

Relation-aware Hierarchical Attention Framework for Video Question Answering

Fangtao Li
Ting Bai
lift@bupt.edu.cn
baiting@bupt.edu.cn
Beijing University of Posts and
Telecommunications
Beijing, China

Chenyu Cao
Zihe Liu
www.caochenyu@bupt.edu.cn
ziheliu@bupt.edu.cn
Beijing University of Posts and
Telecommunications
Beijing, China

Chenghao Yan
Bin Wu[†]
yanch@bupt.edu.cn
wubin@bupt.edu.cn
Beijing University of Posts and
Telecommunications
Beijing, China

ABSTRACT

Video Question Answering (VideoQA) is a challenging video understanding task since it requires a deep understanding of both question and video. Previous studies mainly focus on extracting sophisticated visual and language embeddings, fusing them by delicate hand-crafted networks. However, the relevance of different frames, objects, and modalities to the question are varied along with the time, which is ignored in most of existing methods. Lacking understanding of the dynamic relationships and interactions among objects brings a great challenge to VideoQA task. To address this problem, we propose a novel Relation-aware Hierarchical Attention (RHA) framework to learn both the static and dynamic relations of the objects in videos. In particular, videos and questions are embedded by pre-trained models firstly to obtain the visual and textual features. Then a graph-based relation encoder is utilized to extract the static relationship between visual objects. To capture the dynamic changes of multimodal objects in different video frames, we consider the temporal, spatial, and semantic relations, and fuse the multimodal features by hierarchical attention mechanism to predict the answer. We conduct extensive experiments on a large scale VideoQA dataset, and the experimental results demonstrate that our RHA outperforms the state-of-the-art methods.

CCS CONCEPTS

• Information systems → Question answering; • Computing methodologies → Computer vision.

KEYWORDS

Video Question Answering, Hierarchical Attention, Multimodal Fusion, Relation Understanding

ACM Reference Format:

Fangtao Li, Ting Bai, Chenyu Cao, Ziheliu, Chenghao Yan, and Bin Wu. 2021. Relation-aware Hierarchical Attention Framework for Video Question Answering. In *Proceedings of the 2021 International Conference on Multimedia*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMR '21, August 21–24, 2021, Taipei, Taiwan

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8463-6/21/08...\$15.00

<https://doi.org/10.1145/3460426.3463635>

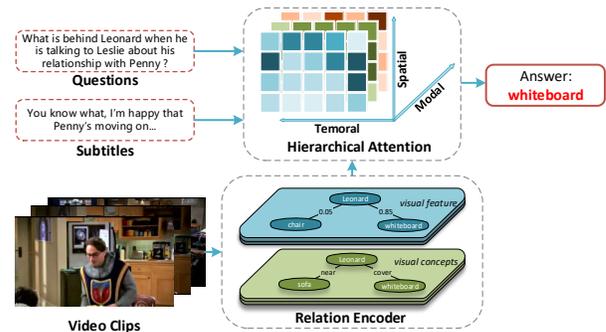


Figure 1: The illustration of Relation-aware Hierarchical Attention framework.

Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3460426.3463635>

1 INTRODUCTION

With the rapid development of deep learning in Computer Vision (CV) and Natural Language Processing (NLP). Video Question Answering (VideoQA), as an interdisciplinary area between language and vision, has been an active research topic in recent studies in video understanding. Given a video clip and a video-related question, VideoQA is expected to answer the question based on video grounding, reasoning, and translating. As a practical and comprehensive task, VideoQA can benefit other video understanding tasks such as video retrieval and storytelling.

The general VideoQA architecture consists of a detector, an embedding module, and a predictor. Usually, objects in the video are firstly detected and then fed into an embedding module. The embedding modules mainly utilize the Convolutional Neural Network (CNN) or pre-trained R-CNN [10] to obtain the visual features of objects in video, and sequential models such as Long Short-Term Memory (LSTM) to encode the texts in the questions. After aligning the question and video by attention mechanism or bilinear pooling [4, 16], the multimodal features of both video and question are jointly learned to obtain the answer to the question by the predictor.

However, most of these methods lack the ability of question localization in videos, which is vital for getting the proper representations to answer the question. In temporal, only some key frames

are closely related to the question. In spatial, only crucial areas and objects are useful to answering. Hence, both the static alignment in a single frame and the dynamic alignment among all frames in the video between the question and objects are significant. Besides, a video consists of multimodal features, i.e., visualizations, conceptions, and subtitles, and the intra-relations of different modalities should also be taken into consideration.

To address the above problems, we propose a novel Relation-aware Hierarchical Attention (RHA) framework to learn both the static and dynamic relations of the visual objects in videos. For the static relations of visual objects in a single frame, we construct a graph to build the spatial and semantic relationships among objects. Then we measure the modality importance by taking their relevance to the question into consideration. For example, in Figure 1, given a question as "What is behind Leonard when he is talking to Leslie about his relationship with Penny?", previous methods [19, 20] incorporate all information into modeling, including the relationship between "Leslie" and "Penny". While such relationship is usefulness and what we focused is the object behind "Leonard", which can be learned from the visual information of video. The importance of different modalities should be considered at the static feature fusion stage. As for learning the dynamic changes of visual objects in different video frames, we characterize the temporal, spatial, and semantic relations. We adopt attention mechanism to learn from each dynamic relations. Our proposed RHA model learn the static relation among objects, as well as the dynamic relations in time, space and modality dimensions. On the temporal dimension, the hierarchical attention locates essential time stamps. On the spatial dimension, it identifies important areas, and in the modality dimension, the importance of different modality features can also be learned.

