-Focal Loss as described in [37]. This loss dynamically assigns lower weights for language-biases based on the question and thereby reduces this training instances contribution to the model. Hence, we believe that this could be a good alternative to the cross-entropy loss which equally computes loss for each instance.

Further, similar to [13], we can apply the principles of Multi-Task Learning to the Loss Function to learn 'easier' tasks earlier, i.e. we can give higher weightage w_v and w_s for the unimodal losses L_v and L_s in earlier epochs, and then give a higher weightage w_f to the multimodal loss L_f for later epochs.

3 Team Member Contributions

Bhavuk Sharma Papers read: [16], [8], [23], [30], [10], contributed to model design. Section 2.2, 2.3, 2.4.

Debayan Ghosh Papers read: [7], [3], [13], [32], [33], contributed to model design. Section 2.2, 2.3, 2.4

Rituraj Singh Papers read: [4], [5], [29], [38], [21], [14], contributed to model design. Section 2.2, 2.3, 2.4

Saahithya Yendluri Papers read: [6], [18], [34], [28], [20], contributed to model design. Section 2.2, 2.3, 2.4

Wallace Dalmet Papers read: [11], [26], [22], [31], [15], [12], contributed to model design. Section 2.2, 2.3, 2.4

References

[1] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018.

[2] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Tech Report, arXiv*, 2019.

[3] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. *CoRR*, abs/1512.02902, 2015.

[4] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. *CoRR*, abs/1708.01641, 2017.

[5] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: toward spatio-temporal reasoning in visual question answering. *CoRR*, abs/1704.04497, 2017.

[6] Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. LifeQA: A real-life dataset for video question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4352–4358, Marseille, France, May 2020. European Language Resources Association.

[7] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *CoRR*, abs/2104.06039, 2021.

[8] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. *CoRR*, abs/1906.02467, 2019.

[9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[10] Yueting Zhuang, Dejing Xu, Xin Yan, Wenzhuo Cheng, Zhou Zhao, Shiliang Pu, and Jun Xiao. Multichannel attention refinement for video question answering. *ACM Trans. Multimedia Comput. Commun. Appl.*, 16(1s), mar 2020.

[11] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. *CoRR*, abs/2012.00451, 2020.

[12] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *CoRR*, abs/1906.03327, 2019.

[13] 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. *CoRR*, abs/1905.13540, 2019.

[14] Aisha Urooj Khan, Amir Mazaheri, Niels da Vitoria Lobo, and Mubarak Shah. MMFT-BERT: multimodal fusion transformer with BERT encodings for visual question answering. *CoRR*, abs/2010.14095, 2020.

[15] Shijie Geng, Ji Zhang, Zuohui Fu, Peng Gao, Hang Zhang, and Gerard de Melo. Character matters: Video story understanding with character-aware relations. *CoRR*, abs/2005.08646, 2020.

[16] Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. On modality bias in the TVQA dataset. *CoRR*, abs/2012.10210, 2020.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[18] Deniz Engin, Yannis Avrithis, Ngoc Q. K. Duong, and François Schnitzler. On the hidden treasure of dialog in video question answering. *CoRR*, abs/2103.14517, 2021.

[19] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. Knowit VQA: answering knowledge-based questions about videos. *CoRR*, abs/1910.10706, 2019.

[20] Huda AlAmri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Peter Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. Audio-visual scene-aware dialog. *CoRR*, abs/1901.09107, 2019.

[21] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang Dong Yoo. Progressive attention memory network for movie story question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8329–8338, 2019.

[22] Junyeong Kim, Minuk Ma, Trung X. Pham, Kyungsu Kim, and Chang D. Yoo. Modality shifting attention network for multi-modal video question answering. *CoRR*, abs/2007.02036, 2020.

[23] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. Bert representations for video question answering. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1545–1554, 2020.

[24] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: hierarchical encoder for video+language omni-representation pretraining. *CoRR*, abs/2005.00200, 2020.

[25] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. *CoRR*, abs/2011.07735, 2020.

[26] Zhicheng Guo, Jiaxuan Zhao, Licheng Jiao, Xu Liu, and Fang Liu. A universal quaternion hypergraph network for multimodal video question answering. *IEEE Transactions on Multimedia*, pages 1–1, 2021.

[27] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836*, 2017.

[28] Fangtao Li, Ting Bai, Chenyu Cao, Zihe Liu, Chenghao Yan, and Bin Wu. Relation-aware hierarchical attention framework for video question answering. *CoRR*, abs/2105.06160, 2021.

[29] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *CoRR*, abs/1606.00061, 2016.

[30] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. *CoRR*, abs/1803.10906, 2018.

[31] Peng Gao, Hongsheng Li, Haoxuan You, Zhengkai Jiang, Pan Lu, Steven C. H. Hoi, and Xiaogang Wang. Dynamic fusion with intra- and inter- modality attention flow for visual question answering. *CoRR*, abs/1812.05252, 2018.

[32] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. *CoRR*, abs/1611.00471, 2016.

[33] Jasdeep Singh, Vincent Ying, and Alex Nutkiewicz. Attention on attention: Architectures for visual question answering (VQA). *CoRR*, abs/1803.07724, 2018.

[34] Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. Object-centric representation learning for video question answering. *CoRR*, abs/2104.05166, 2021.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[36] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *CoRR*, abs/1812.03982, 2018.

[37] Mingrui Lao, Yanming Guo, Yu Liu, and Michael S. Lew. A language prior based focal loss for visual question answering. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021.

[38] Noa Garcia and Yuta Nakashima. Knowledge-based video question answering with unsupervised scene descriptions. *CoRR*, abs/2007.08751, 2020.