Learnable Aggregating Net with Diversity Learning for Video Question Answering

Xiangpeng Li
Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China
xiangpengli.cs@gmail.com

Lianli Gao*
Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China
lianli.gao@uestc.edu.cn

Xuanhan Wang
Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China
wangxuanhan@uestc.edu.cn

Wu Liu
AI Research of JD.COM
liuwu@live.com

Xing Xu
Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China
xing.xu@uestc.edu.cn

Heng Tao Shen
Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China
shenhengtao@hotmail.com

Jingkuan Song
Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China
jingkuan.song@gmail.com

ABSTRACT

Video visual question answering (V-VQA) remains challenging at the intersection of vision and language, where it requires joint comprehension of video and natural language question. Image-Question co-attention mechanism, which aims at generating a spatial map highlighting image regions relevant to answering the question and vice versa, has obtained impressive results. Despite the success, simply applying co-attention to video visual question answering results in unsatisfactory performance due to the complexity and temporal nature of videos. In this paper, we proposed a novel architecture, namely Learnable Aggregating Net with Diversity learning (LAD-Net), for V-VQA. In the proposed method, we address two central problems: 1) how to deploy co-attention to V-VQA task considering the complex and diverse content of videos; and 2) how to aggregate the frame-level features without destroying the feature distributions and temporal information. To solve these problems, our LAD-Net first extends single-path based co-attention mechanism to a multi-path pyramid co-attention structure with a novel diversity learning to explicitly encourage attention diversity. For video-level (or question-level) descriptor, instead of taking a simple temporal pooling (i.e., average pooling), we propose a new learnable aggregation method with a set of evidence gates. It automatically aggregates adaptively-weighted frame-level features (or word-level features) to extract rich video (or question) context semantic information by imitating Bags-of-Words (BoW) quantization. With evidence gates, it then further chooses the most related signals representing the evidence information to predict the answer.Extensive validations on the two challenging video visual question answering datasets TGIF-QA and TVQA show that LAD-Net achieves the state-of-the-art performance under various settings and metrics. Our proposed strategies are of particular importance for improving the performance of the baseline co-attention V-VQA.

CCS CONCEPTS

• **Information systems** → **Question answering**.

KEYWORDS

Video Visual Question Answering; Diversity Learning; Learable Aggregation; Muti-path Pyramid Co-attention

ACM Reference Format:
Xiangpeng Li, Lianli Gao, Xuanhan Wang, Wu Liu, Xing Xu, Heng Tao Shen, and Jingkuan Song. 2019. Learnable Aggregating Net with Diversity Learning for Video Question Answering. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350971

*Lianli Gao is the corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350971

1 INTRODUCTION

Free-form and *open-ended* image visual question answering (I-VQA) has been introduced in [2], which requires a model to take an image and a free-form, open-ended, natural language question about the image as input to produce a natural language answer. It aims to mirror real-world scenarios to help visually-impaired users or intelligence analysts to simultaneously understand visual content and question context information [2]. To date, we have witnessed a significant attention [3, 4] from the computer vision and natural language processing communities to solve the I-VQA problem, and great success has been achieved [8, 15, 18, 20, 22, 24, 26, 30, 34]. Most of the existing methods are focusing on boosting the performance by providing an efficient attention mechanism to tell which regions in an RGB image to attend for each word [11, 15, 18, 22, 24, 30, 34]. Among them, co-attention mechanisms [18, 21, 22, 24, 30, 34], which are designed to jointly learn both image and question attention, are extremely powerful and have obtained impressive results.

Essentially, I-VQA [2] is primarily focusing on images and thus it would be difficult to mirror the real-world scenarios. As a result, Jang *et al.* [14], later on, formally proposed a video visual question answering (V-VQA) task by extending I-VQA to the video domain. Compared with I-VQA where a question is just about one RGB image, V-VQA task is more challenging due to three reasons. First, a video tends to have substantial redundant information. Second, a video related question, such as repetition count, repeating action and state transition, is related to some key frames which are difficult to be localized. Three, how to efficiently aggregate video and sentence feature to capture the true distribution of the answer has been rarely studied. Existing V-VQA approaches [9, 14] are focusing only on constructing visual attention map for appearance feature, temporal features or concatenated spatial-temporal feature, which does not jointly reason about the content and context information for both video and question. Inspired by the success of co-attention in I-VQA task, we aim to extend the co-attention to the video domain to boost the performance of V-VQA.

Despite the success, image-question co-attention cannot directly be applied for V-VQA for two deficiencies. First, a co-attention with single transformation path lacks the diversity and thus it cannot capture distinct, complementary and informative features, especially for input videos with length variety. For instance, when asked "what does the bear on right do after sitting", it requires multi-path attention maps to remove the irrelevant video clips, to increase-weight of the video clips describing "a bear on right doing sitting action", and to localize the the specific video clips with then answer (i.e., "stand" action). Secondly, existing co-attention adopts the common feature aggregation methods, such as weighted-sum, which may change the feature distributions and result in semantic context information loss and enable inaccurate answer prediction. Furthermore, recurrent-based aggregation may provides an solution, but its training is cumbersome due to requiring large amount of video samples. Therefore, a new non-recurrent aggregation approach is required to efficiently and effectively pick the evidence for the final answer prediction.