The architecture of our framework is shown in Figure 2. A pre-trained BERT [7] is applied to extract the features of questions, candidate answers, and subtitles. For each frame in the video, a pre-trained Faster R-CNN [30] is applied for object detection and embedding. All the objects are then fed into a customized graph attention networks (GATs) [35] with shared weights as a relation encoder to learn the relationships between objects. The updated embeddings will be injected into a multimodal attention module, along with subtitle and question embeddings. Finally, we use a Multi-Layer-Perceptron (MLP) as an answer prediction module to generate the correct answer index. We evaluate our framework on TVQA+ [20] dataset and conduct experiments on temporal grounding, showing the the ability of our model in pointing out relevant temporal span.

To summarize, our main contributions are:

- We propose a novel RHA framework which adopts a question-guided hierarchical attention module to capture both static and dynamic relations of multimodal objects.
- We introduce a graph-based relation encoder to model the static relationships between visual objects in videos, and three dynamic relations: temporal relation, spatial relations, and semantic relations are considered to predict the answer.
- We verify the proposed framework RHA on a large scale TVQA+ dataset. Experimental results show that our model outperforms other state-of-the-art methods.

2 RELATED WORKS

2.1 Video Question Answering

As a typical multimodal task, VideoQA requires thorough visual and textual understanding. In recent years, some more restricted sub-tasks have also been proposed to enhance the interpretability, such as Knowledge-based VideoQA [9] and Spatio-temporal grounding VideoQA [20]. Nevertheless, the VideoQA framework generally consists of a video encoder, a question encoder, an embedding alignment module, and a predictor. With the domination of deep learning, early methods [8, 41] use CNN to extract features at frame level. However, most of the CNN-based extractors cannot seize temporal information in the video. [15, 38] apply LSTM as a substitute to model temporal context. As for question encoder, Glove [27] and LSTM are generally applied. As the kernel of the whole VideoQA framework, embedding alignment module could be quite sophisticated. Early works [41] rely on hand-craft CNN architecture to further embed video and question. Inspired by [34], [13] utilizes attention-based methods to focus on relevant video clips. Attention mechanism also enhances the interpretability of these models, since it works in a simple but intuitive way. Another line of research focuses on graph-based learning. With the popularity of Graph Neural Networks (GNNs) [17], some works focus on modeling video from the perspective of topology and embed video by GCNs. [15] constructs a heterogeneous graph which regards both words in the question and frames in the video clip as nodes and aligns them by GCNs [17].

The most related work is STAGE model [20], which proposes a VideoQA framework with spatial and temporal grounding. Compared to STAGE, our RHA contains some critical distinctions: (1) STAGE designs an elaborate convolutional kernel for encoding, which is highly customized and cannot capture any relation attributes. In comparison, RHA utilizes a graph-based encoder to learn both objects embedding and their relations; (2) STAGE fuses multimodal features by a heuristic method, while we propose an interpretable hierarchical attention module to fuse multimodal features adaptively.

2.2 Relation Understanding

Relation understanding contains many sub-tasks, in which relation extraction and relation reasoning are two of the most important task. Relation extraction aims to detect relationships between given objects. Prior works recognize relations between objects by co-occurrence [32] and position [37]. This kind of methods is generally based on statistics and can only identify spatial relations (such as *behind*, *below*, and *cover*). Another line of work is semantic relation extraction, which generally requires a deeper understanding of video. [12] proposes a neural network to extract semantic relations on a single image. [6] designs a novel Two-Stage Model to extract the social relationships between characters. Undeniably, as a stepping-stone of video understanding, most relation extraction methods are designed as a task-specific module. [5] builds a trainable cell named MuRel to model pair-wise object relationships, while [12] discovers implicit relation by adopting an attention-based object relation detector.

In contrast, relation reasoning aims to represent objects based on their relations. [18] proposes a Relation Network as a general

solution of relation reasoning in an unsupervised manner. [33] designs a novel Interaction Canonical Correlation Network for cross-modal relation reasoning. With the explosive development of GNNs, recent research suggests that objects and their relationship can be represented by nodes and edges in the graph. [31] builds a fully-connected graph for a given image, discovering interactions with a self-attention mechanism.

Our work is also inspired by [22], which builds a relation-aware graph network to discover explicit and implicit relations in the image. We make some targeted improvements for VideoQA task. The first difference is that we explore objects relations in a video rather than a single image. Second, we use both visual features and visual concepts for relation encoding, which enhances the interpretability and robustness of our framework.

2.3 Multimodal Fusion

As one of the original topics in multimodal learning, multimodal fusion aims at gaining joint representation of two or more modalities. Some studies [19, 20, 36] applied vector operations between single-modal features, including vector concatenation, element-wise multiplication, and element-wise addition. Such researches are referred to as early fusion since they fuse multimodal information before the decision. In contrast, late fusion uses unimodal decision results and merges the results by a fusion mechanism such as averaging or voting [2]. With the development of deep learning, some works [11, 29] have proposed to use trainable model to enhance the performance of multimodal fusion. As for unsupervised learning, [26] presents a multimodal autoencoder to fuse features adaptively without any supervision. Network Architecture Search (NAS) is also applied in multimodal fusion [28]. More recently, bilinear pooling is a practical pathway of fusion [16]. [39] calculates the outer-product of video, acoustic and textual features to gain the multimodal joint representation. [25] proposes a low-rank method to build tensor networks to reduce computational complexity caused by tensor outer-product. With the domination of bilinear pooling, attention mechanism [40] is also regarded as an effective method to enhance the interaction between modalities and avoid redundancy. [24] proposes a Multimodal Attention (MMA) module, which reweights modalities through the Gram matrix, and [21] optimizes MMA by design a deeper attention convolutional layer.

In terms of multimodal fusion, [15] and [20] are the most similar works. However, [20] simply assumes that each modality has the same weight, ignoring the difference between questions. Co-attention proposed by [15] lacks the ability to analyze fine-grained object-level importance. To the best of our knowledge, we are the first to measure the importance of different time, objects, and modalities at the same time in this task.

