Multiple Choice Video Question Answering: A Comparative Study of Model Training Approaches

Aarti Balana
WorldQuant Brain
balana.1@iitj.ac.in

Abstract

Recent research has introduced attention based models for Visual Question Answering (VQA), which generate spatial maps highlighting relevant visual regions to answer some question/s. The VQA task combines the complexities of processing visual and linguistic data to provide commonsense answers to questions. While extensive work has focused on image captioning and image question answering, the temporal nature of videos needs special algorithms which can effectively capture the temporal aspects in the videos. Understanding video question answering holds significance for various human activities but also poses significant challenges in the field of artificial intelligence. This paper is a part of Perception Test Challenge 2023 organized by DeepMind team. In this paper, we approach this problem as a classification task and conduct a comparative study of various models. Notably, our LSTM-based model achieves a top-1 accuracy of 34%, matching the performance of the baseline flamingo model. This research contributes to the ongoing exploration of effective methods for visual question answering in the context of videos. The code is available at : ***https://github.com/aarti-b/mc-vQA_deepmind_perception_challenge_ICCV_2023***

1. Introduction

Visual Question Answering (VQA) refers to a challenging task which lies at the intersection of image understanding and language processing. The VQA task has witnessed a significant progress the recent years by the machine intelligence community. The aim of VQA is to develop a system to answer specific questions about an input image. The types of tasks in VQA varies across different tasks. Agarwal *et al*. [2] presented a novel way of combining computer vision and natural language processing concepts of to achieve Visual Grounded Dialogue, a system mimicking the human understanding of the environment with the use of visual observation and language understanding. Li *et al*. [9] have employed the attention based mechanism through transfer learning alongwith a cross-modal gating approach to improve the VQA performance. Dealing with video poses temporal, large computational resources, and others challenges. In the task of multiple-choice video question-answering (mc-vQA), presented in the Perception Test Challenge 2023 by DeepMind, the model is presented with a video, a question, and three possible answers, of which only one is correct. The model's objective is to select the correct answer. These questions span four skill areas: Memory, Abstraction, Physics, and Semantics, and they demand various forms of reasoning, including Descriptive, Explanatory, Predictive, and Counterfactual. The questions encompass video, audio, and text components. The more information about mc-vQA task and several other tasks is provided in Perception Test: A Diagnostic Benchmark for Multimodal Video Models by Viorica Pătrăucean *et al*., [10]. The figure 1 provides examples of multiple-choice questions paired with selected video frames from the dataset, with the correct answers highlighted in green.

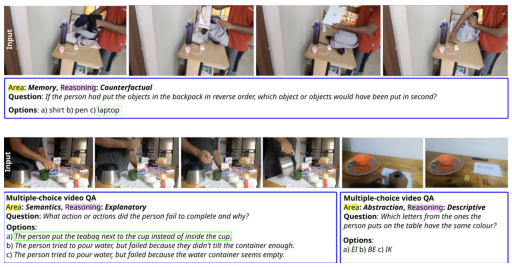

Figure 1. This image shows examples of multiple-choice questions paired with selected video frames from the dataset, with the correct answers highlighted in green.

2. Related Work

There have been works in the field of deep learning for Visual Questions Answering particularly image-text as a data input. Working on videos is still recent and requires a great deal of work for reliable VideoQA system,

especially multiple-choice VideoQA. The latest study on Video Question Answering by Shoubin Yu *et al*. introduced a novel framework called SeViLA, Self-Chained Image-Language Model for Video Localization and Question Answering [14]. It addresses the challenge of finding relevant video moments for answering questions without the need for expensive annotations. SeViLA leverages a single image-language model (BLIP2 [8]) and consists of two modules: a Localizer and an Answerer. The Localizer identifies language-aware keyframes in the video, and the Answerer uses these keyframes to predict answers. Another recent work by Shen Yan *et al*. VideoCoCa: Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners, [13], introduced an efficient approach for building a fundamental video-text model called VideoCoCa. This method maximizes the reuse of a pretrained image-text contrastive captioner (CoCa) model, requiring minimal additional training for video-text tasks. Another imageQA/mc-videoQA model by DeepMind Jean-Baptiste Alayrac *et al*. presented flamingo [1], a family of Visual Language Models (VLM) designed to swiftly adapt to new tasks with minimal annotated examples. Flamingo models can be trained on large-scale multimodal datasets containing interleaved text and images, enabling in-context few-shot learning capabilities. Their vision encoder is based on a pretrained NormalizerFree ResNet (NFNet) [5], and for video inputs, frames are processed independently and combined with temporal embeddings.

3. Models for the Comparative Study

Before we list down the models used for study let us have some idea about gated attention mechanism [12].

3.1. Gated Attention

Gated attention is a mechanism commonly used in neural networks to enhance the learning of important information while suppressing less relevant information. It can be applied in various contexts, including natural language processing and computer vision. Given a set of input vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and a set of corresponding attention weights $\alpha_1, \alpha_2, \ldots, \alpha_n$, the gated attention output $\mathbf{y}$ can be computed as follows:

$$\mathbf{y} = \sum_{i=1}^{n} \alpha_i \cdot \mathbf{x}_i \tag{1}$$

The attention weights α_i are often computed using a softmax function over a set of learnable parameters $\mathbf{w}$ and $\mathbf{b}$, and some context vector $\mathbf{c}$:

$$\alpha_i = \frac{e^{\mathbf{w}^T \cdot \tanh(\mathbf{W}\mathbf{x}_i + \mathbf{c}) + \mathbf{b}}}{\sum_{j=1}^{n} e^{\mathbf{w}^T \cdot \tanh(\mathbf{W}\mathbf{x}_j + \mathbf{c}) + \mathbf{b}}} \tag{2}$$

The attention mechanism computes these weights for each input vector and then linearly combines the input vectors according to these weights to produce the final output $\mathbf{y}$.