To address the above mentioned issues, we propose a novel architecture, namely Learnable Aggregation Net with Diversity Learning (LAD-Net), to improve the performance of V-VQA. Specifically, our LAD-Net consists of two major components. The first component is a novel multi-path pyramid co-attention modular with diversity learning to adaptively learn different attention maps to locate key video clips in the given video or important words in the given sentence. In this work we demonstrate that diversity attention maps are extremely beneficial in the context of V-VQA task. The second component is a new learnable aggregation modular with evidence gates. Firstly, it automatically fuses weighted frame-level feature or word-level features to obtain a descriptor to describe the global context information of a video or sentence. This is done by imitating the non-recurrent Bags-of-Words (BoW) quantization. Next, the two global descriptors are concatenated and further recalibrated with a set of learned evidence gates for capturing signals, which are on-topic contributions to the answer prediction. In this work, we also show that jointly conducting the development of better attention mechanism and improvement of aggregation of features extracted from an input video and question has potential to boost V-VQA performance considerably.

In summary, we make the following contributions in this paper: 1) To handle the complexity of videos in V-VQA, we propose a multi-path pyramid co-attention mechanism with a diversity learning to explicitly encourage attention diversity. This strategy benefits the capturing of distinct, complementary and informative features; 2) To aggregate the sequential features without destroying the feature distributions and temporal information, we propose a new learnable aggregation component. It imitates Bags-of-Words (BoW) quantization mechanism to automatically aggregate adaptively-weighted frame-level feature (or word-level feature). 3) We extensively evaluate the effectiveness of the overall model on two publicly available datasets (i.e., TGIF-QA and TVQA) for V-VQA task. The experimental results demonstrate that our model outperforms the existing state-of-the-art by a large margin.

2 RELATED WORK

In this section, we review the recent studies about visual question answering. In general, visual question answering can be divided into two sub-categories according to their inputs. The first one is image visual question answering [1, 15, 23, 25]. The second one is the video visual question answering [9, 14, 29, 32, 36].

2.1 Image Visual Question Answering

For image visual question answering task, researchers are focused on two lines. The first line is mainly about developing better attention mechanisms to facilitate the understanding of content and context of both image and question as well as their relations to the answer. The second line is to design better aggregation approaches to fuse features extracted from an input image and question.

For the first research line, the effective role of attention has been verified in [10, 11, 15, 20, 22, 24, 31, 37]. In general, an attention mechanism is conducted by firstly computing weights and then assigning them to image regions based on their relevance to the input question. For instance, Yang *et al.* [31] proposed a stack attention network (SAN) and it uses the question representation as a query to search for the most relevant regions in an image. Also, a multi-step reasoning process is introduced to imitate human behavior, and thus it produces a sequential attention maps to infer

the final answer step-by-step. However, it only considers the visual-modality data without considering the text-modality data, which plays an important role in I-VQA task.

Motivated by this, co-attention mechanism was firstly introduced to I-VQA by Lu *et al.* [22]. They introduced a hierarchical co-attention for jointly modeling visual and textual attention at the same time. It not only teaches the model "where to look", but also guides the model "which word should to focus on". Performance of co-attention demonstrates that both visual attention and question attention are essential for visual reasoning. Inspired by the success of co-attention, Nguyen *et al.* [24] proposed an improved Dense Co-attention Network (DCN) for image visual question answering, which enables the input question and image information interacting with each other. Specifically, DCN is also regarded as a multi-step reasoning approach, since it adopts a multi-layer co-attention to enhance final prediction accuracy. Considering that attention mechanism was often adopted to process image regions separately and ignored the cross relations of these regions, Zhu *et al.* [37] proposed a structure attention model to address this issue. They proposed a model, namely structured attention with Conditional Random Field (CRF) over image regions, to explore the relations between image regions to further improve the prediction accuracy.

For the second research line, several feature aggregation approaches are proposed to address the image question answering tasks [6, 12, 16, 17, 35]. Specifically, they focused on embedding visual information and question representation into a joint space, which aims to obtain a better fused representation of question-image pair. Multimodal Residual Learning (MRN) proposed by Kim *et al.* [16] employed element-wise multiplication for the joint residual mappings, and further exploited the residual learning of the attentional models. It learns the joint representation from both vision and language information. Following the work of Gao *et al.* [12], Fukui *et al.* [6] compressed the outer product feature using Multimodal Compact Bilinear (MCB) pooling, which can efficiently and expressively combine image and question features. Later, Multimodal Low-rank Bilinear (MLB) [17] pooling was proposed to reduce parameters. These feature aggregation methods obtain a better representation based on interaction of question and image. However, we observe that both two lines of research have been conducted independently so far. This could lead to sub-optimal results. Therefore, in this paper, we aim to combine both lines to facilitate the study of video visual question answering.

2.2 Video Visual Question Answering

Compared with image visual question answering, video visual question answering [2] task is a new updated VQA version, which is more challenging and it requires a model not only to understand the spatial cues of a video but also to exploit its complex temporal cues. To date, it has attracted a few attentions [9, 14, 29, 32, 36]. For instance, Jang *et al.* [14] presented a spatial-temporal model for the video visual question answering task. It captured the visual-textural association between an input video and question using two dual-layer LSTMs, and one for each input. This is a baseline benchmark approach, which is proposed together with the V-VQA dataset. In addition, a motion-appearance co-memory network [9] is proposed by extracting useful cues from both spatial and temporal memories to separately generate spatial and temporal attention maps. The two attended features are computed with weighted sum and then concatenated together as the final descriptor to infer the final answer. Both of them [9, 14] adopt the recurrent networks (i.e., LSTM or GRU) to encode the features, but two recurrent networks are difficult to train, since V-VQA dataset does not provide enough training samples. Therefore, in this paper we aim to propose a novel non-recurrent aggregation model to locate the evidence information to boost the performance of V-VQA.

