Progressive Attention Memory Network for Movie Story Question Answering

Junyeong Kim[1*] Minuk Ma[1] Kyungsu Kim[2] Sungjin Kim[2] Chang D. Yoo[1]
[1] Korea Advanced Institute of Science and Technology (KAIST)
[2] Samsung Research
[1]{junyeong.kim, akalsdnr, cd_yoo}@kaist.ac.kr [2]{ks0326.kim, sj9373.kim}@samsung.com

Abstract

This paper proposes the progressive attention memory network (PAMN) for movie story question answering (QA). Movie story QA is challenging compared to VQA in two aspects: (1) pinpointing the temporal parts relevant to answer the question is difficult as the movies are typically longer than an hour, (2) it has both video and subtitle where different questions require different modality to infer the answer. To overcome these challenges, PAMN involves three main features: (1) progressive attention mechanism that utilizes cues from both question and answer to progressively prune out irrelevant temporal parts in memory, (2) dynamic modality fusion that adaptively determines the contribution of each modality for answering the current question, and (3) belief correction answering scheme that successively corrects the prediction score on each candidate answer. Experiments on publicly available benchmark datasets, MovieQA and TVQA, demonstrate that each feature contributes to our movie story QA architecture, PAMN, and improves performance to achieve the state-of-the-art result. Qualitative analysis by visualizing the inference mechanism of PAMN is also provided.

1. Introduction

Humans have an innate cognitive ability to infer from different sensory inputs to answer questions of 5W's and 1H involving *who, what, when, where, why* and *how*, and it has been a quest of mankind to duplicate this ability on machines. In recent years, studies on question answering (QA) have successfully benefited from deep neural networks, and showed remarkable performance improvement on textQA [24, 30], imageQA [2, 3, 19, 31], videoQA [8, 11, 32, 34]. This paper considers movie story QA [15, 18, 21, 26, 29] that aims at a joint understanding of vision and language by answering questions about movie contents and storyline after observing temporally-aligned video and subtitle. Movie story QA is challenging compared to VQA in following two aspects: (1) pinpointing the temporal parts relevant to answer the question is difficult as the movies are typically longer than an hour and (2) it has both video and subtitle where different questions require different modality to infer the answer.

The first challenge of movie story QA is that it involves long videos that are possibly longer than an hour which hinders pinpointing the required temporal parts. The information in the movie required to answer the question is not distributed uniformly across the temporal axis. To address this issue, memory networks [24] have widely been accepted in QA tasks [21, 24, 26, 30]. The attention mechanism have widely been adopted to retrieve the information relevant to the question. We observed that single-step attention on memory networks [21, 26] often generates blurred temporal attention map.

The second challenge of movie story QA is that it involves both video and subtitle where different questions require different modality to infer the answer. Each modality may convey essential information for different questions, and optimally fusing them is an important problem. For example in the movie *Indiana Jones and the Last Crusade*, answering the question *"What does Indy do to the grave robbers at the beginning of the movie?"* would require video modality rather than subtitle modality while the question *"How has the guard managed to stay alive for 700 years?"* would require subtitle modality. Existing multi-modal fusion methods [7, 14, 15] only focus on modeling rich interactions between the modalities. However, these methods are question-agnostic in that the fusion process is not conditioned on the question.

To address the aforementioned challenges, this paper proposes Progressive Attention Memory Network (PAMN) for movie story QA. PAMN contains three main features; (1) progressive attention mechanism for pinpointing required temporal parts, (2) dynamic modality fusion for adaptively fusing modalities conditioned on question and (3) belief correction answering scheme. Progressive attention mechanism utilizes cues from both question and an-

*This research was supported by Samsung Research

swers to prune out irrelevant temporal parts for each memory. While iteratively taking question and answers for temporal attention generation, memories are progressively updated to accumulate cues to locate relevant temporal parts for answering the question. Compared to stacked attention [6, 31], progressive attention considers multiple sources (e.g., Q and A) and multiple targets (e.g., video and subtitle memory) in a single framework. Dynamic modality fusion aggregates the outputs from each memory by adaptively determining the contribution of each modality. Conditioned on the current question, the contribution is obtained by soft-attention mechanism. Fusing multi-modal data by bilinear operations [4, 7, 14] often requires heavy computation or large number of parameters. Dynamic modality fusion efficiently integrate video and subtitle modality by discarding worthless information from unnecessary modality. Belief correction answering scheme successively corrects the prediction score of each candidate answer. When humans solve questions, they typically read content, question and answers multiple times in an iterative manner [10]. This observation is modeled by belief correction answering scheme. The prediction score (logits), which this paper refers to a belief, is equally likely initialized and successively corrected compared to existing answering scheme [15, 21, 29] which uses single-step answering scheme.

The main contribution of this paper is summarized as follows. (1) This paper proposes a movie story QA architecture referred to as PAMN that tackles major challenges of movie story QA with three features; progressive attention, dynamic modality fusion and belief correction answering scheme. (2) PAMN achieves the state-of-the-art results on MovieQA dataset. Both the quantitative and qualitative results exhibit the benefits and potential of PAMN.

2. Related Work

2.1. Visual Question Answering

Despite the short history, imageQA enjoys large number of datasets including VQA [3], COCO-QA [23] and Visual7W [35]. Attention mechanism is widely used to locate the visual clues relevant to the question. Stacked Attention Network (SAN) [31] utilizes stacked attention module to query an image multiple times to infer the answer progressively. The Dual Attention Network (DAN) [22] jointly leverages visual and textual attention mechanisms to localize key information from both image and question. Recently, applying bilinear operation showed promising results on imageQA. Multimodal Compact Bilinear pooling (MCB) [7] utilized bilinear operation to fuse image and question features in imageQA. To reduce the computational complexity, MCB uses the sampling-based approximation. To further reduce the feature dimension, Multimodal Low-rank Bilinear Attention Network (MLB) [14] utilizes Hadamard product in the common space with two low-rank projection matrices. Multimodal Tucker Fusion [4] utilizes tucker decomposition [27] to efficiently parameterize bilinear interactions between visual and textual representation.