3 METHODS

3.1 The General Framework of RHA

Our work mainly focuses on multiple-choice VideoQA task, which needs to choose the right answer in a set of candidate answers. Given (1) a question q ; (2) 5 candidate answers $\{a_k|k = 1, \dots, 5\}$; (3) a video clip that consists of keyframes $\{F_t|t = 1, \dots, T\}$ (4) the subtitles of corresponding video $\{P_t|t = 1, \dots, T\}$, our goal is to

predict the index of right answer \hat{a} :

$$\hat{a} = \arg \max_{a \in a_k} p(a|q, P, F). \quad (1)$$

The RHA framework consists of four modules. As shown in Figure 2, all visual and text inputs are first embedded by the input encoder. Then a relation encoder is utilized to discover relationships between visual objects. All these representations are projected into the same dimension after relation modeling. To learn the relevance between video and question, we apply the hierarchical attention module to reweight and fuse representations. Finally, we predict the right answer and its relevance video clip by an answer predictor.

3.2 Input Encoder

For a given video containing keyframes $\{F_t\}$, we use Faster R-CNN to extract the features of each detected object, followed by PCA to downsize the dimension of object proposals into a low dimension representation $\mathbf{O} = \{\mathbf{o}_t^i \in \mathbb{R}^{d_o}, i \leq N_o, t \leq T\}$, where \mathbf{o}_t^i refers to the i -th objects of t -th frame. The bounding-boxes $\mathbf{B} = \{\mathbf{b}_t^i \in \mathbb{R}^4, i \leq N_o, t \leq T\}$ are represented by the coordinate of the top-left and bottom-right point of the bounding-boxes. The label of objects is embedded to a vector $\mathbf{L} = \{\mathbf{l}_t^i \in \mathbb{R}^{d_l}, i \leq N_o, t \leq T\}$ by Glove [27], where d_l is set to 300. For subtitles $\{P_t\}$, we use a pre-trained BERT to extract embeddings $\mathbf{S} = \{\mathbf{s}_t \in \mathbb{R}^{L_s \times d_s}, t \leq T\}$. The candidate answers are first concatenated with the question to compose a qa-hypothesis $\{\mathbf{Q}_k\}$. The same pre-trained BERT is used to extract embeddings which are denoted as $\mathbf{H} = \{\mathbf{h}_k \in \mathbb{R}^{L_q \times d_q}, k \leq 5\}$, where d_s and d_q denote the dimension of word embedding, while L_s and L_q refer to the length of subtitles and hypothesis, respectively.

3.3 Relation Encoder

Given a video clip, we discover explicit (spatial) and implicit (semantic) relationships between different objects in the video. Understanding these relationships is the key to understanding the video underlying information. To capture these relationship attributes in the video, we build an encoder based on GATs [35] to capture the relationship between objects, processing the spatial relationship and semantic relationship, respectively.

3.3.1 Spatial Graph Construction. Spatial relation in the video refers to the position relationship between objects. The spatial relation may be varied with the variance of the position of objects and movement of camera. The frame-level spatial relation is denoted as a graph $\mathbf{G}_{spa}^t = (\mathbf{V}_{spa}^t, \mathbf{E}_{spa}^t)$, and the video-level spatial relation can be represented as the concatenation of \mathbf{G}^t . We denote $\mathbf{v}_{spa}^{t,i}$ as the embedding of i -th node in frame t , and $\mathbf{e}_{spa}^{t,i,j}$ refers to the spatial relation between object i and j . Inspired by [37], the spatial relations between objects are classified into 11 categories based on bounding-boxes \mathbf{B}_t . Each category refers to a kind of spatial relations such as *cover*, *in*, and *near*. Note that we use the label of objects $\{\mathbf{l}_t \in \mathbb{R}^{N_o \times d_l}\}$ as node embeddings, rather than the visual features of objects. We argue that alignment from words in question to objects in video is the key for understanding spatial relation, while the visual features such as shape, color is irrelevant. And it is more difficult to align visual features to text features than to align text features, which affects the accuracy of answering questions.

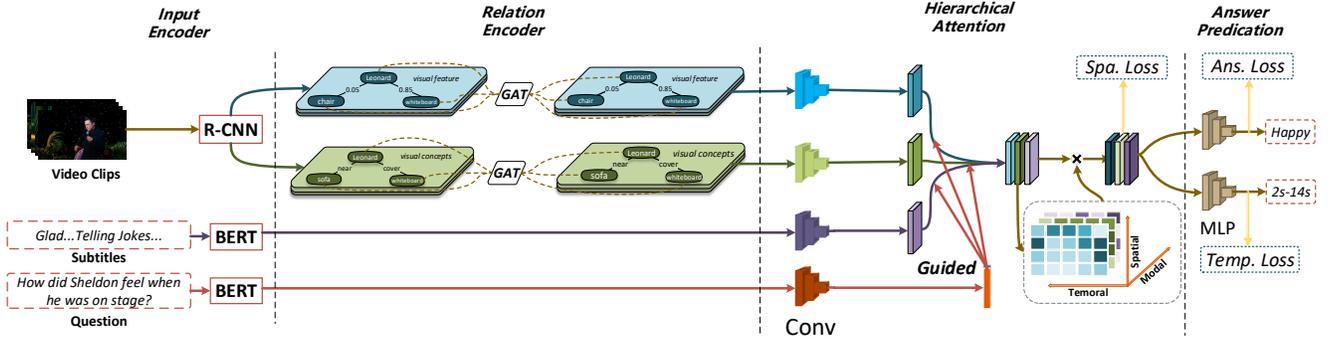


Figure 2: An overview of Relation-aware Hierarchical Attention framework

3.3.2 *Spatial Graph Update.* For each keyframe in the video, we get a spatial graph \mathbf{G}_{spa}^t using the methods above. Then we utilize a customized GAT [35] to update the node embedding. We adopt multi-head attention to generalize the learning progress of GAT. All output features of heads are concatenated. In order to encode spatial relation into GAT, the bias is set to be independent for different spatial relations, and the projection matrix represents the direction of relation (*from objects or to objects*):