3 IMAGE-QUESTION CO-ATTENTION REVISITING

Given an image $\mathbf{I}$ and a related question $\mathbf{Q}$, the target of image-level visual question answering (I-VQA) is to provide an answer a. An I-VQA problem can be formulated as:

$$a = f(\mathbf{I}, \mathbf{Q}, \theta) \tag{1}$$

where $f(.)$ denotes the I-VQA model and θ is the corresponding parameters. In addition, $f(.)$ is often trained by minimizing the cross entropy loss.

To efficiently and effectively solve I-VQA problem, various attention mechanisms have been proposed to meet the requirement of a joint comprehension of images and natural language questions [24, 27]. Among them, image-question co-attention has achieved a great success [24]. To generate an attention map, it firstly takes the question representation $\mathbf{Q} = \{\mathbf{q}_1 ... \mathbf{q}_m\} \in \mathbb{R}^{m \times d_q}$ and image representation $\mathbf{I} = \{\mathbf{x}_1 ... \mathbf{x}_n\} \in \mathbb{R}^{n \times d_x}$ as inputs to compute the affinity matrix $\mathbf{S}$:

$$\mathbf{S} = \mathbf{Q}\mathbf{W}_{wq}(\mathbf{I}\mathbf{W}_{wx})^{\mathsf{T}} = \mathbf{Q}\mathbf{W}_{wq}\mathbf{W}_{wx}^{\mathsf{T}}\mathbf{I}^{\mathsf{T}} \tag{2}$$

where $\mathbf{W}_{wq} \in \mathbb{R}^{d_q \times d}$ and $\mathbf{W}_{wx} \in \mathbb{R}^{d_x \times d}$ denote parameters to be learned; and $\mathbf{S} \in \mathbb{R}^{m \times n}$. Next, row-wise softmax and column-wise softmax operations are applied on $\mathbf{S}$ to generate an attention map on question words for each image region and an attention map on visual regions for each question word, respectively. Finally, an average operation is applied on the features to obtain question and video global attended features (i.e., $\mathbf{Q}_a$ and $\mathbf{I}_a$, respectively). Therefore, we formulate the single-path co-attention as below.

$$\mathbf{Q}_a, \mathbf{I}_a = \text{Co-Attention}(\mathbf{Q}, \mathbf{I}) \tag{3}$$

This single-path co-attention framework is a standard component of I-VQA model and has obtained impressive results. However, it is brittle to V-VQA for lacking the diversity and losing the video or question context information. Therefore, in this paper, we focus on modifying single-path co-attention by firstly introducing a multi-path pyramid co-attention and then providing an alternative temporal aggregation method, namely learnable aggregation. It automatically aggregates multi-path weighted representations (i.e., frame-level or word-level) to obtain a global descriptor (i.e., video-level or sentence-level, respectively) with rich context semantic information by imitating BoW quantization mechanism. Compared with single-path co-attention, multi-path co-attention has potential to capture different contextual information [24, 27], but there is no mechanism to guarantee each single-path attention indeed generates distinct attention maps. Therefore, we impose a diversity

constraints on multi-path co-attention to ensure it outputs diverse attention maps for capturing different contexts.

4 APPROACH

In this section, we formally introduce our proposed learnable aggregation network with diversity learning, namely LAD-Net. Specifically, the LAD-Net consists of two major components: 1) multi-path pyramid co-attention (MPC) with a diversity learning to enable the generated attention maps capturing context diversity; and 2) learnable aggregation modular with evidence-gating, which is a two-stage procedure. In the first stage, in parallel to aggregating adaptive-weighted frame-level representations and adaptive-weighted word-level representations. In the second stage, it combines aggregated features and further re-calibrates it with a set of learned evidence gates for capturing signals, which are most relevant to the answer.

4.1 Multi-path Pyramid Co-attention with Diversity Learning

In this subsection, we detail our Multi-path Pyramid Co-attention (MPC) module and show how we extend original co-attention mechanism with diversity learning. Fig.1 (a) shows the structure of our MPC modular.

Multi-path Pyramid Co-Attention. Given a video $\mathbf{V} = \{\mathbf{v}_1 ... \mathbf{v}_n\} \in \mathbb{R}^{n \times d_v}$ with n frames and a question $\mathbf{Q} = \{\mathbf{q}_1 ... \mathbf{q}_m\} \in \mathbb{R}^{m \times d_q}$ with m words, traditional co-attention mechanism requires an attention generator with single-path feature transformation, or a set of independent attention generators with multi-path linear projections under a single-scale condition. As a result, the generated attention maps with similar data distributions lack the property of diversity, which may lead to the suboptimal results for V-VQA analysis. To address this issue, our MPC module extends single-path co-attention mechanism into a multi-path one by firstly transforming the input (i.e., $\mathbf{V}$ or $\mathbf{Q}$) into H subspaces to form a feature pyramid structure to boost diversity. As a result, we have H types of similarity matrix between the $\mathbf{V}$ and $\mathbf{Q}$, each of which can be written as the Eq.4:

$$\mathbf{S}_i = \mathbf{Q}\mathbf{W}_{wqi}(\mathbf{V}\mathbf{W}_{wvi})^{\mathsf{T}} \tag{4}$$

where i denotes subspace index and $\mathbf{S}_i$ is the i-th affinity matrix. $\mathbf{W}_{wqi} \in \mathbb{R}^{d_q \times d_i}$ and $\mathbf{W}_{wvi} \in \mathbb{R}^{d_v \times d_i}$ are learned parameters. Next, two classes of attention maps can be derived from each similarity matrix through a normalization approach at two directions (i.e., vertical and horizontal), which are widely used for attention map generation in I-VQA tasks. For clarity, we define the attended visual features as $\mathbf{V}^i_{att}$ and attended question embeddings as $\mathbf{Q}^i_{att}$ from the i-th subspace. For each modality, the final representations can be formed by a summation operation over all attended features derived from all subspaces, as depicted in Eq.5.

$$\mathbf{Q}^f_{att} = \mathbf{Q} + \sum \mathbf{Q}^i_{att}, \ \mathbf{V}^f_{att} = \mathbf{V} + \sum \mathbf{V}^i_{att} \tag{5}$$

where $\mathbf{Q}^f_{att}$ and $\mathbf{V}^f_{att}$ indicate the final integrated sentence features and video features, respectively.