VideoQA is a natural extension of imageQA as video can be seen as temporal extension of image. Large-scale videoQA benchmarks such as TGIF-QA [11] and 'fill-in-the-blank' [34] have boosted the research on videoQA. Spatio-temporal VQA (ST-VQA) [11] generates spatial and temporal attention to localize which regions in a frame and which frames in a video to attend, respectively. Yu *et al.* [32] proposed Joint Sequence Fusion (JSFusion) that measures semantic similarity between video and language. JSFusion utilizes hierarchical attention mechanism that learns matching representation patterns between modalities.

2.2. Movie Story Question Answering

A recent direction in videoQA leverages text modality such as subtitle in addition to video modality for story understanding. To this end, various video story QA benchmarks such as PororoQA [16], MeMexQA [12], TVQA [17] and MovieQA [26] have been suggested. Numerous researches have tackled MovieQA benchmark which provides movie clip, subtitle and other various textual descriptions. Tapaswi *et al.* [26] divided the movie into multiple sub-shots and utilized memory network (MemN2N) [24] to store video and subtitle features into memory slots. Deep Embedded Memory Network (DEMN) [16] reconstructs stories from a joint stream of scene-dialogue using a latent embedding space and retrieves information which is relevant to the question. Na *et al.* [21] proposed Read-Write Memory Network (RWMN) which is a CNN-based memory network where video and subtitle features are first fused using bilinear operation, then write/read networks store/retrieve information, respectively.

Liang *et al.* [18] proposed Focal Visual-Text Attention (FVTA) that utilizes the hierarchical attention applied to a three-dimensional tensor to localize evidential image and text snippets. Layered Memory Network (LMN) [29] utilizes Static Word Memory module and Dynamic Subtitle Memory module to learn frame-level and clip-level representations. The hierarchically formed movie representation encodes the correspondence between words and frames, and the temporal alignment between sentences and frames. Multimodal Dual Attention Memory (MDAM) [15] utilizes multi-head attention mechanism [28] and question attention to learn the latent concepts of multimodal contents. Multimodal fusion is performed once after the attention process. Compared to existing architectures on movie story QA that adopt single-step reasoning, PAMN provides multi-step reasoning approach to localize necessary information from question, answers, and movie contents.

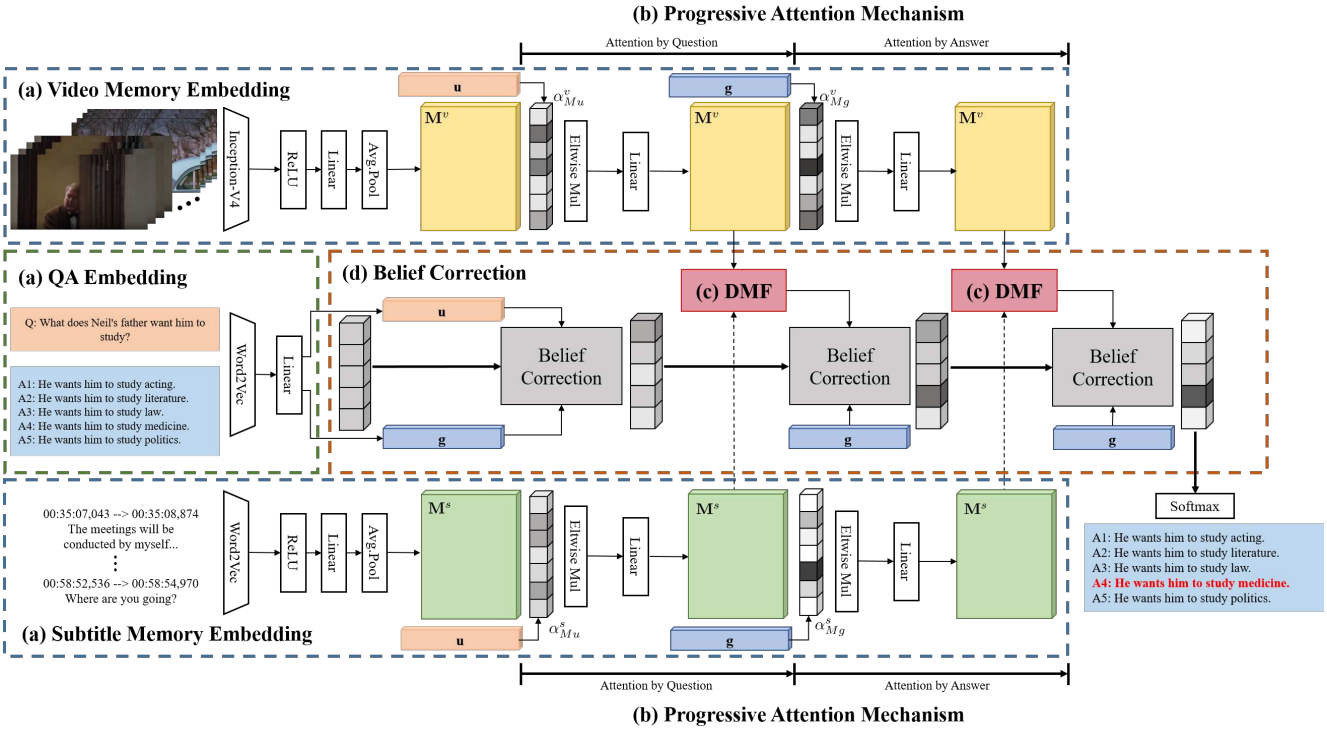

Figure 1. Illustration of the proposed PAMN. The pipeline of PAMN is as follows. (a) Question and candidate answers are embedded into a common space. Video and subtitle are embedded into dual memory that holds independent memories for each modality. (b) Progressive attention mechanism pinpoints temporal parts that are relevant to answering the question. To infer the correct answer, (c) dynamic modality fusion that adaptively integrates outputs of each memory by considering contribution of each modality. (d) Belief correction answering scheme successively corrects the belief of each answer from equally likely initialized belief.

