Holistic Multi-modal Memory Network for Movie Question Answering

Anran Wang, Anh Tuan Luu, Chuan-Sheng Foo, Hongyuan Zhu, Yi Tay, Vijay Chandrasekhar

***Abstract*—Answering questions using multi-modal context is a challenging problem as it requires a deep integration of diverse data sources. Existing approaches only consider a subset of all possible interactions among data sources during one attention hop. In this paper, we present a Holistic Multi-modal Memory Network (HMMN) framework that fully considers interactions between different input sources (multi-modal context, question) at each hop. In addition, to hone in on relevant information, our framework takes answer choices into consideration during the context retrieval stage. Our HMMN framework effectively integrates information from the multi-modal context, question, and answer choices, enabling more informative context to be retrieved for question answering. Experimental results on the MovieQA and TVQA datasets validate the effectiveness of our HMMN framework. Extensive ablation studies show the importance of holistic reasoning and reveal the contributions of different attention strategies to model performance.**

***Index Terms*—Question answering, multi-modal learning, MovieQA.**

I. Introduction

Inspired by the tremendous progress in computer vision and natural language processing, there has been increased interest in building models that enable joint understanding of visual and textual semantics. This has led to the proliferation of works in related topics such as image-text retrieval [20], [7], [36], image/video captioning [43], [24], [26], [23], [39], [45], [44] and visual question answering (VQA) [1], [4], [11], [9]. In particular, the VQA task is challenging as it requires models to understand information encoded in images to answer questions.

Developing question answering (QA) systems that are able to attain an understanding of the world based on multiple sources of information beyond images alone is a natural next step. Several QA datasets incorporating multiple data modalities have recently been developed towards this end [12], [32], [13], [15]. In this work, we focus on the MovieQA [32] and TVQA [15] datasets, which require systems to demonstrate story comprehension by successfully answering multiple choice questions relating to videos and subtitles taken from movies or TV shows.

A key challenge in multi-modal QA is to integrate information from different data sources. Both query-to-context attention and inter-modal attention between videos and subtitles should be considered. Recently developed methods have adopted the classic strategies of early-fusion [22] and late-fusion [32], both of which have their limitations. Early-fusion of different modalities may limit the ability to pick up meaningful semantic correlations due to the increased noise at the feature level, while late-fusion does not allow for cross-referencing between modalities to define the higher level semantic features. Wang *et al.* [35] proposed to utilize inter-modal attention. However, their method does not fully integrate the input data, in that different attention stages consider different subsets of interactions between the question, video, and subtitles for context retrieval.

Moreover, answer choices are only considered at the final step of the system where they are matched against an already integrated representation of the input data. As a result, potentially useful context provided by the answer choices are not effectively utilized to enable the model to focus on relevant parts of the input data.

To address these limitations, we propose a Holistic Multi-modal Memory Network (HMMN) framework that builds upon the End-to-end Memory Network (E2EMN) [31]. E2EMN was originally proposed for textual question answering using context from a single modality, while we propose a multi-modal method to deal with both visual and textual modalities, and consider inter-modal, question-to-context, answer attentions at the same time.

Our framework differs from existing work in two ways. Firstly, it employs both inter-modal and query-to-context attention mechanisms for effective data integration at each hop. Specifically, our attention mechanism holistically investigates videos, subtitles, question to obtain a summarized context at each attention hop, which is different from existing methods that only consider a subset of interactions in each hop. Hence, query-to-context relationship is jointly considered while modeling the multi-modal relationship between context. Secondly, our framework considers answer choices not only at the answer prediction stage, but also during the context retrieval stage. Utilizing answer choices to hone in on relevant information is a common heuristic used by students when taking multiple-choice tests. Analogously, we thought this would help on the QA task by restricting the set of inputs considered by the model thus helping it sieve out signal from the noise. In particular, for correct answers, the retrieved answer-aware context should match the answer choice. Otherwise, the resultant context may convey different semantic meaning from the answer choice. In real-world question answering systems (such as chatbots), it is not uncommon to have an answer bank

Anran Wang, Anh Tuan Luu, Chuan-Sheng Foo, Hongyuan Zhu, and Vijay Chandrasekhar are with Institute for Infocomm Research, A*STAR, Singapore 138632 (e-mail: wang_anran@i2r.a-star.edu.sg; at.luu@i2r.a-star.edu.sg; foo_chuan_sheng@i2r.a-star.edu.sg; zhuh@i2r.a-star.edu.sg; vijay@i2r.a-star.edu.sg). Yi Tay is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: YTAY017@e.ntu.edu.sg). *(Corresponding author: Hongyuan Zhu).*

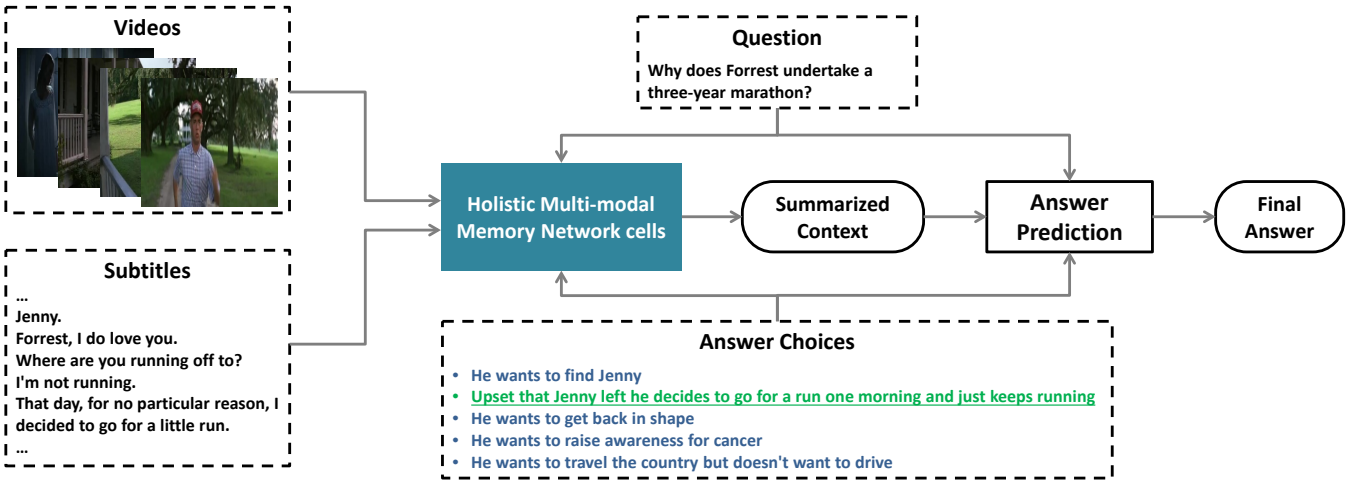

Fig. 1. Illustration of our proposed multi-modal feature learning framework. Holistic Multi-modal Memory Network cells integrate multi-modal context (videos, subtitles), the question, as well as answer choices. The generated summarized context will then be considered together with the question and answer choices for answer prediction. Our framework incorporates both inter-modal and query-to-context attention mechanisms in each attention hop, and incorporates answer choices at both context retrieval and answer prediction stages.

or use answer templates to handle frequently asked questions. In addition, in community question answering, users provide answer choices that are then voted upon by other users. Here, incorporating answer choices in the retrieval stage can be useful for ranking high-quality answers. Finally, some question answering systems have a separate answer generation stage during which several candidate answers are produced. These answers are then scored and ranked; our system may be seen to be addressing the answer ranking task following answer generation.