Diversity Learning. Feature scale pyramid structure provides a way to initially boost the difference between attended feature maps from the input aspect. Here, we further consolidate the diversity of attended features by introducing a diversity learning strategy from the output aspect. Specifically, for each modality, we introduce a diversity learning objective to penalize attended features with high similarity scores. The cosine similarity is adopted as metric to measure the distance and our diversity learning objective is defined as follows.

$$D(\mathbf{O}^i, \mathbf{O}^j) = \frac{1}{MN} \sum_{p=1}^{M} \sum_{q=1}^{N} \frac{\mathbf{o}^i_M {\mathbf{o}^j_N}^{\mathsf{T}}}{\left\|\mathbf{o}^i_M\right\| \left\|\mathbf{o}^j_N\right\|} \tag{6}$$

where $\mathbf{o}^i_M$ and $\mathbf{o}^j_N$ are two different feature vectors in feature matrix $\mathbf{O}^i = [\mathbf{o}^i_1, \mathbf{o}^i_2, ..., \mathbf{o}^i_M]$ and $\mathbf{O}^j = [\mathbf{o}^j_1, \mathbf{o}^j_2, ..., \mathbf{o}^j_N]$. And $i \neq j$ in this equation. Next, we impose this constraint on the attended visual representations and attended question embeddings to compute D_V and D_Q respectively, and the final diversity loss ℓ_D can be written as:

$$\begin{aligned} D_V &= \tfrac{1}{H^2} \sum_{i=1}^{H} \sum_{j=1}^{H} D(\mathbf{V}^i_{att}, \mathbf{V}^j_{att}) \\ D_Q &= \tfrac{1}{H^2} \sum_{i=1}^{H} \sum_{j=1}^{H} D(\mathbf{Q}^i_{att}, \mathbf{Q}^j_{att}) \\ \ell_D &= D_Q + D_V \end{aligned} \tag{7}$$

4.2 Feature Learnable Aggregating

Hand-crafted representations are conventionally used to describe video local statistics, which has been proved to be effective for video analysis [7, 28]. The establishment of hand-crafted feature often consists of two stages. The first stage is local descriptor extraction, where numerous local descriptors, such as dense trajectory and HOF descriptor, are originally computed from the RGB frames. The second stage is a feature encoding process, where a set of quantization based method are designed, such as Fisher vector (FV) and VLAD. Specifically, it first generates several visual words or feature cluster centers, and then assigns local descriptors to clusters according to their distance. However, the visual words are generated through an unsupervised approach, e.g., clustering, which may lead to the non-differentiable problem. To address above issue, we re-design a feature learable aggregating algorithm to make all operations differentiable to adaptively obtain a video-level descriptor to facilitate V-VQA task. The whole aggregation method is conducted in an end-to-end manner. For brevity, we denote it as a learnable aggregator.

Given a sequence of r feature descriptors $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_r\}$, where the number of feature descriptors r can be varied, and this module aims to transform $\mathbf{X}$ into a video-level representation with a fixed-length. To achieve this, a visual dictionary, as mentioned above, is firstly generated, which is followed by a quantization-based approach to further encode the video-level descriptor. Formally, we define the visual dictionary $\Phi(\mathbf{X})$ as a K types of temporal combination, where $\Phi(\mathbf{X}) = \{\mathbf{c}_k\}_{k=1}^{K}$ consisting of K visual words. It is generated through a nonlinear transformation $\Phi(\cdot)$. Specifically, to distribute the feature descriptor $\mathbf{x}_i$ to a visual words $\{\mathbf{c}_k\}$, we construct a temporal similarity representation by a set of differentiable operations, which can be optimized end-to-end. In particular, a temporal similarity encoding e_k is firstly computed through the Eq.8, shown below.

$$\begin{aligned} \mathbf{e}_k &= \mathbf{c}_{1k}\mathbf{x}_1 + \mathbf{c}_{2k}\mathbf{x}_2 +, ..., +\mathbf{c}_{ik}\mathbf{x}_i, ..., +\mathbf{c}_{rk}\mathbf{x}_r \\ \mathbf{E} &= \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_k, ..., \mathbf{e}_K\} \end{aligned} \tag{8}$$