3. Progressive Attention Memory Network

This section describes the proposed Progressive Attention Memory Network (PAMN). Fig. 1 shows the overall architecture of PAMN, which fully utilizes diverse sources of information (video, subtitle, question and candidate answers) to answer the question. The pipeline of PAMN is as follows. First, video and subtitle are embedded into dual memory as in Fig. 1(a) that holds independent memories for each modality. Then, progressive attention mechanism pinpoints temporal parts that are relevant to answering the question as in Fig. 1(b). To infer the correct answer, dynamic modality fusion in Fig. 1(c) adaptively integrates outputs of each memory by considering contribution of each modality. Belief correction answering scheme successively corrects the belief of each answer from equally likely initialized belief as in Fig. 1(d).

3.1. Problem Setup

The formal definition of the problem is as follows. The inputs of PAMN are (1) a question representation $\mathbf{q} \in \mathbb{R}^{300}$, (2) five candidate answer representations $\{\mathbf{a}_i\}_{i=1}^{5} \in \mathbb{R}^{5\times 300}$, (3) temporally aligned video ($\mathbf{v}$) and subtitle ($\mathbf{s}$) representation $\{(\mathbf{v}_i, \mathbf{s}_i)\}_{i=1}^{T}$ on the whole movie. Each element of subtitle representation $\mathbf{s}_i$ corresponds to the dialog sentence of a character and each element of video representation $\mathbf{v}_i$ is extracted from temporally aligned video clip. The number of overall sentences of the movie is denoted as T. The detailed explanation on extracting visual and textual feature is provided in Section 4.2. The objective is to maximize the following likelihood:

$$\arg\max_{\theta} \sum_{\mathcal{D}} \log P(\mathbf{y}|\mathbf{v}, \mathbf{s}, \mathbf{q}, \mathbf{a}; \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ denotes learnable model parameters, $\mathcal{D}$ represents dataset and $\mathbf{y}$ represents the correct answer.

3.2. Dual Memory Embedding

As depicted in Fig. 1(a), the inputs are first mapped to an embedding space. The question representation $\mathbf{q}$ and candidate answer representations $\{\mathbf{a}_i\}_{i=1}^{5}$ are embedded to a common space by weight-shared linear fully connected (FC) layer with parameters $\mathbf{W}_{ug} \in \mathbb{R}^{300\times d}$ and $\mathbf{b}_{ug} \in \mathbb{R}^{d}$, to yield question embedding $\mathbf{u} \in \mathbb{R}^{d}$ and answer embedding $\mathbf{g} \in \mathbb{R}^{5\times d}$ where d denotes the memory dimension.

Video representation $\mathbf{v}$ and subtitle representation $\mathbf{s}$ are embedded independently to generate video memory $\mathbf{M}^v$ and subtitle memory $\mathbf{M}^s$. This dual memory structure enables pinpointing different temporal parts for each modality. To reflect the observation that the adjacent video clips often have strong correlations, we utilized the average pooling (Avg.Pool) layer to store the adjacent representations into a single memory slot.

As the first step of dual memory embedding, feed-forward neural network (FFN) composed of two linear FC layers with ReLU non-linearity in between is applied to embed video and subtitle representation. This operates on every element of $\mathbf{v}$ and $\mathbf{s}$ independently. Then, average pooling layer is applied to model neighboring representations together, forming video memory $\mathbf{M}^v$ and subtitle memory $\mathbf{M}^s$, i.e. *dual memory*:

$$\begin{aligned} \text{FFN}(\mathbf{x}) &= \text{ReLU}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, & (2)\\ \mathbf{M}^v &= \text{Avg.Pool}(\text{FFN}(\mathbf{v}); \theta_p, \theta_s), & (3)\\ \mathbf{M}^s &= \text{Avg.Pool}(\text{FFN}(\mathbf{s}); \theta_p, \theta_s), & (4) \end{aligned}$$

where θ_p and θ_s denotes the size and stride of the pooling, $\mathbf{x}$ indicates the each input and $\mathbf{W}$, $\mathbf{b}$ denotes the weight and bias of feed-forward neural network. Finally, generated video and subtitle memory are $\mathbf{M}^v, \mathbf{M}^s \in \mathbb{R}^{N\times d}$ where $N = \lceil T/\theta_s \rceil$.

3.3. Progressive Attention Mechanism

The progressive attention mechanism in Fig. 1(b) takes dual memory $\mathbf{M}^v, \mathbf{M}^s$, question embedding $\mathbf{u}$ and answer embedding $\mathbf{g}$ as inputs, and progressively attends and updates the dual memory. While iteratively taking question and answers for temporal attention generation, memories are progressively updated to accumulate cues to locate relevant temporal parts for answering the question. We observed that single-step temporal attention on memory networks [26, 21] often generates blurry attention map. The multi-step nature of progressive attention mechanism enables generating sharper attention distribution. Unnecessary information from memory is filtered out at each iteration.