$$\mathbf{l}_i = \|\|_{m=1}^M \sigma \left(\sum_{j \in N_i} \alpha_{ij}^m \cdot \mathbf{W}_{spa}^m \mathbf{l}_j \right), \quad (2)$$

$$\alpha_{ij} = \frac{\exp((\mathbf{U}_{spa} \mathbf{l}_i)^\top \cdot \mathbf{V}_{spa}^{dir(i,j)} \mathbf{l}_j + \mathbf{b}_{spa}^{lab(i,j)})}{\sum_{j \in N_i} \exp((\mathbf{U}_{spa} \mathbf{l}_i)^\top \cdot \mathbf{V}_{spa}^{dir(i,j)} \mathbf{l}_j + \mathbf{b}_{spa}^{lab(i,j)})}, \quad (3)$$

where $\alpha \in \mathbb{R}^{N_o \times N_o}$ is the attention weight, $\mathbf{W}_{spa} \in \mathbb{R}^{M \times d_h \times d_i}$, $\mathbf{U}_{spa} \in \mathbb{R}^{d_h \times d_i}$, $\mathbf{V}_{spa}^{dir} \in \mathbb{R}^{d_h \times d_i}$ is the projection matrix, $\mathbf{b}_{spa}^{lab} \in \mathbb{R}^{d_h}$ is the bias, and d_h is the dimension of hidden layer. M refers to the number of heads of graph attention, which is set to 15 in our implementation. Residual connection is also involved to avoid over smoothing in GAT. Updated frame-level features \mathbf{v}_{spa}^t can be represented as the concatenation of node embeddings.

For different frames in the video, we share the parameters of GAT. One advantage of weight sharing is that it can reduce the number of parameters significantly. On the other hand, this relation encoder can be more robust to the temporal changing of spatial relationship and the changing of the number of objects.

3.3.3 *Semantic Graph Construction.* Semantic relation in the video refers to the relationship that can not be inferred only by position and visual information. Similar to spatial relation, semantic relation between objects may be changed along with the progress of plots. Given visual features $\{\mathbf{o}_t \in \mathbb{R}^{N_o \times d_i}\}$, the frame-level semantic relation between objects i and j in frame t is defined as below.

$$\mathbf{e}_{sem}^{t,i,j} = \frac{\exp(\mathbf{W}_s [\mathbf{o}_t^i; \mathbf{o}_t^j])}{\sum_{j \in N_i} \exp(\mathbf{W}_s [\mathbf{o}_t^i; \mathbf{o}_t^j])}, \quad (4)$$

where $\mathbf{W}_s \in \mathbb{R}^{2d_o \times 1}$ is trainable parameters. Treating each object $\{\mathbf{o}_t^i\}$ as a node, we build the semantic graph $\mathbf{G}_{sem}^t = (\mathbf{V}_{sem}^t, \mathbf{E}_{sem}^t)$ where $\mathbf{e}_{sem}^{t,i,j}$ refers to the semantic relation between object i and j . Note that in this stage, we use the region features $\{\mathbf{O}_t^i\}$ as the

embedding of nodes since the semantic relation between objects is hard to infer only by their categories. More visual information is involved as supplementary of relation understanding.

3.3.4 *Semantic Graph Update.* Following the previous works [22], we utilize graph attention mechanism to update the node embedding after building semantic graph \mathbf{G}_{sem}^t for each keyframe in the video. Multi-head attention is also applied in GAT. All output features of attention heads are concatenated to obtain the updated node embeddings:

$$\mathbf{o}_i = \|\|_{m=1}^M \sigma \left(\sum_{j \in N_i} \beta_{ij}^m \cdot \mathbf{W}_{sem}^m \mathbf{o}_j \right), \quad (5)$$

$$\beta_{ij} = \frac{\exp((\mathbf{U}_{sem} \mathbf{o}_i)^\top \cdot \mathbf{V}_{sem} \mathbf{o}_j)}{\sum_{j \in N_i} \exp((\mathbf{U}_{sem} \mathbf{o}_i)^\top \cdot \mathbf{V}_{sem} \mathbf{o}_j)}, \quad (6)$$

where $\beta_{ij} \in \mathbb{R}^{N_o \times N_o}$ is the attention weight, $\mathbf{W}_{sem} \in \mathbb{R}^{d_h \times d_o}$, $\mathbf{U}_{sem} \in \mathbb{R}^{d_h \times d_o}$, $\mathbf{V}_{sem} \in \mathbb{R}^{d_h \times d_o}$ is the projection matrix, and m is the number of heads of graph attention, which is set to 15 in our implementation. For semantic graph, we do not encode relation direction, since the semantic graph is fully-connected and symmetric. Residual connection is also involved to avoid over smoothing in GAT. For different frames in the video, we also share the parameters of GAT for semantic graph update. Similar to spatial relation, each frame can be represented as the concatenation of node embeddings $\mathbf{v}^{sem} = \|\|_{t=1}^T \mathbf{o}_t^{sem}$.

3.4 Hierarchical Attention Module

As stated before, a video consists of visual feature, visual concepts, and subtitles. In order to measure the relevance to the question, the visual and textual features will be downsized into the same dimension first. Following the previous works [19, 20], we adopt a fully-connected layer with residual connection to build downsize encoding block for object features $\{\mathbf{o}_t\}$, layer normalization [1] is also involved:

$$\mathbf{o}_t^i = \text{Layernorm}(\text{ReLU}(\mathbf{W}_d \mathbf{o}_t^i) + \mathbf{o}_t^i). \quad (7)$$

Similar procedure is applied on visual concepts $\{\mathbf{l}_t\}$, subtitles $\{\mathbf{s}_t\}$, and qa-hypothesis $\{\mathbf{h}_k\}$. Then the question is grounded in temporal, spatial, and modal, generating an updated video features with the question encoded. Note that in our model, we distinguish the visual

features and visual concepts, since although they all represent visual information, the description methods are different.