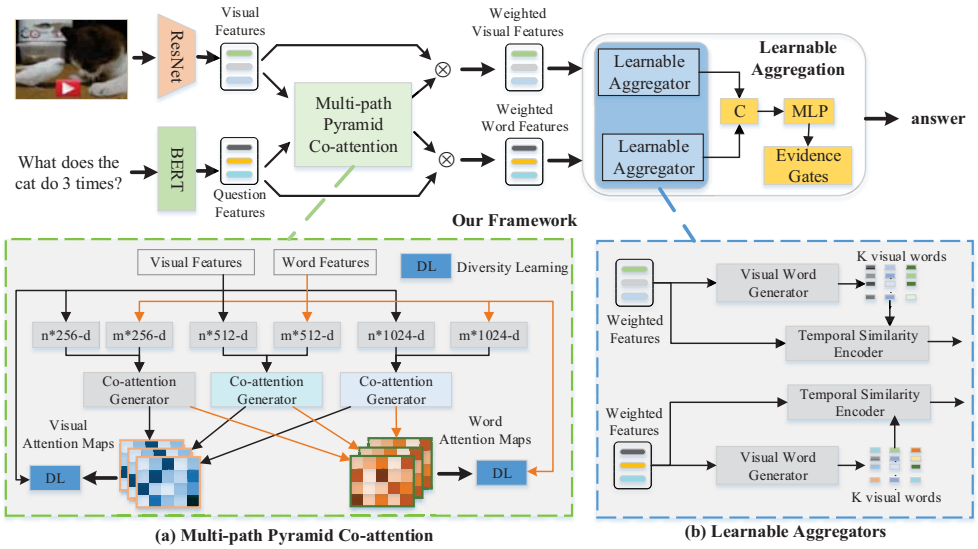

Figure 1: The overview of our proposed framework: Learnable Aggregation network with Diversity learning. There are two key components: multi-path pyramid co-attention with a diversity learning and learnable aggregation modular with evidence gating. C denotes concatenation operation and MLP denotes multi-layer perception.

Next, we combine K temporal similarity encodings through a concatenation operation, and the resulted similarity encoding is used for further V-VQA analysis. In addition, there are two types of feature descriptors for V-VQA analysis, including video local descriptors $\mathbf{V}^f_{att}$ and question local embeddings $\mathbf{Q}^f_{att}$. Both $\mathbf{V}^f_{att}$ and $\mathbf{Q}^f_{att}$ share the same temporal structure. Therefore we are able to make the feature aggregation process sharing straightforward. After the aggregation process, two final encoded features are obtained: $\mathbf{V}_E$ from the video local descriptors $\mathbf{V}^f_{att}$ and $\mathbf{Q}_E$ from the question representation $\mathbf{Q}^f_{att}$.

4.3 Evidence-Gating

A human being has potential to adaptively select the most relevant information to provide an answer, while gates are a way to optionally let information through. To mimic this behavior and to down-weight irrelevant features, in this paper we introduce an evidence-gating to capture long-term dependencies among descriptors and learn prior distribution of the answer probability by recalibrating the aggregated feature. Given a multi-modal representation $[\mathbf{V}_E, \mathbf{Q}_E]$, it conducts a multiplicative interaction between the given feature and a weight vector with values in between 0 and 1 to capture signals, which are most relevant to the answer. A value of zeros indicates "let nothing through", while a value of one denotes "let everything through". To achieve it, the aggregated feature $[\mathbf{V}_E, \mathbf{Q}_E]$, at first, is transformed into a weight gating vector by a nonlinear mapping. Next, feature signals are captured through a multiplicative interaction. We formalize this process as Eq.9.

$$A_E = \sigma(\mathbf{W}[\mathbf{V}_E, \mathbf{Q}_E] + \mathbf{b}) \circ [\mathbf{V}_E, \mathbf{Q}_E] \tag{9}$$

where σ is an element-wise sigmoid activation and $\circ$ is an element-wise multiplication. $\mathbf{A}_E$ denotes the answer embedding, which is used for further answer prediction.

4.4 Answer Inference

In this V-VQA task, there are three types of questions, including Multiple choice, Open-ended-number and Open-ended-word. Therefore, three answer inference processes are proposed to fulfill this task.

For the first type of question, we use a linear regressor to derive the final answer from A_E, depicted below:

$$\mathbf{P}_{mc} = \mathbf{W}^{\mathsf{T}}_{mc}\mathbf{A}_E \tag{10}$$

where $\mathbf{W}_{mc}$ are the parameters of the multi-choice regressor. Next, we apply a hinge-loss to optimize the regression results.

For the second type of question, we take a similar strategy by applying a special regressor on A_E:

$$\mathbf{P}_{opn} = [\mathbf{W}^{\mathsf{T}}_{opn}\mathbf{A}_E + \mathbf{b}_{opn}] \tag{11}$$

where $[\cdot]$ denotes a round operation, $\mathbf{W}_{opn}$ and $\mathbf{b}_{opn}$ are trainable parameters of the regressor. In this task, L2 loss is used to calculate the distance between the ground-truth and predicted values.

For the third type of question, we view it as a classification problem, and thus a linear classifier is applied to predict the answer. Specifically, we define a classifier that takes $\mathbf{A}_E$ as input, and then it selects an answer from the answer vocabulary according to their scores. The scores of the answer vocabulary can be written as:

$$\mathbf{P}_{opw} = softmax(\mathbf{W}^{\mathsf{T}}_{opw}\mathbf{A}_E + \mathbf{b}_{opw}) \tag{12}$$

where $\mathbf{W}_{opw}$ and $\mathbf{b}_{opw}$ denotes learned parameters of the classifier. We train this classifier by minimizing the cross entropy loss function.

For simplicity, we use ℓ to represent the answer prediction loss. Thus, the entire loss function can be written as:

$$\ell_{total} = \ell + \lambda\ell_D \tag{13}$$

ℓ indicates Hinge-loss, L2 loss or cross entropy loss. The choice is based on the condition of the question type. ℓ_D denotes diversity

loss introduced in diversity learning section. The λ is a hyper-parameter.

5 EXPERIMENTS

We validate our model on two public V-VQA datasets: TGIF-QA dataset and TVQA dataset. Firstly, we introduce the two datasets and experimental settings. Secondly, we conduct ablation studies to validate components effectiveness. Then, we compare our model's performance with state-of-the-art methods. Finally, qualitative analysis will also be introduced.