The first step of progressive attention mechanism is temporal attention by question embedding $\mathbf{u}$. The attention weights are obtained by calculating the cosine similarity between each memory slot and question embedding $\mathbf{u}$ as in Eqs.5, 6. The dual memory is multiplied by the attention weights and is followed by linear FC layer to be updated as in Eqs.7, 8. The attention operates independently for the video memory $\mathbf{M}^v$ and the subtitle memory $\mathbf{M}^s$:

$$\begin{aligned} \alpha^v_{Mu} &= \text{softmax}(\mathbf{u}\mathbf{M}^{v\intercal}), & (5)\\ \alpha^s_{Mu} &= \text{softmax}(\mathbf{u}\mathbf{M}^{s\intercal}), & (6)\\ \mathbf{M}^v &\leftarrow (\alpha^v_{Mu} \odot \mathbf{M}^v)\mathbf{W}^v_{Mu} + \mathbf{b}^v_{Mu}, & (7)\\ \mathbf{M}^s &\leftarrow (\alpha^s_{Mu} \odot \mathbf{M}^s)\mathbf{W}^s_{Mu} + \mathbf{b}^s_{Mu}, & (8) \end{aligned}$$

where $\alpha^v_{Mu}, \alpha^s_{Mu} \in \mathbb{R}^N$ denote the temporal attention weight for $\mathbf{M}^v, \mathbf{M}^s$, respectively. The learnable parameters for linear FC layer is denoted by $\mathbf{W}_{Mu}, \mathbf{b}_{Mu}$, $\leftarrow$ indicates the update operation and $\odot$ represents the element-wise multiplication with broadcasting on appropriate axis.

The second step of progressive attention mechanism is temporal attention by answers. This step is similar to the first step except it utilizes answer embedding $\mathbf{g}$ to attend updated dual memory $\mathbf{M}^v, \mathbf{M}^s$:

$$\begin{aligned} \alpha^v_{Mg} &= \text{softmax}(\mathbf{g}\mathbf{M}^{v\intercal}), & (9)\\ \alpha^s_{Mg} &= \text{softmax}(\mathbf{g}\mathbf{M}^{s\intercal}), & (10)\\ \mathbf{M}^v &\leftarrow (\alpha^v_{Mg} \odot \mathbf{M}^v)\mathbf{W}^v_{Mg} + \mathbf{b}^v_{Mg}, & (11)\\ \mathbf{M}^s &\leftarrow (\alpha^s_{Mg} \odot \mathbf{M}^s)\mathbf{W}^s_{Mg} + \mathbf{b}^s_{Mg}, & (12) \end{aligned}$$

where α^v_{Mg} and $\alpha^s_{Mg} \in \mathbb{R}^{5\times N}$ denote the temporal attention weights for dual memory and $\mathbf{M}^v, \mathbf{M}^s \in \mathbb{R}^{5\times N\times d}$ represent the updated video and subtitle memory, respectively.

Multiple Hops Extension. As described above, the progressive attention mechanism attends the dual memory only once for each attention step. In this case, the dual memory may contain much irrelevant information and lack capability to query complicated semantics to answer the question. Progressive attention can be naturally extended to utilize multiple hops [24] for fine-grained extraction of abstract concepts and reasoning of high-level semantics.

Different from the memory network [24] that utilizes the sum of the output o_k and query u_k of k-th hop as the query of next hop, we use the same question embedding $\mathbf{u}$ with updated dual memory $\mathbf{M}^{(k)}$ for k-th hop. Each attention step in Eqs. 5-8, 9-12 is repeated h_{Mu}, h_{Mg} times, respectively. Each attend and update operations can be expressed as:

$$\begin{aligned} \alpha^{(k)} &= \text{softmax}(\mathbf{x}\mathbf{M}^{(k-1)\intercal}), & (13)\\ \mathbf{M}^{(k)} &\leftarrow (\alpha^{(k)} \odot \mathbf{M}^{(k-1)})\mathbf{W}^{(k)} + \mathbf{b}^{(k)}, & (14) \end{aligned}$$

where the subscripts and superscripts corresponding to each equation are omitted to avoid repetition, and $\mathbf{x}$ indicates $\mathbf{u}$ or $\mathbf{g}$ for each step of progressive attention.

3.4. Dynamic Modality Fusion

Dynamic modality fusion in Fig. 1(c) aggregates dual memory into fused output $\mathbf{o}$ at the end of each progressive attention step. Different question requires different modality to infer the answer. Consider the question *"What drink bottle is at the table when Robin, Lily, Marshall and Ted are talking to each other?"*. In this case, the video modality would be more important than subtitle modality. Similar to modality attention [9, 13], dynamic modality fusion is soft-attention based algorithm that determines the contribution of each modality for answering the question.

Given dual memory $\mathbf{M}^v, \mathbf{M}^s$, dynamic modality fusion first sum each memory along temporal axis and compute

cosine similarity with question embedding $\mathbf{u}$ to calculate attention score.:

$$\mathbf{o}^m = \sum_{n=1}^{N} \mathbf{M}^m, \quad (15)$$

$$\alpha_{\text{DMF}} = \text{softmax}(\mathbf{u}[\mathbf{o}^v; \mathbf{o}^s]^\intercal), \quad (16)$$

where m indicates each modality $\mathbf{v}$ or $\mathbf{s}$, $\mathbf{o}^m$ represents the output of each memory, N denotes the temporal length of dual memory, and α_{DMF} denotes attention weights. Finally, the fused output $\mathbf{o}$ is computed by weighted summing between attention weight and memory output:

$$\mathbf{o} = \sum_m \alpha^m_{\text{DMF}} \mathbf{o}^m. \quad (17)$$

The learned attention weight can be interpreted as contribution or importance of each modality on answering the question. By regulating the ratio of each modality on fused output, dynamic modality fusion leads to stable learning by discarding information from unnecessary modality.