We evaluate our HMMN framework on the challenging MovieQA and TVQA datasets for video-based movie question answering. On the MovieQA dataset, HMMN achieves state-of-the-art performance on both validation and test sets; on the TVQA dataset, HMMN is competitive with state-of-the-art approaches. Ablation studies confirm the utility of incorporating answer choices during the context retrieval stage, demonstrating that incorporating answer choices contributes to the context retrieval process. In addition, we include analyses of different attention mechanisms (query-to-context attention, inter-modal attention, intra-modal self attention) and confirm the importance of holistic reasoning.

The rest of this paper is organized as follows. Section II introduces related works on multi-modal question answering and visual question answering. Section III describes the HMMN method in detail. Section IV presents the experimental results together with analyses of different attention mechanisms. Section V concludes the paper.

II. Related Work

A. Visual Question Answering

The release of large visual question answering (VQA) datasets [1], [4], [11] has fueled the popularity of the VQA task. Early works about VQA task use the holistic full-image feature to represent the visual context. For example, Malinowski *et al.* [21] proposed to feed the Convolutional Neural Network (CNN) image features and the question features together into a long short-term memory (LSTM) network and train an end-to-end network. Later, quite a few works have used attention mechanism to pay attention to certain parts of the image, where the alignment between image patches [41], [47] or region proposals [28] with the words in the question has been explored. Several attention mechanisms of connecting the visual context and the question have been proposed [40], [19]. For example, Lu *et al.* [19] presented a co-attention framework which considers visual attention and question attention jointly. Yang *et al.* [42] proposed stacked attention networks (SANs) for the VQA task. This method performs multiple steps of reasoning over the image and predicts the answer progressively. [42] shares a similar idea with the End-to-end Memory Network (E2EMN) [31], where summarized context from the image is fused with the question as the query for the next step. Singh *et al.* [30] performed extensive hyperparameter and architecture searches based on [33] for the VQA task. The results show that the design of attention mechanisms, gated activations and output structures has a significant impact on the performance. Wu *et al.* [37] proposed to incorporate high-level concepts and external knowledge for image captioning and visual question answering. Zhu *et al.* [46] presented an encoder-decoder model with a dual-channel ranking loss for video question answering. Their task is different from ours as we have not only video context but also accompanying subtitles. Answer attention was investigated for the grounded question answering task [6]. Grounded question answering is a special type of VQA, which is to retrieve an image bounding box from the candidate pool to answer the textual question. This method models the interaction between the answer candidates and the question and learns the answer-aware summarization of the question, while our method models the interaction between the answer choices and the context to retrieve more informative context.

B. Multi-modal Question Answering

In contrast to VQA, which only involves visual context, multi-modal question answering takes multiple modalities as context, and has attracted great interest. Kembhavi *et al.* [12] presented the Textbook Question Answering (TextbookQA) dataset that consists of lessons from middle school science curricula with both textual and diagrammatic context. In [13], PororoQA dataset was introduced, which is constructed from children cartoon Pororo with video, dialogue, and description. Tapaswi *et al.* [32] introduced the movie question answering (MovieQA) dataset which aims to evaluate the story understanding from both video and subtitle modalities. Lei *et al.* [15] presented the TVQA dataset that is constructed from popular TV shows. In this paper, we focus on the MovieQA and TVQA datasets, and related approaches are discussed as follows.

Most methods proposed for the MovieQA dataset are based on the End-to-end Memory Network [31], which was originally proposed for the textual question answering task. In [32], they proposed a straightforward extension of the End-to-end Memory Network [31] to multi-modal scenario on MovieQA. In particular, answer is predicted based on each modality separately, and late fusion is performed to combine the answer prediction scores from two modalities. Na *et al.* [22] proposed another framework based on the End-to-end Memory Network [31]. Their framework has read and write networks that are implemented by convolutional layers to model sequential memory slots as chunks. Context from textual and visual modalities are early fused as the input to the write network. Specifically, compact bilinear pooling [3] is utilized to obtain the joint embeddings with subshot and sentence features. Deep embedded memory networks (DEMN) model was introduced by [13], where they aim to reconstruct stories from a joint stream of scene and dialogue. However, their framework is not end-to-end trainable, as their method makes a hard context selection. Liu *et al.* [17] presented a method for MovieQA which propagates attention across different segments of the video and exploits answer attention as well. The main difference between [17] and our method is that their method retrieves context from the visual and textual modalities separately and applies a late fusion before answer prediction, while we explicitly reason multi-modal relationship in each attention hop. Liang *et al.* [16] presented a focal visual-text attention network which captures the correlation between visual and textual sequences for the personal photos and descriptions in the MemexQA dataset [10] and applied this method to the MovieQA dataset. Wang *et al.* [35] proposed a layered memory network with two-level correspondences. Specifically, the static word memory module corresponds words with regions inside frames, and the dynamic subtitle memory module corresponds sentences with frames. However, visual modality is used to attend to the textual modality which is dynamically updated by different strategies. Interactions between the question, videos, subtitles are not holistically considered in each attention stage. In [15], a baseline is provided along with the TVQA dataset that late-fuses different modalities for answer prediction.

TABLE I
SUMMARY OF VARIABLES.

Symbol	Definition
$q \in \mathbb{R}^d$	question feature
$A \in \mathbb{R}^{d\times k}$	answer choice features
$S \in \mathbb{R}^{d\times m}$	subtitle features
$V \in \mathbb{R}^{d\times n}$	video features
m	number of subtitle features
n	number of visual features
k	number of answer choices
d	dimension of visual and subtitle features

III. METHODOLOGY

We will first introduce the notations. Then, we will introduce the End-to-end Memory Network (E2EMN) [31]. After that, we introduce different attention strategies as building blocks. Then, Holistic Multi-modal Memory Network (HMMN) cell will be introduced together with the prediction framework.

A. Notation

We provide a summary of the symbols used in Table I. We let S and V denote feature vectors corresponding to the subtitle modality and video modality respectively, where d is the dimension of the feature vectors, n and m denote the number of frames and subtitles respectively. Feature vectors corresponding to the question q and answer choices A are represented in the same way as subtitle sentences.