5.1 Datasets

5.1.1 TGIF-QA. TGIF-QA dataset [14] is a public V-VQA dataset and it contains four tasks: FrameQA, Action, Transition (Trans.) and Count. Specifically, **FrameQA** task is similar to image question answering task, where it asks questions about objects or their attributes. In particular, the question can be answered by observing one frame only. For instance: "what color is the woman's hair in video?". **Action** provides questions, related to actions appeared in a video. **Trans.** task places great emphasis on the transition relation between actions, such as "What does the man do before laugh?" **Count** task requires a model to predict a number. Both **FrameQA** and **Count** are open-ended tasks, but the later one expects a number ranging from zero to nine. Furthermore, both **Trans.** and **Action** are multi-choice tasks, where each question has five options. For **Count**, we adopt mean square error (MSE) to evaluate the model performance. For other three tasks, we employ the accuracy as the evaluation metric. Note that the lower the MSE score is, the better the performance.

5.1.2 TVQA[19]. TVQA is a large-scale V-VQA dataset based on six popular TV series. It consists of 152, 545 QA pairs and 21, 793 video clips, where it has 122, 039 training pairs, 15, 253 validation pairs and 15, 253 testing pairs. Specifically, it only contains a multi-choice task, and each question has five options. Moreover, each question is attached with a timestamp, which indicates the start and end time of the related clip. Therefore, we run two-experiments, one with full-length video clips (*w/o ts*), and another with localized videos using timestamps (*w ts*) to predict answers. In addition, it does not release the testing split. Therefore we utilize the validation set to evaluate all the models on this dataset.

5.2 Implementation Details

For TGIF-QA, 36 equal-spaced frames are selected from each video to represent the whole one, while for TVQA 64 equal-spaced frames are picked from each video. For each frame, we utilize the pre-trained ResNet-152 [13] network to extract a feature with 2, 048 dimension. For sentence pre-processing, we remove all the punctuations within a sentence and then convert it into lower case. Next, the NLTK toolbox is used to tokenize the sentence into a word list. For sentences with diverse lengths, we pad the short ones and truncate the long ones to make sure each sentence has 25 words. Finally, the pre-trained BERT [5] model is adopted to capture the long range dependencies for each question. As a result, each word is mapped into a 768-dimensional vector.

In the training stage, the Adamax is used as our optimizer. The batch size is set as 64 for TGIF-QA and 128 for TVQA. Moreover, the initial learning rate is set as 0.001. For better convergence, we decay the learning rate every five epochs and the decay rate is 0.25.

scale	FrameQA	Action	Trans	Count
Baseline [512]	56.89	69.17	77.42	4.37
[512,512,512]	57.41	69.04	78.34	4.34
[256,512,1024]	**57.53**	**69.92**	**78.37**	**4.32**

Table 1: Comparison of different variants of our method on the TGIF-QA dataset. Experiments are run to study the effect of MPC modular without imposing diversity learning.

λ	FrameQA	Action	Trans	Count
0	57.53	69.92	78.37	4.32
0.001	57.99	70.10	79.22	4.31
0.01	58.02	**70.43**	**79.54**	**4.29**
0.1	**58.25**	70.27	78.90	4.32
1	57.67	69.96	77.76	4.35

Table 2: The effect of diversity learning. The experiments are conducted on the TGIF-QA dataset.

5.3 Ablation Study

To further demonstrate the effectiveness of the proposed method and to deeply analyze the impact of each major component on the V-VQA task, we conduct an ablation study on the TGIF-QA dataset.

Effect of multi-path pyramid co-attention. To investigate the effect of multi-path pyramid co-attention, in particular the nature of multi-path and pyramid structures. We test the following models: (i) **Baseline**: LAD-Net with single-path co-attention ([512]); (ii) LAD-Net with multi-path co-attention ([512,512,512]) but without pyramid structure; and (iii) LAD-Net with multi-path pyramid co-attention ([256,512,1024]). Furthermore, all three variants do not use diversity learning. For fair comparison, all of the methods adopt the feature learnable aggregating without considering the evidence-gating. The number of visual words is set as 1. The experimental results are shown in Tab.1. It can be observed that in general the multi-path co-attention model without pyramid structure outperforms the single-path co-attention. The results demonstrate that multi-path has ability to encourage the feature diversity but its power is limited. Among them, our multi-path pyramid co-attention performs the best. More importantly, our multi-path pyramid co-attention based model obtains remarkable performance than the single-path co-attention based model, further demonstrating the superiority of the proposed pyramid structure. These results show the importance of multi-path pyramid co-attention.

Effect of diversity learning: λ. Here, we conduct an ablation to exploit the effect of diversity learning. In our method, the value of λ is used to balance the prediction loss and diversity loss during the training stage. To conduct the ablation study, we set λ with different values, ranging from 0, 0.001, 0.01, 0.1 to 1. Following previous ablation study, all models take the pyramid features as input (i.e.,256-D, 512-D, and 1024-D). In addition, the number of visual words is set as 1 and the evidence-gating is removed. The experimental results are shown in Tab.2. From the table it is clear when $\lambda = 0$, its performs worst. When $\lambda = 0.01$, our model performs the best on the Action, Trans. and Count tasks, and it achieves a

visual word number	FrameQA	Action	Trans	Count
0	56.70	63.46	76.80	4.40
1	58.02	70.43	79.54	4.29
2	58.18	70.84	79.41	4.33
4	58.09	71.33	79.72	**4.26**
8	**58.48**	**71.99**	**79.75**	4.30
16	58.24	70.49	79.59	4.32

Table 3: The effect of feature learning aggregation. The experiments are conducted on the TGIF-QA dataset.