3.5. Belief Correction Answering Scheme

Belief correction answering scheme in Fig. 1(d) selects the correct answer among five candidate answers. Rather than determining the prediction score once, belief correction answering scheme successively corrects prediction score by observing diverse source of information. This mimics the multi-step reasoning process of human answering difficult questions [10]. Combined with progressive attention and dynamic modality fusion, this multi-step reasoning approach of PAMN strengthens the model's ability to extract high-level meaning from the multimodal data.

Belief $\mathbf{B} \in \mathbb{R}^5$ denotes the prediction score on the candidate answers. The prediction probability $\mathbf{z} \in \mathbb{R}^5$ is computed by normalizing the belief, and the answer y is predicted with the highest probability:

$$\mathbf{z} = \text{softmax}(\mathbf{B}), \quad (18)$$

$$y = \arg\max_{i \in [5]}(\mathbf{z}_i). \quad (19)$$

One way of initializing belief would be *null initialization* that endows all candidate answers with equal probabilities before observing any information. To reflect this unbiased initialization, the belief $\mathbf{B}$ is initialized as zero vector.

Belief correction answering scheme adopts three-step belief correction; *u*-, *Mu*- and *Mg*-correction. For each correction step, the belief is corrected by accumulating the similarity between answer embedding $\mathbf{g}$ and the observed information. Belief is first corrected by only considering the question, i.e. *u*-correction. The intuition is that human often builds prior biases after skimming through only the question and candidate answers:

$$\mathbf{B}_u = \mathbf{u}\mathbf{g}^\intercal, \quad (20)$$

$$\mathbf{B} \leftarrow \mathbf{B} + \mathbf{B}_u. \quad (21)$$

Then for *Mu*- and *Mg*-correction, the outputs of first and second progressive attention steps, $\mathbf{o}_{Mu}$ and $\mathbf{o}_{Mg}$, are considered. Again, the similarities between answer embedding $\mathbf{g}$ are computed:

$$\mathbf{B}_{Mu} = \mathbf{o}_{Mu}\mathbf{g}^\intercal, \quad (22)$$

$$\mathbf{B}_{Mg,i} = \mathbf{o}_{Mg,i}\mathbf{g}_i^\intercal. \quad (23)$$

Finally, the belief is corrected to infer correct answer:

$$\mathbf{B} \leftarrow \mathbf{B} + \beta_{Mu}\mathbf{B}_{Mu}, \quad (24)$$

$$\mathbf{B} \leftarrow \mathbf{B} + \beta_{Mg}\mathbf{B}_{Mg}, \quad (25)$$

where the correction weights β_{Mu}, β_{Mg} are hyper parameters that scales corresponding belief correction. Note that the belief is normalized to have unit norm after each correction.

4. Experiments

4.1. Dataset

MovieQA [26] benchmark is constructed for movie story QA which consists various sources of information such as movie clip, subtitle, plot synopses, scripts and DVS transcriptions. MovieQA dataset contains 408 movies with corresponding 14,944 multiple-choice questions. MovieQA benchmark consists of 6 tasks according to which sources to be used. This paper focuses on video+subtitles task which is the only task utilizing movie clip. Since only 140 movies contain video clips, there are 6,462 question-answer pairs which splits into 4,318 training, 886 validation and 1,258 test samples.

TVQA [17] benchmark is video story QA dataset on TV show domain. It consists of total 152.5k question-answer pairs on six TV shows: *The Big Bang Theory, How I Met Your Mother, Friends, Grey's Anatomy, House, Castle*. Each split of TVQA contains 122k, 15.25k, 15.25k for train, validation and test, respectively. Unlike MovieQA which considers whole movie as input, TVQA contains 21,793 short clips of 60/90 seconds segmented from the original TV show for question-answering.

4.2. Feature extraction

For fair comparison, we extracted visual and textual features similar to previous works [21, 26] and fixed them during training.

Textual feature Each sentence from question, candidate answers and subtitle are divided into sequence of words, then each word is embedded by skip-gram model [20] provided by Tapaswi *et al.* [26] which is trained on MovieQA plot synopses. In order to encode the order of words within a sentence, position encoding (PE) [24] is utilized to obtain textual feature. For example in the case of question,

Methods	valid Acc.	test Acc.
SSCB w/o Sub	21.60	-
SSCB w/o Vid	22.30	-
SSCB [26]	21.90	-
MemN2N w/o Sub	23.10	-
MemN2N w/o Vid	38.00	-
MemN2N [26]	34.20	-
DEMN [16]	44.70	29.97
RWMN [21]	38.67	36.25
FVTA [18]	41.00	37.30
LMN [29]	42.50	39.03
MDAM [15]	-	41.41
PAMN w/o Sub	42.33	-
PAMN w/o Vid	42.56	-
PAMN	43.34	**42.53**

Table 1. Accuracy comparison on the validation and test set of MovieQA benchmark of Video+Subtitles task. PAMN achieves the state-of-the-art performance. The test set accuracy is obtained from online evaluation server. And '-' indicates that the performance is not provided.

Methods	Video Feat.	test Acc.
Longest Answer	-	30.41
TVQA [17]	img	63.57
TVQA [17]	reg	63.19
TVQA [17]	cpt	65.46
PAMN	img	64.61
PAMN	cpt	**66.77**

Table 2. Accuracy comparison on the test set of TVQA benchmark without timestamp annotation. We utilized the video and text features extracted by Lei *et al.* [17].