In the MovieQA dataset, each question is aligned with several relevant video clips. We obtain features for frames and sentences following [35]. For the subtitle modality, we not only gather the subtitle sentences within the relevant video clips, but also incorporate nearby (in time) subtitle sentences to make use of the contextual information. The word2vec representation of each word in the subtitle is projected to d-dim with a projection matrix W_1. Then, a mean-pooling is performed among all words in each sentence to get the final sentence representation. Regarding the video modality, We select n frames from the relevant video clips for each question. Frame-level representations are generated by investigating attention between regional features and word representations in the vocabulary, where a projection matrix W_2 is utilized to project the d_r dimension regional VGG [29] features to match the d_w dimension of word representations. With regional features represented by the vocabulary word features, frame-level representations are generated with average pooling followed by a projection with W_1. We refer the reader to the original paper [35] or their released code for more details. The structures to generate representations for the subtitle and video modalities for MovieQA are shown in Fig. 2(a).

For the TVQA dataset, words in the subtitle modality are represented with GloVe embeddings [25]. For the video modality, we follow [15] to use visual concepts (vcpt) detected from frames to represent the video modality. Visual concepts are represented with GloVe embeddings. Visual concepts of

each image are mean pooled to generate frame-level representations. We use a Bi-directional LSTM to encode subtitle and video modalities.

B. End-to-end Memory Network

The End-to-end Memory Network (E2EMN) [31] is originally proposed for a question answering task where the aim is to pick the most likely word from the vocabulary as the answer according to the textual context. In [32], E2EMN is adapted to multi-modal question answering with multi-choice answers. In particular, scores from two modalities are late-fused to make the final prediction. As E2EMN is designed for textual question answering, this method only deals with context from a single modality. Here we use the subtitle modality S for explanation.

In E2EMN, input features of context S are treated as memory slots. With both memory slots and query (question is used as query here) as input, a summary of context is derived according to the relevance between the query and memory slots. In particular, the match between a query q and i-th memory slot $S_{:i}$ of subtitle S is calculated with the inner product followed by a softmax:

$$\alpha_i = softmax(q^T S_{:i}) \tag{1}$$

where $S_{:i}$ is the i-th column of S, and α_i indicates the importance of the i-th subtitle sentence to the query.

The summarized context u is computed as the weighted sum of subtitle sentence features based on α_i:

$$u = \sum_{i=1}^{m} \alpha_i S_{:i} \tag{2}$$

Finally, the answer prediction is made by comparing the answer choice a_i with the sum of query representation q and the summarized context u:

$$p = softmax((q+u)^T A) \tag{3}$$

where $p \in \mathbb{R}^k$ is the confidence vector. Here k is the number of answer choices.

The process to derive the summarized context can be performed in multiple hops, where the output of one layer can be used as part of the query of the next layer.

C. Different Attention Strategies

Conceptually, we consider two types of attention – using a feature matrix to attend to another feature matrix, and using a feature vector to attend to a feature matrix.

Matrix-matrix attention ($\mathcal{A} \rightarrow \mathcal{B}$). We use $\mathcal{A} \rightarrow \mathcal{B}$ to denote using feature matrix $\mathcal{A}$ to attend to feature matrix $\mathcal{B}$. Each column of the resultant representation $\mathcal{A}_{new}$ is computed as a weighted sum over columns of $\mathcal{B}$, where weights are given by the relevances (attention weights) to columns in $\mathcal{A}$.

Vector-matrix attention ($\sigma \Rightarrow \mathcal{B}$) We use $\sigma \Rightarrow \mathcal{B}$ to denote using feature vector σ to attend to feature matrix $\mathcal{B}$. Each column in $\mathcal{B}$ is re-weighted according to its relevance (or attention weight) to σ to generate a updated matrix $\mathcal{B}_{new}$.

In the following, we describe in detail how these two attention types are applied to the various input modalities.

(i) Query-to-context Attention

Query-to-context attention indicates which memory slots are more relevant to the query. Here the subtitle modality S is used as context for illustration. We denote this process as $q \Rightarrow S$, and the output is S'. With the calculated similarity between the query and each memory slot of S in Eq. 1, more relevant subtitle sentences can be highlighted with:

$$S'_{:i} = \alpha_i S_{:i} \tag{4}$$

(ii) Inter-modal Attention

The inter-modal attention, which is denoted as $S \rightarrow V$, indicates for each subtitle sentence, we intend to find the most relevant frames. The retrieved frame features will be fused to represent the subtitle sentence. The output can be interpreted as the video-aware subtitle sentence representation.

First, the coattention matrix between frames and subtitle sentences can be defined as:

$$\beta_{ij} = S_{:i}{}^T V_{:j} \tag{5}$$

where β_{ij} indicates the relevance between the j-th frame and the i-th subtitle sentence. Then the i-th subtitle sentence can be represented by the weighted sum of all frames based on β_{ij}:

$$\bar{S}_{:i} = \sum_{j=1}^{n} \beta_{ij} V_{:j} \tag{6}$$

The resulted representation for subtitle modality $\bar{S}$ is of the same size as S. Similarly, the result for $(V \rightarrow S)$ is $\bar{V}$.

(iii) Intra-modal Self Attention

Self attention has shown its power in tasks such as question answering and machine translation [8], [34]. The intuition is that contextual information among other memory slots can be exploited. Similar with the inter-modal attention, the intra-modal self attention, denoted as $S \rightarrow S$, considers attention between different memory slots in the same modality. The coattention matrix can be defined as:

$$\gamma_{ij} = I(i \neq j) S_{:i}{}^T S_{:j} \tag{7}$$

Noted that the correlation between one sentence with itself is set to zero. The resulted representation with self-attention is $\hat{S}$. Each subtitle sentence will be represented by the weighted sum of features of all the subtitle sentences based on γ_{ij}:

$$\hat{S}_{:i} = \sum_{j=1}^{m} \gamma_{ij} S_{:j} \tag{8}$$

D. Holistic Multi-modal Memory Network (HMMN)

Different from E2EMN, our HMMN framework takes multi-modal context as input. HMMN framework investigates interactions between multi-modal context and the question jointly. By doing this, query-to-context relationship is jointly considered while modeling the multi-modal relationship between context. In addition, it not only exploits answer choices for answer prediction but also in the process of summarizing the context from multiple modalities.