Method	FrameQA	Action	Trans	Count
W/O gate	58.48	71.99	79.75	4.30
W/ gate	**58.19**	**71.99**	**80.70**	**4.24**

Table 4: Effect of evidence-gating

comparable result on the task of FrameQA compared with that of $\lambda = 0.01$. In our model, diversity learning brings a significant improvement (i.e., 58.25% vs 57.53 on FrameQA, 70.43% vs 69.92% on Action, 79.54% vs 78.37%, and 4.29 vs 4.32 on Count). The results clearly show the advantage of diversity learning and its power in encouraging capturing diversity information. In the following experiments, we set $\lambda = 0.01$.

Effect of learnable aggregation: K**.** Here, we evaluate the impact of the learnable aggregation and the effect of K value, where K indicates the number of visual words. All variant models remove the evidence-gating, which will be discussed later. When $K = 0$, we apply a traditional and simple pooling strategy, sum pooling, to aggregate the attended features. The rest of the variants adopt our proposed learnable aggregation mechanism, with different settings of K. The results of our analysis are shown in Tab.3. As expected, our learnable aggregation performs significantly better than sum pooling on all tasks (58.48% vs 56.7% on FrameQA, 71.99% vs 63.46% on Action, 79.75% vs 76.8% on Trans, and 4.26 vs 4.40 on Count, respectively), which proves that our learnable aggregation provides an efficient and effective way to describe the global-level feature for both question and video. More specifically, as the number of visual words increases, the performance of our model increases for ($0 \leq K \leq 8$), then decreases for $K = 16$. In particularly, when $K = 8$, the aggregation performs the best on the FrameQA, Action and Trans., while when $K = 4$ the Count task obtains the highest scores. Therefore, we set $K = 8$ for the following experiments.

Effect of evidence-gating. Two experiments are conducted: with and without evidence-gating. The experimental results are shown in Tab.4. Specifically, on the task of FrameQA, with evidence-gating the performances is slightly lower than without gating. This might be due to the nature of the FrameQA, where the answer can be inferred by observing only one frame. Moreover, from the Tab.4 we can conclude that evidence-gating plays a fundamental role for the tasks of Trans and Count.

5.4 Comparisons with the state of the arts

In this section, we present the results of our evaluation, comparing our approach with several state-of-the-art methods on TGIF-QA dataset and then on the recently released TVQA dataset.

Model	FrameQA	Action	Trans	Count
Random Chance	0.06	20.0	20.0	20.4
VIS+LSTM(aggr)	34.6	46.8	56.9	5.09
VIS+LSTM(avg)	35.0	48.8	34.8	4.80
VQA-MCB(aggr)	25.7	58.9	24.3	5.17
VQA-MCB(avg)	15.5	29.1	33.0	5.54
Yu et al.[33]	39.6	56.1	64.0	5.13
ST(R+C) [14]	48.2	60.1	65.7	4.38
ST-SP(R+C)	45.5	57.3	63.7	4.28
ST-SP-TP(R+C)	47.8	57.0	59.6	4.56
ST-TP(R+C)	49.3	60.8	67.1	4.40
Co-memory(R+F) [9]	51.5	68.2	74.3	**4.10**
Ours(R)	**58.2**	**72.0**	**80.7**	4.24

Table 5: TGIF-QA dataset: comparison with state of the art. R denotes Resnet-152 feature. C denotes C3D motion feature. F indicate Flow CNN features.

Method	W/O TS	W/ TS
Random Chance	20.00	20.00
Question Only	30.18	30.18
Multi-Stream[19]	41.55	43.22
Ours	**42.34**	**44.00**

Table 6: TVQA dataset. Comparison with the state-of-the-art methods. W/TS indicates taking the localized video clips as input, while W/O TS indicates taking the full-length video clips as input.

TGIF-QA. We firstly consider the TGIF-QA dataset and compare the performance of our approach with several traditional V-VQA methods, including Random Chance, *VIS+LSTM*, *VQA-MCB*, Yu et al.[33], *ST* [14] and *Co-memory* [9]. Note that both ST [14] and Co-memory [9] adopt two-stream features: spatial and temporal. R indicates spatial features extracted by ResNet, C is the C3D features, and F represents the optical flow features. Both C and F are temporal features. Compared with them, we only consider the spatial feature.

The results of this comparison are shown in Tab. 5. Specifically, *Random Chance* is conducted by randomly selecting an answer from the answer vocabulary. No surprisingly, it achieves the worst performance. Compared with *Random Chance*, two conventional I-VQA models *VIS+LSTM* and *VQA-MCB* perform better. In particular, *aggr.* is operated by averaging all video frame features to obtain a video global descriptor. *avg.* is conducted by firstly using every frame to infer an answer and then averaging all answers to generate the final answer. Moreover, Yu et al.[33] further improve the performance by utilizing concept words to assist V-VQA task and it achieves 39.6%, 56.1%, 64.0% and 5.13 on FrameQA, Action, Trans and Count, respectively. Compared with Yu et al.[33], two-stream based methods, including *ST* and *Co-memory*, perform significantly better by introducing spatial and temporal attentions and constructing a motion appearance co-memory network, respectively. However, applying two-stream feature has potential to improve the V-VQA performances, but it requires more computational resources than utilizing one-stream feature. On the Count task, compared with existing state-of-the art *Co-memory*, our model achieves a comparable

Figure 2: Qualitative comparison examples extracted from the TGIF-QA dataset. Answers are predicted by ST-TP(R+C) model and our LAD-Net(R). The correct answers are marked as green, while the wrong ones are marked with red.

results with less computational resources. On the task of FrameQA, Action and Trans, our model performs the best, with a clear and considerable performance gap, 6.7%, 3.8% and 6.4%, respectively.