$\mathbf{q} = \sum_n \text{PE}(\mathbf{q}_n) \in \mathbb{R}^{300}$ where each $\mathbf{q}_n$ indicates word vector.

Visual feature Movies are divided into video clips that are temporally aligned with each sentence of the subtitle. The frames are sampled from each video clip with the rate of 1 fps. Then, frame feature of size 1536 is extracted from "Average Pooling" layer on Inception-v4 [25]. Finally, mean-pooling over all frame features from the corresponding video clip produces the visual feature, $\mathbf{v}_i \in \mathbb{R}^{1536}$.

4.3. Implementation details

The entire architecture was implemented using Tensorflow [1] framework. All the results reported in this paper were obtained using the Adagrad optimizer [5] with a mini-batch size of 32 and the learning rate of 0.001. All the experiments were performed under CUDA acceleration with single NVIDIA TITAN Xp (12GB of memory) GPU. In all the experiments, the recommended train / validation / test split was strictly observed.

4.4. Quantitative Results

Table 1 compares the validation and test accuracy on the MovieQA benchmark of Video+Subtitles task. We compare the performance of PAMN with other state-of-the-art architecture. The ground-truth answers for MovieQA test set are not observable and the evaluation on the test set can only be performed once every 72 hours through an online evaluation. On MovieQA benchmark, PAMN exhibits the state-of-the-art results by attaining test accuracy of 42.53%. It outperforms the runner-up, MDAM [15] (41.41%) by 1.12% and the third place, LMN [29] (39.03%) by 3.50%. Note the MDAM is an ensemble of 20 different models, while PAMN is a single model.

In order to evaluate the effectiveness of each modality, experiments based on using only video and subtitle were also conducted: PAMN w/o Sub and PAMN w/o Vid. From near random-guess performances of SSCB w/o Sub [26] and MemN2N w/o Sub [26] as shown in Table. 1, it is noticed that movie story understanding is difficult using only video. The PAMN w/o Sub attains large performance gain of 19.23% compared to MemN2N w/o Sub. It even achieves performance comparable to LMN [29] which exploits both video and subtitle. PAMN understands movie story even without observing subtitle. From Table.1, it is noticed that PAMN performs better than PAMN w/o Vid and PAMN w/o Sub which indicates both video and subtitle provides conducive information in improving prediction.

Table 2 shows performance comparison on TVQA benchmark without timestamp annotation. In this experiment, we utilized the video and text features extracted by Lei *et al.* [17] (i.e. ImageNet and visual concept feature for video and GloVe feature for text) for fair comparison. Further, we encoded the sentence feature using LSTM instead of position encoding. On TVQA benchmark, PAMN outperforms state-of-the-art result by attaining test accuracy of 66.77% with visual concept feature.

4.5. Ablation Study

Table. 3 summarizes the ablation analysis of PAMN on the validation set of MovieQA benchmark in order to measure the validity of the key components of PAMN. To measure to effectiveness of progressive attention mechanism, each temporal attention step of PAMN w/o PA utilizes dual memory obtained in Eqs. 3,4, i.e. PAMN w/o PA do not accumulate cues and each attention step operates in a parallel manner. PAMN w/o Multiple Hop attends dual memory only once for each temporal attention step. As shown in the first block of Table. 3, PAMN w/o PA underperforms PAMN, which shows that the attention accumulation by progressive attention mechanism is important in understanding movie story. Multiple hops extension is also crucial in attaining the best possible performance. For ablating dynamic modality fusion, we experiment with

Methods	valid Acc.	Δ
PAMN w/o PA	42.03	-1.31%
PAMN w/o Multiple Hop	42.67	-0.67%
PAMN w/o DMF	42.09	-1.25%
PAMN w/ MCB [7]	42.89	-0.45%
PAMN w/ MFB [33]	42.55	-0.79%
PAMN w/ Tucker [4]	42.89	-0.45%
PAMN w/o *Mu,Mg*-correction	39.50	-3.84%
PAMN w/o *Mg*-correction	41.76	-1.58%
PAMN w/o *Mu*-correction	40.86	-2.48%
PAMN	**43.34**	-

Table 3. Ablation studies of the proposed PAMN on the validation set of MovieQA benchmark. The last column shows the performance drop.

four variants: PAMN w/o DMF take the mean of the outputs of dual memory $\mathbf{o}^v$, $\mathbf{o}^s$, PAMN w/ MCB, MFB, Tucker use MCB [7], MFB [33], Tucker decomposition [4, 27] instead of dynamic modality fusion, respectively. As shown in the second block of Table. 3, fusing modalities by averaging or bilinear operations show lower performance than dynamic modality fusion. This implies that question dependent modality weighting (i.e. dynamic modality fusion) helps strengthens conducive modality. To measure the effectiveness of belief correction answering scheme, the third block of Table. 3 shows the experimental results of three variants: PAMN w/o *Mu,Mg*-correction, PAMN w/o *Mg*-correction, and PAMN w/o *Mu*-correction. It is noteworthy that only using QA pairs shows much higher performance that the random baseline of 20%. Considering Mu- and Mg-correction, PAMN w/o *Mg*-correction shows 2.26% and PAMN w/o *Mu*-correction shows 1.36% performance boosts, respectively.