3.4.1 Spatial and Temporal Attention. At the first stage, we locate the qa-hypothesis spatially and temporally. Given the encoded hypothesis $\{\mathbf{h}_k\}$ and encoded visual features $\{\mathbf{o}_t\}$, the attention scores of visual features $M_{k,t} \in \mathbb{R}^{L_q \times N_o}$ and visual representation $\mathbf{o}_{t,att} \in \mathbb{R}^{L_q \times d_h}$ is computed as:

$$M_{k,t} = \text{softmax}(\mathbf{h}_k \cdot \mathbf{o}_t^\top), \quad (8)$$

$$\mathbf{o}_{t,att} = M_{k,t} \mathbf{o}_t. \quad (9)$$

The question-guided attention scores actually represent the relevance between the qa-hypothesis and the objects in different frames so that the predictor can focus on important visual objects. The same process is performed on visual concepts and subtitles to compute representation $\mathbf{l}_{t,att} \in \mathbb{R}^{L_q \times d_h}$ and $\mathbf{s}_{t,att} \in \mathbb{R}^{L_q \times d_h}$.

3.4.2 Multimodal Attention. In the ideal case, all modalities are as same important for answering the question. However, the questions are often changeable and comprehensive, resulting in the redundancy of video. Different questions may only focus on the information of a certain modal in the video. In this case, we need to focus on some modalities to answer the question better. Motivated by this, we design a multimodal attention mechanism at the second stage of the hierarchical attention. First, we concatenate $\mathbf{o}_{t,att}$, $\mathbf{l}_{t,att}$, and $\mathbf{s}_{t,att}$ as multimodal features $X_t \in \mathbb{R}^{L_q \times d_h \times 3}$. The feature X_t is transformed into low dimension space by trainable parameters $\mathbf{W}_F \in \mathbb{R}^{d_h \times d_h}$. Then we adopt the Gram matrix of X_t to capture modal correlation by multiplying X_t with its transpose. The weight of each modal is yielded through the convolution layer and a softmax activation:

$$Y_i = \sum_{j=1}^3 X_j \cdot \frac{\exp((\mathbf{W}_F X)^T (\mathbf{W}_F X))_{j,i}}{\sum_i \exp((\mathbf{W}_F X)^T (\mathbf{W}_F X))_{j,i}}. \quad (10)$$

Note that at the second stage, we apply a self-attention mechanism for multimodal fusion rather than question guided attention, since the question has been encoded into the multimodal features in the first stage so that the second stage mainly focuses on the importance of the modalities themselves.

3.5 Answer Predictor

The last module is an answer and localization predictor. The RHA is required to predict the answer based on multimodal features $Y \in \mathbb{R}^{T \times L_q \times d_h}$. Minimum time spans related to the question are also predicted based on the joint representation. We first apply a convolutional layer with max-pooling layer to obtain the output $A \in \mathbb{R}^{T \times d_h}$. For temporal prediction, $A \in \mathbb{R}^{T \times d_h}$ is sent into two linear layers with softmax to produce start probabilities $\mathbf{p}_k^1 \in \mathbb{R}^T$ and end probabilities $\mathbf{p}_k^2 \in \mathbb{R}^T$ for each frame k . For answer prediction, an additional linear layer is first utilized to further encode video-text representation A_k . Then a global representation $G_k^g \in \mathbb{R}^{d_h}$ is generated by max-pooling across all the time steps. Taking temporal prediction into consideration, we generate temporal proposals using dynamic programming. For each proposal, we generate a local representation $G_k^l \in \mathbb{R}^{d_h}$ by max-pooling A_k , concatenating with G_k^g to obtain $G \in \mathbb{R}^{2d_h \times 5}$. Finally, concatenated features G is sent to

a softmax function to generated the answer scores $\mathbf{P}^{ans} \in \mathbb{R}^5$. Note that we mainly implemented this module based on STAGE [20], and we only make some essential changes to fit the input of visual features, visual concepts, and subtitles for a fair comparison.

3.6 Training Loss

As we need to answer the question with temporal grounding, the RHA framework is trained with supervision from ground truth (GT) bounding boxes, GT time proposal, and GT answer. For spatial supervision, we define a box as positive for spatial prediction if it has an IoU larger than 0.5 with the GT box. The attention weights of positive objects should be higher than negative ones. So LSE [23] loss function is applied since it is easier to optimize [3]:

$$\text{loss}_{Spa} = \frac{1}{N} \sum_{i=1}^N \sum_{r_p \in \Omega_p, r_n \in \Omega_n} \log(1 + \exp(M_{i,t,r_n} - M_{i,t,r_p})), \quad (11)$$

where $M_{i,t,r}$ is the r -th element of the matching scores $M_{i,t}$. For temporal supervision, cross-entropy loss is applied to measure the probabilities of start and end time:

$$\text{loss}_{Temp} = \frac{1}{2N} \sum_{i=1}^N (y_i^{st} \log \mathbf{P}^1 + y_i^{ed} \log \mathbf{P}^2), \quad (12)$$

where y^{st} and y^{ed} is ground truth start and end indices. Similarly, given answer probabilities \mathbf{P}^{ans} , we also apply a cross-entropy loss as answer prediction loss:

$$\text{Loss}_{Ans} = \frac{1}{N} \sum_{i=1}^N y_i^{ans} \log \mathbf{P}^{ans}, \quad (13)$$

where y^{ans} is the index of ground truth answer. Finally, all three losses are summed up as:

$$\text{Loss} = \omega_{Ans} \text{Loss}_{Ans} + \omega_{Spa} \text{loss}_{Spa} + \omega_{Temp} \text{loss}_{Temp}, \quad (14)$$

where ω_{Ans} , ω_{Spa} and ω_{Temp} are the weights of different loss. In our case, ω_{Ans} is set to 1, while ω_{Spa} and ω_{Temp} are both 0.5.