The inference process is performed by stacking small building blocks, called HMMN cell. The structure of HMMN cell

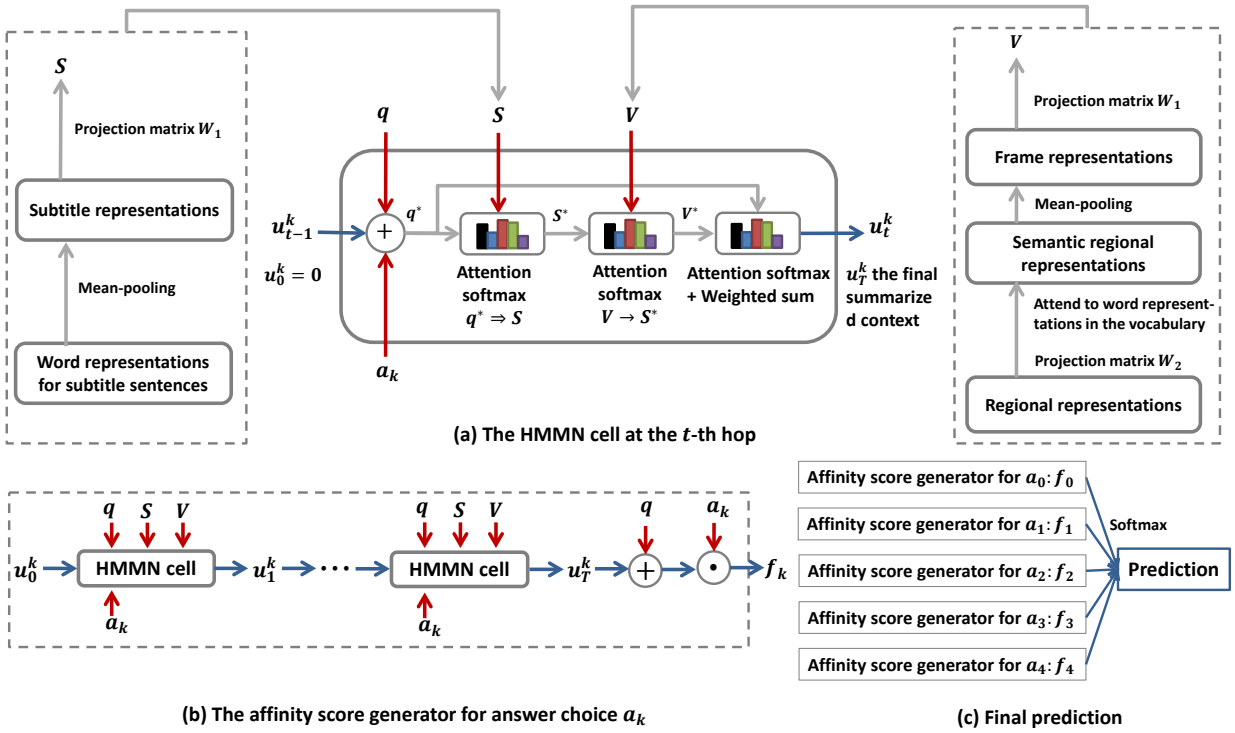

Fig. 2. Illustration of (a) the HMMN cell at the t-th hop, showing how representations for the subtitle and video modalities are generated (in the context of the MovieQA dataset), (b) the affinity score generator for answer choices a_k, and (c) how the final prediction is made. In (b), T denotes the number of hops, which is also the number of stacked HMMN cells. $\oplus$ denotes element-wise addition and $\odot$ denotes element-wise multiplication.

is shown in Fig. 2(a). The process to generate S and V is only for illustration, while our main contribution lies in the HMMN structure. Each HMMN cell takes as input the question, one answer choice, context from videos and subtitles, and derives the answer-aware summarized context. We call this process as one hop of reasoning. Let u_t^k be the output of the t-th reasoning hop with respect to answer choice k. The output of the t-th hop will be utilized as the input of the $(t+1)$-th hop.

1) Involving answers in context retrieval: The HMMN cell incorporates the answer choice as a part of the query in the context retrieval stage. The query involving k-th answer choice for the t-th hop is calculated by combining the output of the previous hop u_{t-1}^k, the question q, answer choice a_k:

$$q^* = u_{t-1}^k + a_k + \lambda q \tag{9}$$

where λ is a tradeoff parameter between the question and the rest of the query.

The intuition of incorporating answer choices in the context retrieval stage is to mimic behaviors of students who take a reading test with multi-choice questions. When the context is long and complicated, a quick and effective way to answer the question is to locate relevant information with respect to each answer choice. For one answer choice, if the retrieved answer-aware context conveys similar ideas with the answer choice, it tends to be the correct answer. Alternatively, if the retrieved context has a different semantic meaning, the answer is likely to be wrong.

2) Holistically considering different attention mechanisms in each hop: Instead of only taking a subset of interactions between the query and multi-modal context, our framework jointly considers inter-modal and query-to-context attention strategies in each hop.

The HMMN cell takes the query q^* to gather descriptive information from multi-modal context, where interactions between the question, answer choices, videos, subtitles are exploited holistically. In particular, we utilize the updated query to highlight the relevant subtitle sentences in S by performing the query-to-context attention (denoted as $(q^* \Rightarrow S)$). The resulted re-weighted subtitle modality is represented as S^*:

$$\begin{aligned} \delta_i &= softmax(q^{*T} S_{:i}) \\ S^*_{:i} &= \delta_i S_{:i} \end{aligned} \tag{10}$$

where more relevant subtitle sentences are associated with large weights.

Inter-modal attention reasoning is applied by using the video modality V to attend to the subtitle modality S (denoted as $(V \rightarrow S^*)$), which aims to generate the subtitle-aware representations for frames as V^*. Each frame is represented with the weighted sum of all the subtitle sentence features according to the relevance:

$$\begin{aligned} \varepsilon_{ij} &= V_{:i}^T S^*_{:j} \\ V^*_{:i} &= \sum_{j=1}^{m} \varepsilon_{ij} S^*_{:j} \end{aligned} \tag{11}$$

The resulted V^* can be summarized with respect to the query q^* as the hop output. In particular, the t-hop summarized

context with respect to the k-th answer choice is caculated as u_t^k:

$$\zeta_i = softmax(q^{*T} V_{:i}^*)$$
$$u_t^k = \sum_{i=1}^{n} \zeta_i V_{:i}^* \tag{12}$$

In each reasoning hop, the output of previous hop, the answer choice, the question, multi-modal context are holistically integrated. The reason of using V to attend to S (not using S to attend to V) is that, the subtitle modality is more informative than the video modality for the MovieQA and TVQA tasks. Typically, the subtitle modality includes descriptions of the story such as character relationships, story development. By attending to S, the feature representations in S are used to form the summarized context.

3) Predicting answer with affinity scores: It is shown in the original E2EMN that multiple hops setting yields improved results. We stack the HMMN cells to do T hops of reasoning. For each answer choice a_k, an affinity score f_k is generated by comparing the sum of the question q and answer-aware summarized context u_T^k with the answer choice a_k:

$$f_k = (q + u_T^k)^T a_k \tag{13}$$

The structure of generating the affinity score is shown in Fig. 2(b). This score indicates whether the retrieved context has the consistent semantic meaning with the answer choice. Then the affinity scores for all the answer choices $[f_0, f_1, f_2, f_3, f_4]$ are passed to a softmax function to get the final answer prediction as shown in Fig. 2(c), where each rectangle on the left corresponds to a structure in Fig. 2(b). The cross-entropy loss is minimized with the standard stochastic gradient descent. The key idea is that if one answer choice matches with the answer-aware summarized context, it is likely to be the correct answer.