Table. 4 summarizes the performance variation depending on three sets of hyperparameters; the number of hops for attention by question $\mathbf{u}$ and answer $\mathbf{g}$, θ_p, θ_s: size and stride of Avg. Pool layer, and β_{Mu}, β_{Mg}: correction weights for belief correction module. The multiple hops extension with 2-repetitions exhibits the best validation performance for PAMN. The multiple hops extension with more than three repetitions may suffer from overfitting due to the small size of dataset. The performance is positively affected by increasing θ_p and θ_s, but it degrades for large θ_p and θ_s due to information blurring of Avg. Pool. We observed that there is no best-performing optimal correction weights. If the question representation $\mathbf{u}$ has enough information about where in the movie to focus on, β_{Mu} should be higher, and vice versa. Furthermore, it is preferable to have smaller β_{Mg} than β_{Mu} since large value of β_{Mg} dilates the effect of ug and Mu correction since the normalization is applied in between every belief correction.

# hops		Avg. Pool		Correction		Acc.
h_{Mu}	h_{Mg}	θ_p	θ_s	β_{Mu}	β_{Mg}	
1	1	1	1	1	0.5	38.94
1	1	12	8	1	0.5	40.18
1	1	24	16	0.5	0.5	40.07
1	1	24	16	1	0.1	42.10
1	1	24	16	1	0.5	42.67
1	1	40	30	0.5	0.5	40.97
1	1	40	30	1	0.1	42.66
1	1	40	30	1	0.5	42.55
1	1	80	60	1	0.5	41.20
2	2	24	16	1	0.5	**43.34**
2	2	40	30	1	0.1	42.89
3	3	24	16	1	0.5	42.55
3	3	40	30	1	0.1	42.77

Table 4. Performance variation of PAMN on the validation set of MovieQA benchmark depending on three sets of hyper parameters. h_{Mu}, h_{Mg}: the number of hops for attention by question $\mathbf{u}$ and answer $\mathbf{g}$, θ_p, θ_s: size and stride of Avg. Pool layer, and β_{Mu}, β_{Mg}: correction weights for belief correction module.

4.6. Qualitative analysis

The Fig. 2 illustrates the selected qualitative examples of PAMN. Each example provides the temporal attention map $\alpha^v_{Mg}, \alpha^s_{Mg}$ from progressive attention mechanism, the ground-truth (GT) temporal part where the question was generated from, the attention weights $\alpha^v_{DMF}, \alpha^s_{DMF}$ from dynamic modality fusion, and the inference path of belief correction answering scheme. The generated temporal attention well matches with the GT which indicates that PAMN successively learns where to attend. The weights $\alpha^v_{\text{DMF}}, \alpha^s_{\text{DMF}}$ adaptively scales depending on the question type which implies that PAMN learns what modality to use without additional supervision. For some cases, PAMN predicts the correct answer at the *u*-correction step while for other cases the correct answer is determined at the last (*Mg*) step. PAMN is an interpretable architecture in that the inference path and the attention map provide the trace of where PAMN attends and what information source it used to answer the question.

The Fig. 3 exhibits the accuracy comparison with respect to the first word of the question between MemN2N [26], RWMN [21] and PAMN on the validation set of MovieQA benchmark. The results on 5W1H question types: *Who, Where, When, What, Why* and *How* are analyzed. Typically, answering *who, where, when, what* questions require pinpointing the temporal parts relevant to the question (e.g., When do the loyalists take over Air Force One?, What does Korshunov demand from Vice President Bennett?). On the other hand, answering *why, how* questions require understanding the contextual information over the