4 EXPERIMENT

4.1 Dataset

TVQA+ [20] is a large scale multiple-choice VideoQA dataset with spatio-temporal grounding. All data is collected from "The Big Bang Theory". The TVQA+ dataset is the augmented version of TVQA dataset [19], with 21.8K 60-90 seconds long video clips and 29.4K multiple-choice questions grounded in both the temporal and the spatial domains. Each question is followed by 5 candidate answers, in which only one of them is the right answer. For spatio-temporal grounding, there are 310.8K bounding boxes linked with referred objects and people, spanning across 2.5K categories. All questions are composed of a question part ("Where/Why/What") and a localization part ("when/before/after").

4.2 Implementation Details

In our implementation, Layer Normalization [1] and Dropout is applied between every two full-connected layers. As for hyper-parameters, dimension of object features d_o is set to 300, while textual feature dimension d_s and d_q are 768. The dimension of hidden layers d_h is set to 128, while \hat{d}_h in hierarchical attention is

Table 1: Comparison with state-of-the-art methods on TVQA+ test set.

Model	Acc	Temp. mIoU	ASA
ST-VQA [14]	48.28	-	-
two-stream [19]	68.13	-	-
STAGE-video [20]	52.75	10.90	2.76
STAGE-sub [20]	67.99	30.16	20.13
STAGE [20]	72.14	30.68	20.99
FAMF(Ours)	74.34	31.53	21.77

32. The dropout rate is set to 0.1. At the training stage, batch size is set to 16 to balance the training speed and memory cost, while at the inference stage, we set batch size to 1. Adam optimizer is applied for training, the initial learning rate is set to 0.001 and will be decayed by 0.1 for every 10 epochs.

4.3 Evaluation Metrics

In this work, we use three evaluation metrics to measure performance. First, we use classification accuracy to measure QA performance. We also consider temporal localization performance, which is evaluated by temporal mean Intersection-over-Union (Temp. mIoU). Finally, we evaluate QA accuracy and temporal localization jointly by Answer-Span joint Accuracy (ASA) [20]. For this metric, we regard a prediction as positive only if the predicted temporal localized span has an IoU ≥ 0.5 with the ground-truth span and the answer is correctly predicted at the same time.

4.4 Experimental Results

We mainly compare with state-of-the-art methods on TVQA+ [20]. ST-VQA [14] is designed for question answering on short videos or GIFs. Two-stream model [19] is a method to predict the answer based on videos and subtitles, respectively. The two-stream model is retrained based on official code and TVQA+ data since the original two-stream uses Glove rather than BERT [7] to embed subtitles, which may result in worse performance. STAGE [20] also encodes frame-wise regional visual representations and neural language representations, which also implements temporal localization and spatial grounding. STAGE-sub means only the subtitle branch is activated for question answering, while STAGE-vid means only video features are taken as input. STAGE is also retrained with the official code. All models are trained on the train set and tested on the test set. The experimental results are shown in Table 1. We select STAGE as our main baseline method. Experimental results show that our RHA model outperforms other methods by a significant margin(74.34 vs. 72.14).

We also report the performance of our model on different question types. As shown in Table 2, we classify the questions according to the first word and select the most frequent five classes. Surprisingly, our RHA performs quite well for questions started with "Why", which are generally regarded as hard cases in VideoQA. This phenomenon indicates that RHA captures implicit information for reasoning and answering. However, for questions started with "Who" or "How", RHA still leaves a lot to be desired, which may be the aftermath of the absence of acoustic analysis.

Table 2: Evaluation by question type on TVQA+ valid set

Model	Acc	Temp. mIoU	ASA
What	72.23	31.68	20.81
Why	81.60	31.81	21.75
Where	73.63	31.38	23.97
Who	69.57	26.96	15.21
How	69.23	30.68	20.99

Table 3: Experimental results of using different modalities on valid set.

model	Acc	Temp. mIoU	ASA
baseline	70.16	30.05	19.52
RHA-re	71.72	30.43	19.98
RHA-vc	71.89	30.76	20.11
RHA-vs	72.08	30.88	20.15
RHA-ha	72.45	30.93	20.31
RHA-full	72.58	31.30	20.64

4.5 Ablation Study

To measure the effectiveness of our proposed RHA framework, we drop the relation encoder and hierarchical attention, respectively. We report the ablation results on the TVQA+ valid set, which are shown in Table 3. As stated before, we select STAGE as our main baseline method. RHA-re means we only applied relation encoder and no attention fusion is activated. RHA-vc and RHA-vs mean we use visual concepts and visual features for relation encoder, respectively. RHA-ha means only the hierarchical attention module is applied, and the relation encoder is dropped. RHA-full is the full version of RHA framework.

4.5.1 The effect of Relation Encoder. First, we analyze the effect of the relation encoder. From lines 1 and 2 in Table 3, we can find the relation encoder brings an accuracy improvement of about 1% compared to the baseline model. The results show that introducing relation modeling, which can discover the potential relationships between objects, helps better video understanding. For semantic relation, commonsense (such as traffic rules, physics rules, etc.) is less considered in the previous VideoQA models, and this knowledge can be described in terms of the internal connections between objects. For example, given a question like "Why Sheldon stop the car", we need to build a connection between traffic lights and vehicles, answering the question based on the fact that the traffic light is red. These kinds of questions are exactly what the previous models difficult to learn by convolutional layer, but can be learned through our implicit relation module.