TVQA dataset. We secondly consider the recently released TVQA dataset. In order to test the language bias of this dataset, we run the *Question Only* experiment by not considering visual information. The Multi-Stream model [19] is proposed as a benchmark method together with the TVQA dataset. With its released code, we reproduce its performance on the validation dataset since the test dataset is not released. The experimental results are shown in Tab.6. For both Multi-Stream model [19] and our method, models with localized video shots perform better than models with full-length video. This is because the localized video shots are specifically related to the question, while a full-length video contains loads of irrelevant video clips. Again, our approach outperforms all previous methods, reaching 42.34% with full-length video and 44% with localized video shots.

5.5 Qualitative Comparison

Fig. 2 shows several examples of our method and ST-TP (R+C) [14]. Note that, we cannot compare with Co-memory [9], because the code is not released. All the examples are collected from the TGIF-QA dataset. In Fig.2, the left column show four examples and each of them corresponds to the four V-VQA tasks in TGIF-QA. From the left column, we can observe that both ST-TP (R+C) and our method are able to provide correct answers. In addition, the right column of Fig.2 show some negatives examples, where ST-TP (R+C) fails to infer the right answer but our model is successful. From the top right example, it can be seen that our model has potential to correctly identify objects in videos as well as to understand the object's attributes. Where the "shirt" is and what the color of the "shirt" is. The middle two-examples on the right column show that our approach can precisely identify the specific body postures, e.g., "lift right hand" and "remove a bar from a game", but the ST-TP (R+C) provides inaccurate answers. This demonstrates the effectiveness of our proposed method. The bottom right shows another example, where both our method and ST-TP (R+C) conduct the count task. From it, we can see that our method can precisely count the number of repeated actions, while ST-TP (R+C) makes a mistake. To conclude, our model performs considerably better than the ST-TP (R+C), even though the later one takes more input information.

6 CONCLUSIONS

In this work, we present a novel network for video visual question answering named learnable aggregation network with diversity learning. Firstly, we extend single-path co-attention to multi-path pyramid co-attention with diversity learning, which explicitly encourage attended more diverse. Besides, we propose a new learnable aggregation methods with a set of evidence gates to replace traditional aggregation methods. Experimental results on TGIF-QA and TVQA datasets demonstrate that our LAD-Net outperforms state of the art significantly. Ablation studies also indicate the particular importance of our proposed model for V-VQA.

ACKNOWLEDGEMENTS

This work is supported by the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2014J063, No. ZYGX2016J085), the National Natural Science Foundation of China (Grant No. 61772116, No. 61872064, No. 61632007, No. 61602049) and Sichuan Science and Technology Program (No. 2018GZDZX0032, No. 2019ZDZX0008).

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*.

[3] Yi Bin, Yang Yang, Chaofan Tao, Zi Huang, Jingjing Li, and Heng Tao Shen. 2019. MR-NET: Exploiting Mutual Relation for Visual Relationship Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8110–8117.

[4] Yi Bin, Yang Yang, Jie Zhou, Zi Huang, and Heng Tao Shen. 2017. Adaptively attending to visual attributes and linguistic knowledge for captioning. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 1345–1353.

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

[6] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016).

[7] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. 2012. Recognizing activities with cluster-trees of tracklets. In *BMVC British Machine Vision Conference*. BMVA Press.

[8] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. 2017. VQS: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1811–1820.

[9] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[10] Lianli Gao, Xiangpeng Li, Jingkuan Song, and Heng Tao Shen. 2019. Hierarchical LSTMs with adaptive attention for visual captioning. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[11] Lianli Gao, Pengpeng Zeng, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. 2018. Examine before you answer: Multi-task learning with adaptive-attentions for multiple-choice VQA. In *Proceedings of the ACM international conference on Multimedia*. ACM.

[12] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. 2016. Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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

[15] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*.

[16] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Multimodal residual learning for visual qa. In *Advances in Neural Information Processing Systems*.

[17] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325* (2016).

[18] M. Lao, Y. Guo, H. Wang, and X. Zhang. 2018. Cross-Modal Multistep Fusion Network With Co-Attention for Visual Question Answering. *IEEE Access* 6 (2018).

[19] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*.

[20] Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, and Alexander G Hauptmann. 2018. Focal visual-text attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[21] Xiang Long, Chuang Gan, and Gerard de Melo. 2018. Video captioning with multi-faceted attention. *Transactions of the Association of Computational Linguistics* 6 (2018), 173–184.

[22] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering.

[23] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[24] Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[25] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[26] Damien Teney, Lingqiao Liu, and Anton van den Hengel. 2017. Graph-structured representations for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

[28] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2013. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision* 103, 1 (2013).

[29] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the ACM international conference on Multimedia*. ACM.

[30] C. Yang, M. Jiang, B. Jiang, W. Zhou, and K. Li. 2019. Co-Attention Network With Question Type for Visual Question Answering. *IEEE Access* 7 (2019).

[31] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[32] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. 2017. Video question answering via attribute-augmented attention network learning. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

[33] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[34] Z. Yu, J. Yu, J. Fan, and D. Tao. 2017. Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*.

[35] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*.

[36] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Video Question Answering via Hierarchical Spatio-Temporal Attention Networks.. In *International Joint Conferences on Artificial Intelligence*.

[37] Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. 2017. Structured attentions for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*.