In summary, we utilize the answer choice together with the question as the clue for context retrieval. We derive the re-weighted subtitle modality representation S^* with respect to the question as information screening. Then our method uses V to attend to S^* to generate subtitle-aware video representation which forms the summarized context for answer prediction.

IV. Experiments

A. Dataset and Setting

The MovieQA dataset [32] consists of 408 movies and 14,944 questions. Diverse sources of information are collected including video clips, plots, subtitles, scripts, and Descriptive Video Service (DVS). Plot synopses from Wikipedia are utilized to generate questions. For multi-modal question answering task with videos and subtitles, there are 6,462 questions with both videos clips and subtitles from 140 movies. We follow the public available train, validation, test split.

The TVQA dataset [15] is constructed from 6 popular TV shows, such as "Friends" and "How I met your mother". It contains compositional questions that require both visual and textual understanding. Each question is accompanied by 5 answer choices, just as in the MovieQA dataset. There are 122,039 questions for training, 15,253 for validation, and 15,253 for testing. For both datasets, we report accuracy – how many multiple-choice questions are answered correctly.

TABLE II
Performance of HMMN variants on the validation set of the MovieQA dataset.

Method	Accuracy (%)
HMMN (1 layer) w/o answer attention	43.35
HMMN (2 layers) w/o answer attention	**44.47**
HMMN (3 layers) w/o answer attention	44.24
HMMN (1 layer)	45.71
HMMN (2 layers)	**46.28**
HMMN (3 layers)	44.13

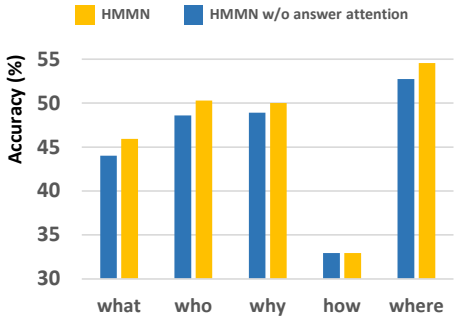

Fig. 3. Effect of answer attention on different question types in the MovieQA dataset.

B. Implementation Details

For the MovieQA dataset, for the subtitle modality, we consider the subtitle sentences which fall into the time interval derived by extending the starting and ending time points of video clips by 300 seconds. For the video modality, 32 frames are selected from the relevant video clips following [35]. We use the word2vec representations provided by [32]. The dimension of the word2vec representation d_w is 300. The dimension of the regional features from 'pool5' of VGG-16 d_r is 512. 10% of the training examples are kept as the development set. The batch size is 8. The learning rate is set to 0.005. The tradeoff parameter λ is set to be 0.45. The dimension of features d is set to 300. Our model is trained up to 50 epochs, and early stopping is performed.

For the TVQA dataset, we consider subtitle sentences that fall into the time interval derived by extending the starting and ending time points of video clips by 10 seconds. For the video modality, all frames in the dataset for each question are considered. The batch size is 32. The size of the hidden state in the Bi-directional LSTM is set to 150. The learning rate is 0.0003. We keep the setting for the tradeoff parameter λ the same as MovieQA.

C. Quantitative Analysis

1) The MovieQA dataset: Table II presents results for HMMN structures with and without answer attention with different numbers of layers. The HMMN structure without

TABLE III
COMPARISON OF STATE-OF-THE-ART METHODS ON THE MOVIEQA DATASET.

Method	Accuracy on validation set (%)	Accuracy on test set (%)
Tapaswi *et al.* [32]	34.20	-
Na *et al.* [22]	38.67	36.25
Kim *et al.* [13]	44.7	34.74
Liang *et al.* [16]	41.0	37.3
Liu *et al.* [17]	41.66	41.97
Wang *et al.* [35]	42.5	39.03
Our HMMN framework w/o answer attention	44.47	41.65
Our HMMN framework	**46.28**	**43.08**

TABLE IV
PERFORMANCE OF BASELINES WITH DIFFERENT ATTENTION STRATEGIES ON THE VALIDATION SET OF THE MOVIEQA DATASET.

Method	Accuracy (%)
V	37.69
S	39.62
V' $(q \Rightarrow V)$ Query-to-context Attention	37.92
S' $(q \Rightarrow S)$ Query-to-context Attention	40.86
$\bar{V}$ $(V \rightarrow S)$ Inter-modal Attention	42.73
$\bar{S}$ $(S \rightarrow V)$ Inter-modal Attention	35.10
$\hat{V}$ $(V \rightarrow V)$ Intra-modal Self Attention	37.92
$\hat{S}$ $(S \rightarrow S)$ Intra-modal Self Attention	40.29

TABLE V
PERFORMANCE OF BASELINES USING VARIOUS HIGHER-LEVEL INTER-MODAL ATTENTION STRATEGIES ON THE VALIDATION SET OF THE MOVIEQA DATASET.

Method	Accuracy (%)	Method	Accuracy (%)
$V \rightarrow S$	42.73	$S \rightarrow V$	35.10
$V \rightarrow S'$	**43.35**	$S' \rightarrow V$	35.21
$V \rightarrow \bar{S}$	37.47	$\bar{S} \rightarrow V$	35.10
$V \rightarrow \hat{S}$	41.08	$\hat{S} \rightarrow V$	35.10
$V' \rightarrow S$	43.12	$S \rightarrow V'$	35.21
$V' \rightarrow S'$	**43.35**	$S' \rightarrow V'$	35.44
$V' \rightarrow \bar{S}$	38.14	$\bar{S} \rightarrow V'$	35.32
$V' \rightarrow \hat{S}$	37.02	$\hat{S} \rightarrow V'$	35.44
$\bar{V} \rightarrow S$	41.76	$S \rightarrow \bar{V}$	40.29
$\bar{V} \rightarrow S'$	41.08	$S' \rightarrow \bar{V}$	40.18
$\bar{V} \rightarrow \bar{S}$	39.84	$\bar{S} \rightarrow \bar{V}$	40.85
$\bar{V} \rightarrow \hat{S}$	37.47	$\hat{S} \rightarrow \bar{V}$	38.60
$\hat{V} \rightarrow S$	43.12	$S \rightarrow \hat{V}$	34.55
$\hat{V} \rightarrow S'$	**43.35**	$S' \rightarrow \hat{V}$	35.77
$\hat{V} \rightarrow \bar{S}$	37.58	$\bar{S} \rightarrow \hat{V}$	35.10
$\hat{V} \rightarrow \hat{S}$	38.15	$\hat{S} \rightarrow \hat{V}$	34.98

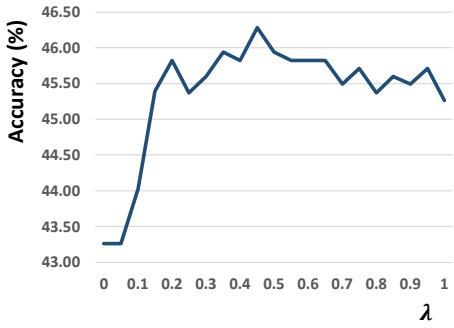

Fig. 4. Effect of λ as evaluated on the validation set of the MovieQA dataset. A large λ assigns more weight to the question q.