On the other hand, some questions may refer to the position of objects. Similarly, taking "What is near Sheldon" as a toy example, previous models may use large amounts of training data for pattern recognition, answering the question based on objects that often appear with Sheldon. However, these methods do not really understand the concept of "near", but only choose the candidate answer based on the principle that "what you see is what you answer". Our spatial relation module encodes the position of objects

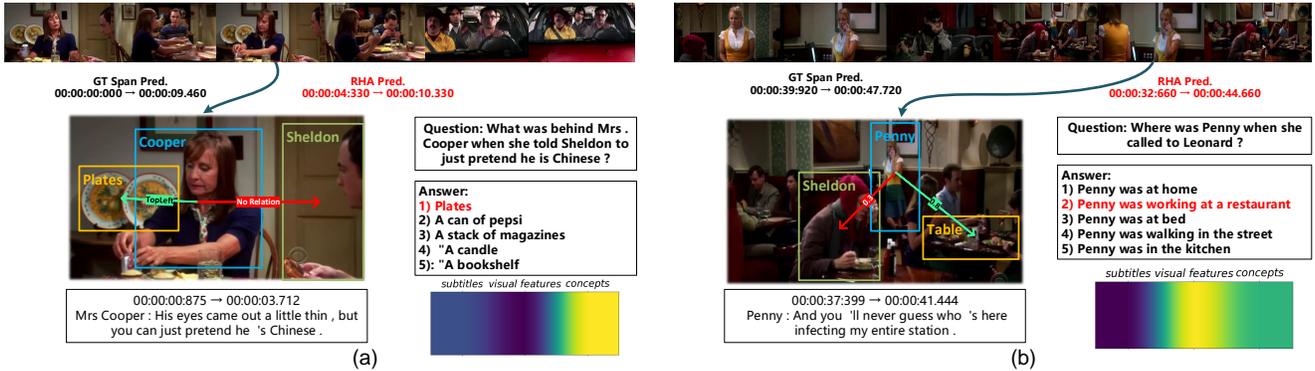


Figure 3: Positive examples to illustrate the process of the RHA framework. Darker color refers to lower modality weight.

by their bounding boxes so that the RHA can explicitly modeling spatial relationships and answer these kinds of questions from an interpretable perspective.

4.5.2 The usefulness of Visual Concepts and Visual Features. One difference between the RHA framework and previous models is that we use visual concepts rather than visual features to model spatial relations. Compared to visual features, visual concepts are more abstract and concise as a text-based category. The 2nd and 3rd lines in Table 3 show that using visual concepts for the spatial relation module can improve performance. The results validate our assumption that the abstract features that are easy to align with the question are more suitable than the specific features that are not in the same space as questions. In the ideal case that every object mentioned in question has its annotations, we can focus on understanding spatial relationship, ignoring the bias caused by object alignment.

As shown in the 2nd and 4th lines in Table 3, accuracy will decrease when the visual concepts are applied in the semantic relation encoder. We argue that using only categories in implicit relation reasoning is too hard. The absence of visual semantic information leads to wrong or unreasonable reasoning. For example, it is hard to judge why a vehicle is stopped if we only know there is a traffic light but do not know its direction and color. Compared to using too abstract embedding that may lead to misunderstanding, it is a better solution to use more detailed visual features in the reasoning stage and then use a convolutional layer or a fusion module to perform the alignment.

4.5.3 The effect of Hierarchical Attention. At last, we discuss the influence of the Hierarchical Attention (HA) module. Lines 1 and 5 in Table 3 shows that HA brings an improvement of about 2% in accuracy. In the first stage of HA, the question is used to measure the weight of frames and subtitles temporally. In general, only part of the time periods is related to the questions. By magnifying the weight of important time periods, the RHA framework can discover information better. Although in these periods, only a few objects and words in subtitles are strongly related (e.g. some words are cited in the question, or the question mentions specific objects). Thus the scores of each words and objects are calculated. The question is

actually encoded into the representation of the video in this stage so that the video representation is adapted to the specific questions.

In the second stage of HA, we reweight different modalities. One disadvantage of the question-guided spatio-temporal attention mentioned in the first stage is that all modalities actually have the same contribution for question answering. However, it is common that the question is unrelated to some modalities (e.g. the problem may be purely visually related and not related to subtitles). Previous research lacks a discussion on the importance of modality. According to the multimodal attention mechanism, our RHA framework learns the weight of different modalities adaptively, reaching a more precise localization of questions. Another scenario like "What does Sheldon say..." is also a typical case. Although it refers to a person, what did the person say is more important. For these questions, the key point is searching in candidate answers that have a similar representation with subtitles.

4.6 Case Study

To illustrate the RHA framework performance better, we select two right-answered examples randomly to better illustrate the process of RHA framework. As shown in Figure 3, we visualize the main objects of each frame and their relations. The weight of modalities is also visualized by color. Figure 3(a) shows the influence of spatial modeling. Given the question as "What was behind Mrs. Cooper when she told Sheldon to just pretend he is Chinese?", ground truth time annotation is 0s-9.46s, and the true answer is "Plates". At the 3.71 seconds of the video, the subtitles mentioned, "Mrs. Cooper: His eyes came out a little thin, but you can just pretend he's Chinese". The difference between ground truth and our result (0s vs. 4.33s) may because before 4.33s, although Cooper and Sheldon had occurred, Cooper did not say anything, so the RHA regards these frames as irrelevant. After 10s, plates never appeared again, and the conversation between Sheldon and Cooper shifted to other topics. RHA labeled 10.33s as the end time, which is still in the acceptable error range. As for spatial relation modeling, two main objects related to "Mrs. Cooper" are "Sheldon" and "plates". Among them, Sheldon was sitting on the right side in frame 2 and 4, and the plates were first overlapped with Cooper at frame 1, then at frame 3, the spatial relation between them is recognized as "close

to the left". Based on the spatial adjective "behind" in the question, RHA infers "Plates" as the right answer, indicating that RHA can understand spatial relations in video.