3.2. Models

In this study, we utilized both visual and text based models. Specifically, we applied a gated attention mechanism to the output embeddings of both video and text. Additionally, we trained these models in their vanilla form, meaning without the gated attention mechanism. We have experimented with PerceiverModel (PerceiverForImageClassificationLearned, HuggingFace) [7] + BertModel [6] , VivitModel [3] + BertModel, TimeSformerModel [4] + BertModel, and ResNext_LSTM [15,16] + miniLM. Before proceeding to the implementation section let us discuss about the models chosen for training -

- Perceiver model by DeepMind, [7], is flexible across modalities and has strong few shot learning capabilities.
- ViViT, A Video Vision Transformer, [3] model leverage the Transformer architecture which allows the model to capture long-range dependencies in temporal data, making it well-suited for video dataset.
- TimeSformer, [4] provides an efficient video classification framework.
- ResNext_LSTM ensemble model is used for simplicity where the convolutional network extracts the embeddings of video frames with MAX_FRAMES=100. These embeddings are then passed into the LSTM layer for temporal learning.

4. Implementation

4.1. DataSet

There are 2100 videos in the training set and 3146 videos in the test set. All videos have annotations and the data is in COCO style.

4.2. Training Strategies and Experiments

In this study, the training for the multiple-choice video question-answering (mc-vQA) task involves the use of two distinct strategies:

- Fine tuning the model using few-shot learning.
- Training the model with first 100 frames of whole video dataset.

As mentioned in section 3.2 about different models, our approach leverages the PerceiverModel [7] for processing

video data, while employing the Bert model for textual data. In the case of answer options, their embeddings are aggregated through averaging. Similar averaging is done for video embeddings. Subsequently, gated attention mechanism is applied to the resulting visual and text embeddings, followed by a series of layers, including a fully connected layer, a batch normalization layer, and a linear layer, for the purpose of classification. For comparison to see if gated attention is doing any improvement we have trained PerceiverModel + BertModel without using gated attention mechanism and is named as Vanilla training.

Other models, ViViTModel [3]+ BertModel, TimesformerModel + BertModel, are used in a similar fashion with few-shot learning and their respective Vanilla model training. Whereas ResNext [16] model is used for extracting the features/embeddings of first 100 frames of the videos and then these features are passed into the LSTM model with a chunk size of 50 followed by averaging it out and feeding the resultant embedding into the fully connected layer.

It shall be noted that some experimental results have not been included in the result table due to the needing of extra resources (not every trained model is used on test set). A100 GPU, Google Colab and paperspace, was used for training of all of the methods.

5. Results

For evaluation, top-1 accuracy metric is used, [11]. Table 1 and Table 2 shows the result of the models while testing and training. Table 1 shows the evaluation metrics calculated by the platform EvalAI, where the challenge is held, after submission. Table 1 show that after using whole dataset, up to 100 frames, we got 1% increase in top-1 accuracy with respect to Gated TimeSFormer+Bert. This is worth experimenting with full frame videos in future work.

6. Conclusion and Future Work

As mentioned earlier, it is worth experimenting with the whole dataset of full length videos for training ResNext_LSTM and other models. Another direction that can be used is using audio dataset in the training for mcvQA, but this will require huge computational resources.

7. Acknowledgement

I would like to extend my heartfelt gratitude to the organizers of this challenge. This challenge has not only provided us with an opportunity to showcase our skills and knowledge but has also raised a sense of community and shared purpose among participants.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 2, 3

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 2

[5] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning*, pages 1059–1071. PMLR, 2021. 2

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[7] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[9] Wei Li, Jianhui Sun, Ge Liu, Linglan Zhao, and Xiangzhong Fang. Visual question answering with attention transfer and a cross-modal gating mechanism. *Pattern Recognition Letters*, 133:334–340, 2020. 1

[10] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, et al. Perception test: A diagnostic benchmark for multimodal video models. *arXiv preprint arXiv:2305.13786*, 2023. 1

[11] sklearn. to-1. 3

[12] Lanqing Xue, Xiaopeng Li, and Nevin L Zhang. Not all attention is needed: Gated attention network for sequence data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6550–6557, 2020. 2

[13] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022. 2

top-1 accuracy of models trained on mc-vQA	
Testing Phase	
Models	top-1
ResNext_LSTM+miniLM	**34.03**%
Gated TimeSFormer+Bert	33.07%
Gated ViViT+Bert	34.02%

Table 1. top-1 accuracy values of different visual+text models for mc-vQA task trained with the Standard Cross Entropy loss (testing phase)

top-1 accuracy of models trained on mc-vQA	
Training Phase	
Models	top-1
Gated Perceiver+Bert	**35.61** %
Vanilla Perceiver+Bert	33.73%
Gated TimeSFormer+Bert	33.91%
Vanilla TimeSFormer+Bert	**35.32**%
Gated ViViT+Bert	33.64%
Vanilla ViViT+Bert	32.32%
ResNext_LSTM+miniLM	32.12%

Table 2. top-1 accuracy values of different visual+text models for mc-vQA task trained with the Standard Cross Entropy loss (training phase)

[14] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023. 2

[15] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019. 2

[16] Tianyan Zhou, Yong Zhao, and Jian Wu. Resnext and res2net structures for speaker verification. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 301–307. IEEE, 2021. 2, 3