TABLE VI
PERFORMANCE ON THE VALIDATION SET OF THE TVQA DATASET.

Method	Accuracy (%)
V	46.10
S	61.38
V' $(q \Rightarrow V)$ Query-to-context Attention	45.42
S' $(q \Rightarrow S)$ Query-to-context Attention	61.69
$\bar{V}$ $(V \rightarrow S)$ Inter-modal Attention	62.65
$\bar{S}$ $(S \rightarrow V)$ Inter-modal Attention	44.94
$\hat{V}$ $(V \rightarrow V)$ Intra-modal Self Attention	45.83
$\hat{S}$ $(S \rightarrow S)$ Intra-modal Self Attention	61.17
Our HMMN framework w/o answer attention	63.44
Our HMMN framework	65.03
Lei *et al.* [15]	67.70
Lei *et al.* [15] + ours	**69.42**

answer attention means not considering answer choices when retrieving the context. This model can be easily extended for the opening question answering tasks by replacing the answer prediction step in Eq. 3 with a decoder that generates free-form answers as done in [2]. We can see that by incorporating the answer choices in the context retrieval stage, the performance is significantly improved. And 2-layer structures achieve the best performance, thus the number of layers which is also the number of hops T is set to 2. Fig. 3 shows the comparison of HMMN w/ and w/o answer attention for different question types, with the starting word as 'what', 'who', 'why', 'who', 'where'. It can be seen that HMMN framework performs consistently better than the HMMN framework w/o answer attention. In Fig. 4, effects of using different settings of λ are shown.

TABLE VII
EFFECT OF BI-LSTM ON THE VALIDATION SET OF THE TVQA DATASET.

Method	Accuracy (%)
Our HMMN framework w/o Bi-LSTM	64.17
Our HMMN framework with Bi-LSTM	65.03

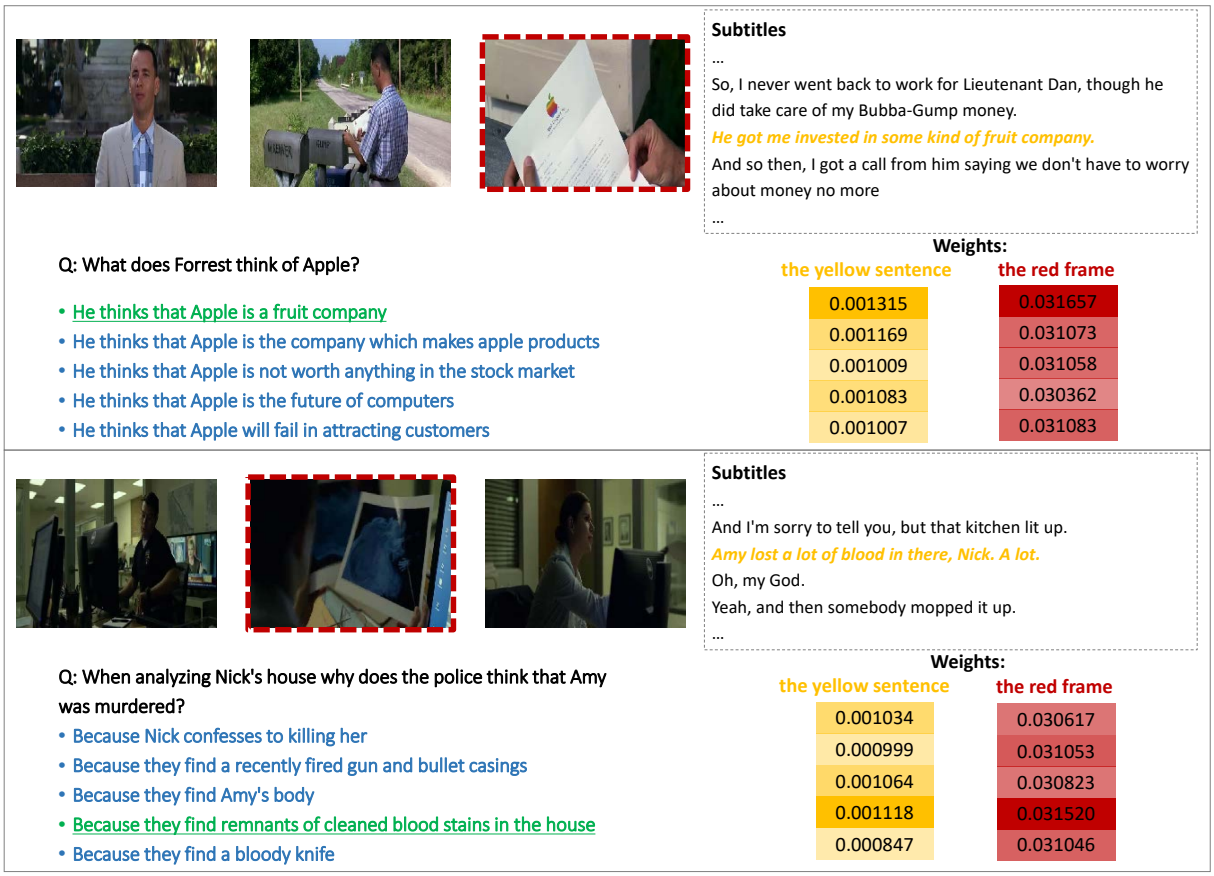

Fig. 5. Visualization of model weights for relevant image frames and sentences with respect to different answer choices. We show two correct predictions from the HMMN framework on the MovieQA dataset. The correct answer choice is colored in green. The relevant frame and subtitle sentence are highlighted in red and yellow respectively and their weights according to different answer choices are shown in red and yellow blocks. When using the correct answer choice to search for the relevant context, the relevant frame and sentence are associated with higher weights than those of other answer choices.

Table III shows the comparison with state-of-the-art methods. We compare with [32], [22], [13], [17], [16], [35]. It can be seen that our proposed method significantly outperforms state-of-the-art methods on both validation and test sets.

In particular, [32], [17] performs a late fusion. [22] conducts an early fusion. Neither late fusion nor early fusion can well exploit the relationship between modalities, which results in suboptimal results. [13] is not end-to-end trainable. [16] assumes sequences of the visual and textual representations have the same length, which is not true for MovieQA. Similar with [22], [16] cuts off visual information of those frames without accompanying subtitles. [35] takes the multi-modal relationship into consideration, however, in each attention stage, a different subset of interactions between the question, videos, subtitles are considered. In comparison, our HMMN framework w/o answer attention holistically incorporates the output of previous hop, the question, videos, subtitles in each hop, which leads to superior performance.