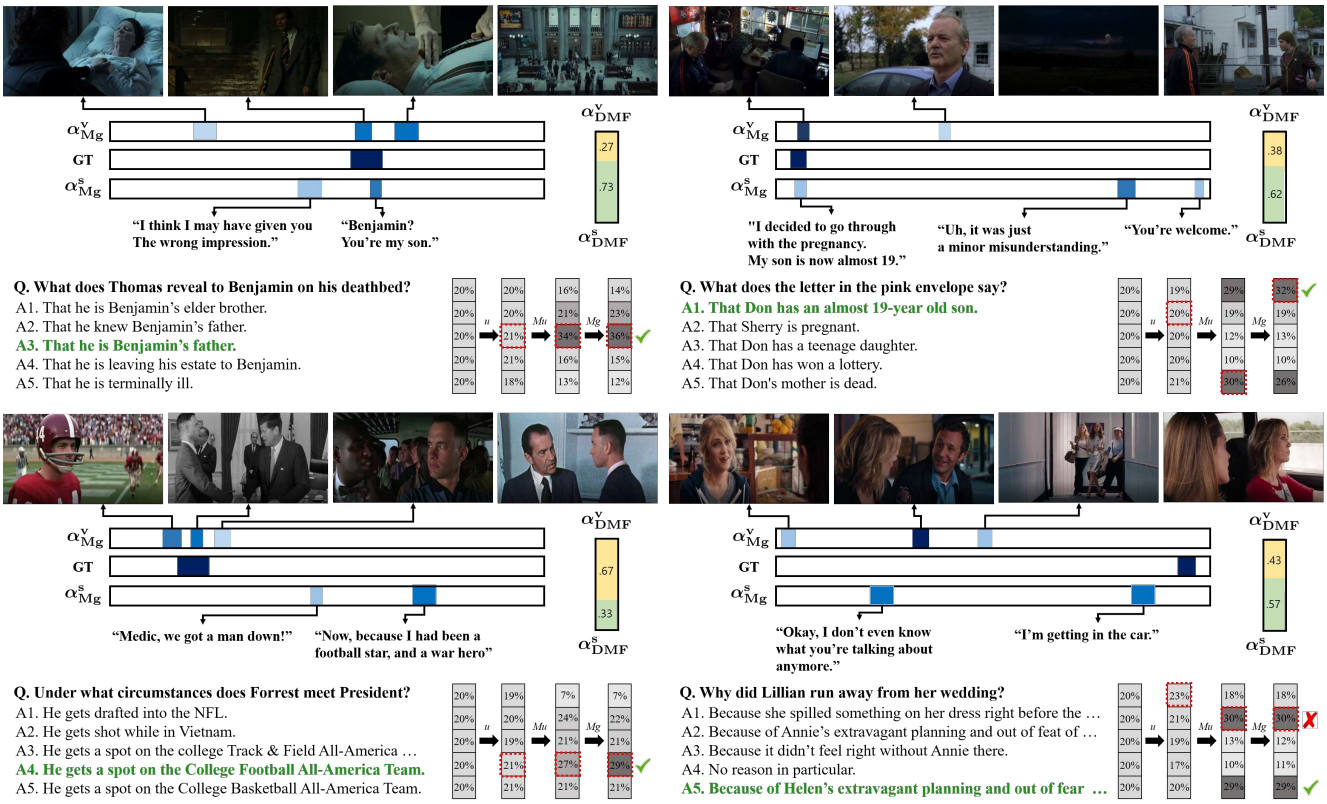

Figure 2. Qualitative examples of MovieQA benchmark solved by PAMN (the last example is failure case). Green sentences and check symbols indicate correct answers and red dotted boxes highlight PAMN's prediction at each belief correction step. For failure cases, red 'x' symbols indicate the the incorrect selection. $\alpha^v_{Mg}, \alpha^s_{Mg}$ represents temporal attention obtained by progressive attention mechanism, $\alpha^v_{\text{DMf}}, \alpha^s_{\text{DMF}}$ denotes attention obtained by dynamic modality fusion. The temporal attention by PAMN matches well with groundtruth (GT) where the question is generated. Observing diverse source of information, PAMN successfully corrects the belief toward correct answer.

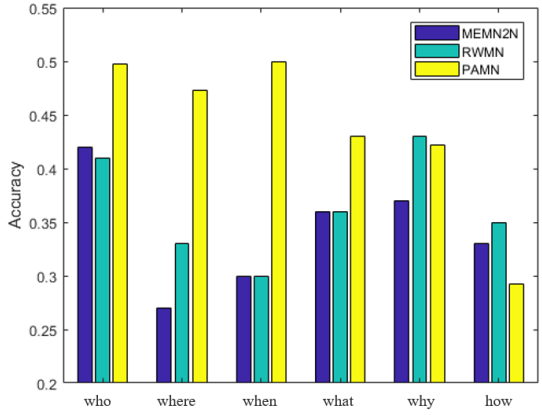

Figure 3. Accuracy comparison with respect to the first word of the question between MemN2N [26], RWMN [21] and PAMN on the validation set of MovieQA. PAMN outperforms on the majority of the question types.

whole movie (e.g., How do Schmidt and Jenko's fake identities end up getting switched?, Why does Mozart's financial situation get worse and worse?). We observed that PAMN outperforms MemN2N and RWMN on the majority of question types. Especially, PAMN attains 20% and 13% performance boosts on *when, where* questions, respectively which implies the superiority of PAMN to pinpoint the movie story.

5. Conclusion

In this paper, a movie story question answering (QA) architecture referred to as Progressive Attention Memory Network (PAMN) was proposed. The main challenges of movie story QA were summarized as: (1) pinpointing the temporal parts relevant to answer the question is difficult (2) different questions require different modality to infer the answer. Proposed PAMN make use of three main features to tackle aforementioned challenges: (1) progressive attention mechanism, (2) dynamic modality fusion and (3) belief correction answering scheme. We empirically demonstrated that proposed PAMN is valid by showing the state-of-the-art performance on MovieQA and TVQA benchmark dataset.

References

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016.

[2] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[4] Hedi Ben-younes, Remi Cadene, Matthieu Cord, and Nicolas Thome. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[5] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[6] Haoqi Fan and Jiatong Zhou. Stacked latent attention for multimodal reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[8] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[10] Andrew Howe and Phong Nguyen. Sat reading analysis using eye-gaze tracking technology and machine learning. In *Intelligent Tutoring Systems*, 2018.

[11] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Lu Jiang, Junwei Liang, Liangliang Cao, Yannis Kalantidis, Sachin Farfade, and Alexander G Hauptmann. Memexqa: Visual memex question answering. *arXiv:1708.01336*, 2017.

[13] Sunghun Kang, Junyeong Kim, Hyunsoo Choi, Sungjin Kim, and Chang D. Yoo. Pivot correlational neural network for multimodal video categorization. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[14] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations (ICLR)*, 2017.

[15] Kyung-Min Kim, Seong-Ho Choi, and Byoung-Tak Zhang. Multimodal dual attention memory for video story question answering. In *The European Conference on Computer Vision (ECCV)*, 2018.

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

[17] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[18] Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, and Alexander Hauptmann. Focal visual-text attention for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6135–6143, 2018.

[19] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[21] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[22] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] Mengye Ren, Ryan Kiros, and Richard Zemel. Image question answering: A visual semantic embedding model and a new dataset. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[24] Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[25] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, 2016.

[26] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] Ledyard R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, Sep 1966.

[28] 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 *Advances in Neural Information Processing Systems (NIPS)*. 2017.

[29] Bo Wang, Youjiang Xu, Yahong Han, and Richang Hong. Movie question answering: Remembering the textual cues for layered visual contents. In *AAAI Conference on Artificial Intelligence*, 2018.

[30] Jason Weston, sumit Chopra, and Antoine Bordes. Memory networks. In *International Conference on Learning Representations (ICLR)*, 2015.

[31] Zichao Yang, Xiaodong He, Jianfeng Gao, and Alex Smola. Stacked attention networks for image question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[33] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, 2017.

[34] 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.

[35] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.