2) Ablation study: In this paper, we also explore different attention strategies. In particular, we show the results with memory slots derived from: original features: S, V; "Query-to-context Attention": S', V'; "Inter-modal Attention": $\bar{S}$, $\bar{V}$; and "Intra-modal Self Attention": $\hat{S}$, $\hat{V}$. Each of them can be treated as the representation of memory slots in the original E2EMN framework. Similarly, Eq. 1, Eq. 2 and Eq. 3 can be applied to predict the answer. Table IV shows results of baselines with different attention strategies. The baseline S performs better than V, as the subtitle modality contains more informative descriptions of character relationships and story development. By using each attention strategy to obtain the enhanced representations, the performance improvement is achieved. Particularly, the inter-modal attention $(V \rightarrow S)$ brings a significant improvement.

To explore higher-level inter-modal attention, we use V, V', $\bar{V}$, $\hat{V}$ to attend to S, S', $\bar{S}$, $\hat{S}$, and vice versa with Eq. 5 and Eq. 6. The results are shown in Table V. Typically, the baselines on the left side perform better than ones on the right side. As mentioned in the methodology section, it is because when we use $V \rightarrow S$ attention, the video modality will be represented by the subtitle features, which are more descriptive. According to this observation, when designing the structure of attention mechanism for any general multi-modal tasks, using the less discriminative modalities as clues to attend to the more discriminative modalities tends to achieve better performance. We can observe that the intra-modal self attention does not bring much improvement to this task. Baselines of using the video modality to attend to re-weighted subtitle modality $(V \rightarrow S')$ perform considerably well, and our 1-layer HMMN framework w/o answer attention degenerates

Fig. 6. Visualization of model weights for image frames and sentences with respect to the correct answer choice, from the HMMN framework on the MovieQA dataset. The correct answer choice is colored in green. The most relevant frames and subtitles for question answering are assigned large weights.

to the $V \rightarrow S'$ baseline.

3) The TVQA dataset: We present results on the validation set of the TVQA dataset in Table VI. Performance on the TVQA dataset follows a similar pattern to that on MovieQA. Models using the subtitle modality outperform those using the video modality, and inter-modal attention provides significant improvements compared to query-to-context and intra-modal self attention. Our HMMN framework achieves improved performance compared to the baseline without answer attention. However, our method does not outperform the baseline presented in [15]. By combining our prediction scores with theirs by averaging the two, we obtain a 1.72% improvement, indicating that our method brings important complementary information. In Table VII, it can be observed that using a Bi-LSTM to encode features improves performance on the TVQA dataset.

D. Qualitative Analysis

1) Attention Weight Visualization: To demonstrate that relevant context can be well captured by the HMMN framework, we visualize the attention weights of frames and subtitles of two success cases of the MovieQA dataset in Fig. 5. The observed relevant frame and subtitle are highlighted in red and yellow respectively. For the question "What does Forrest think of Apple?", the first answer choice is correct. Following attention weights of the second layer are visualized: 1) the attention weight of the subtitle sentence with respect to the answer choice and question; 2) the attention weight of the frame with respect to the answer choice and question. By using the first answer choice to retrieve the context, the relevant frame and subtitle sentence are associated with larger weights compared to those of other answer choices. Thus the key information can be captured and a high affinity score will be generated. On the other hand, the HMMN framework w/o answer attention picks the second answer choice as the correct one. This is because that the word "Apple" is not mentioned in the subtitle, thus by using the question only to retrieve the information, important context about "Apple" is missed. Similar remarks can be made for the second example question in Fig. 5, the relevant frame and subtitle sentence are given high weights. In Fig. 6, we visualize the weights of different sentences and image frames with respect to the correct answer choice. It can be observed that relevant sentences and frames for question answering such as "My name's Forrest Gump" and the 4-th image are highlighted.

2) Success and Failure Cases: Fig. 7(b) shows 3 typical failure cases on the MovieQA dataset. In the first example, although both video and subtitle modalities contain the information that Harry's father made fun of Snape, it is difficult to associate them with the word "bully" that has a high-level semantic meaning, where common sense knowledge reasoning is required. The second example is from "Gone girl". Although the husband is not happy with the main character Amy, the actors act as having a happy ending. This question also requires common sense to answer. In the third example, the revolver appears in the frames, but is not mentioned in the subtitles. Although we use conventional CNN features to generate frame-level representations, the associations between

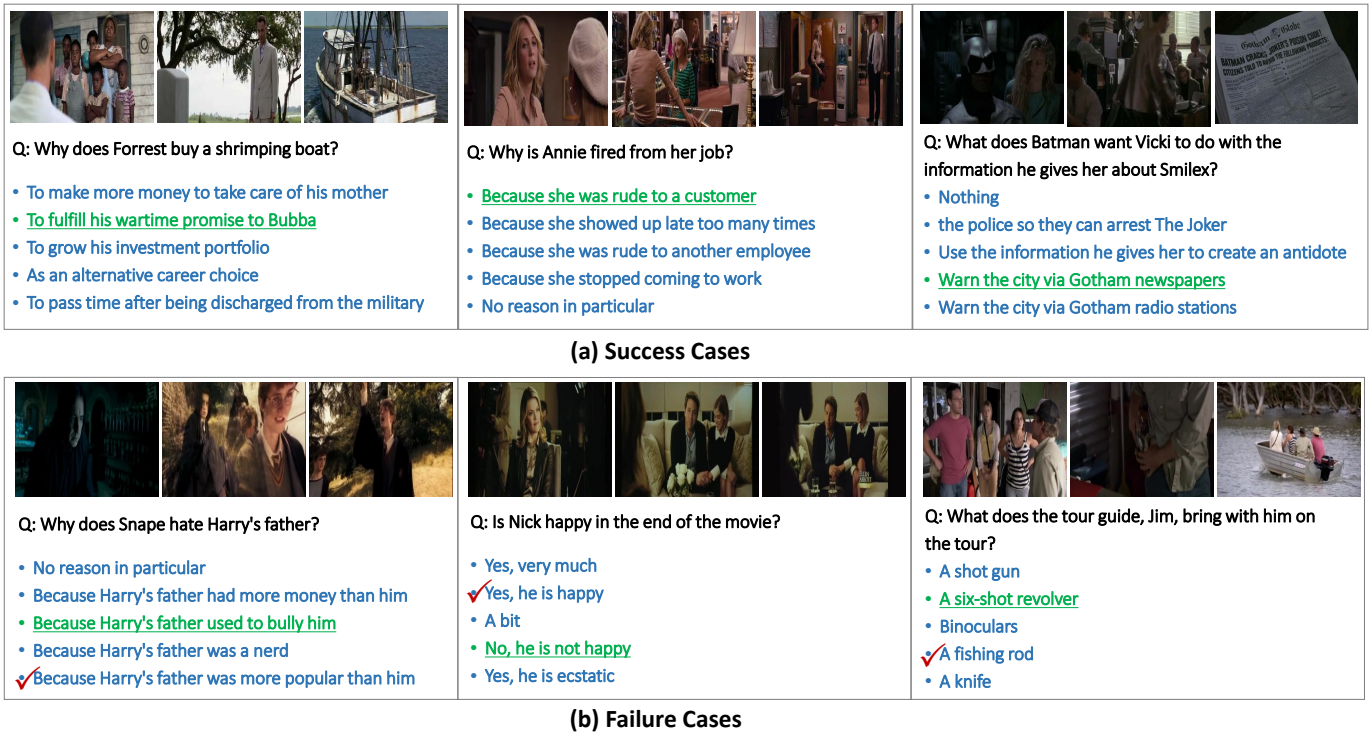

Fig. 7. Success and failure cases for HMMN on the MovieQA dataset. Correct answer choices are in green while mistakes made by our framework are indicated by red check symbols.

Fig. 8. Examples where the method in [35] fails but HMMN makes correct predictions, taken from the MovieQA dataset. Correctly predicted answer choices from HMMN are in green and mistakes made by [35] are indicated by red check symbols.

visual patterns and object labels are not enforced during training. More success and failure cases can be found in Fig. 7. In addition, we show some cases where our method succeeds while [35] fails in Fig. 8.

E. Future Work

The failure cases indicate the necessity of exploiting common sense knowledge, which we leave for future work. High-level semantic information can be injected to the network by leveraging well-built knowledge graph ConceptNet [18] as done in [14]. In addition, the question answering tasks on the MovieQA and TVQA datasets require a significant amount of supervised data that is hard to obtain. A future direction is to develop a framework for the few-example question answering tasks as what has been explored on the person re-identification task [38]. To build the unlabeled data, questions can be generated by question generation methods [27], [5], and answer choices can be generated by pre-trained free-form answer generators (wrong answer choices can be derived from answers generated for other questions). A progressive learning framework can be proposed to gradually exploit the unlabeled data.

V. Conclusion

We presented a Holistic Multi-modal Memory Network framework that learns to answer questions with context from multi-modal data. In our proposed HMMN framework, we investigate both inter-modal and query-to-context attention mechanisms to jointly model interactions between multi-modal context and the question. In addition, we explore the benefits of answer attention by incorporating answer choices during the

context retrieval stage. Our HMMN framework achieves state-of-the-art results on the MovieQA dataset and competitive results on the TVQA dataset. We also presented a detailed ablation study for different attention mechanisms, which could provide guidance for future model design.

REFERENCES

[1] 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.

[2] Zhiwei Fang, Jing Liu, Yong Li, Yanyuan Qiao, and Hanqing Lu. Improving visual question answering using dropout and enhanced question encoder. *Pattern Recognition*, 90:404–414, 2019.

[3] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *CVPR*, pages 317–326, 2016.

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, pages 6904–6913, 2017.

[5] Vrindavan Harrison and Marilyn Walker. Neural generation of diverse questions using answer focus, contextual and linguistic features. *arXiv preprint arXiv:1809.02637*, 2018.

[6] Junjie Hu, Desai Fan, Shuxin Yao, and Jean Oh. Answer-aware attention on grounded question answering in images. In *AAAI Fall Symposium*, 2017.

[7] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *CVPR*, pages 2310–2318, 2017.

[8] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *ICLR*, 2018.

[9] Unnat Jain, Svetlana Lazebnik, and Alexander Schwing. Two can play this game: Visual dialog with discriminative question generation and answering. In *CVPR*, pages 5754–5763, 2018.

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

[11] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017.

[12] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, pages 4999–5007, 2017.

[13] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: video story qa by deep embedded memory networks. In *IJCAI*, pages 2016–2022, 2017.

[14] Leo Laugier, Anran Wang, Chuan-Sheng Foo, Guenais Theo, and Vijay Chandrasekhar. Compositional attention networks for machine reasoning. In *ICLR workshop*, 2019.

[15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, pages 1369–1379, 2018.

[16] Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, and Alexander Hauptmann. Focal visual-text attention for visual question answering. In *CVPR*, pages 6135–6143, 2018.

[17] Chen Ding-Jie Chen Hwann-Tzong Liu, Chao-Ning and Tyng-Luh Liu. A2A: Attention to attention reasoning for movie question answering. In *ACCV*, pages 404–419, 2018.

[18] Hugo Liu and Push Singh. Conceptnet?a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[19] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, pages 289–297, 2016.

[20] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, pages 2623–2631, 2015.

[21] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, pages 1–9, 2015.

[22] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *CVPR*, pages 677–685, 2017.

[23] Fudong Nian, Teng Li, Yan Wang, Xinyu Wu, Bingbing Ni, and Changsheng Xu. Learning explicit video attributes from mid-level representation for video captioning. *Computer Vision and Image Understanding*, 163:126–138, 2017.

[24] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *CVPR*, pages 1029–1038, 2016.

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[26] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 6077–6086, 2017.

[27] Mrinmaya Sachan and Eric Xing. Self-training for jointly learning to ask and answer questions. In *NAACL*, pages 629–640, 2018.

[28] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, pages 4613–4621, 2016.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

[30] Jasdeep Singh, Vincent Ying, and Alex Nutkiewicz. Attention on attention: Architectures for visual question answering (vqa). *arXiv preprint arXiv:1803.07724*, 2018.

[31] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *NIPS*, pages 2440–2448, 2015.

[32] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *CVPR*, pages 4631–4640, 2016.

[33] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *CVPR*, pages 4223–4232, 2018.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[35] Bo Wang, Youjiang Xu, Yahong Han, and Richang Hong. Movie question answering: Remembering the textual cues for layered visual contents. In *AAAI*, pages 7380–7387, 2018.

[36] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018.

[37] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1367–1381, 2018.

[38] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wei Bian, and Yi Yang. Progressive learning for person re-identification with one example. *IEEE Transactions on Image Processing*, 28(6):2872–2881, 2019.

[39] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *ACM Multimedia Conference on Multimedia Conference*, pages 1029–1037, 2018.

[40] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, pages 2397–2406, 2016.

[41] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, pages 451–466. Springer, 2016.

[42] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016.

[43] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016.

[44] Mingxing Zhang, Yang Yang, Hanwang Zhang, Yanli Ji, Heng Tao Shen, and Tat-Seng Chua. More is better: Precise and detailed image captioning using online positive recall and missing concepts mining. *IEEE Transactions on Image Processing*, 28(1):32–44, 2019.

[45] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, pages 8739–8748, 2018.

[46] 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.

[47] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, pages 4995–5004, 2016.