Figure 3(b) is another example. The given question is "Where was Penny when she called to Leonard?", time annotation is from 39.92s to 47.72s, while the ground truth answer is "Penny was working at a restaurant.", main object annotations include Penny, Leonard, and some tableware. RHA proposed 32.66s-42.66s as temporal localization. We find that at 32s, Sheldon was sitting in the restaurant, and at 42s, Penny called to Leonard. This case shows that RHA is inclined to use the moment of scene transition as the key point. RHA caught some objects that only appeared in the restaurant and connected them to Penny with a higher implicit relation score to answer the question. We also find our RHA gives subtitles a lower weight, which proves that RHA could focus on specific modalities by Hierarchical Attention.

Plus, we also visualize some negative examples predicted by our RHA framework, as shown in Figure 4. Most of the negative cases can be classified into four categories: (1)Temporal ambiguity; (1)Audio understanding; (2)Causal reasoning; (3)Counting. Figure 4(a) shows a typical case of temporal ambiguity. For given question "What is Amy drinking when evaluating the monkey?", we need to understand the meaning of "evaluating", which is obscurely hidden in Amy's dialogue. In Figure 4(b), although the question "Who knocked the door when Bernadette, Amy, and Penny were chatting?" is not quite difficult, the moment of knocking requires listening to the sound. In case that RHA only takes keyframes and subtitles as input, it can not understand any acoustic information, which finally leads to time dislocating (0-14s vs. 11-22s). In Figure 4(c), the question "Why did Raj tell himself to turn his pelvis when Penny was giving him a hug?" can be answered by the fact that Raj likes Penny and he is glad to have physical contact with her. However, commonsense and causal reasoning need external knowledge as a supplement, which is not involved in our framework. For these questions, knowledge-based methods provide a feasible idea. In Figure 4(d), we show a negative example caused by counting error. Given a question "How many times does Amy bounce the quarter into the glass when Amy and Penny are playing the game?", RHA locates the event precisely (14.33s-35.66s vs. 13.52s-32.08s). However, RHA fails to count the times of bouncing. For temporal counting, there is not a well-performed method yet, since it needs a deep understanding of action and number.

5 CONCLUSION

In this paper, we propose a novel RHA framework for VideoQA task. As a challenging video understanding task, VideoQA needs a comprehensive understanding of both visual and textual information. To address the video redundant phenomenon, we design a novel hierarchical attention module. The hierarchical attention module firstly measures the temporal and spatial importance based on their relevance to the question. Then scores of different modalities are calculated at the second stage to fuse multimodal features efficiently. As an important underlying semantic information, relations between objects reflect interaction and connections. To involve relation understanding into VideoQA, We build a graph-based relation encoder to capture such relation information and

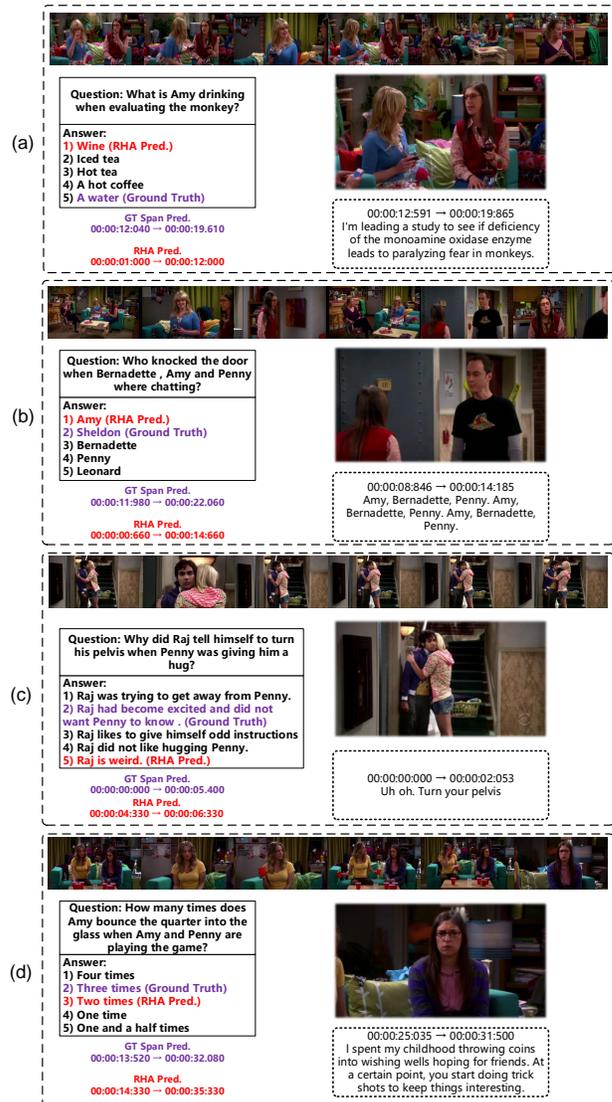


Figure 4: Some negative examples of TVQA+ valid set. RHA predictions are colored in red, while the ground truth predictions are colored in purple.

embed it into objects by weight-shared GATs. Experimental results on TVQA+ dataset validate the performance of RHA.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant No.61972047, the National Key Research and Development Program of China (2018YFC0831500), the Fundamental Research Funds for the Central Universities (500420824), the NSFC-General Technology Basic Research Joint Funds under Grant U1936220 and the Fundamental Research Funds for the Central Universities (2019XD-D01).

REFERENCES

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.
- [3] Peter L Bartlett and Marten H Wegkamp. 2008. Classification with a Reject Option using a Hinge Loss. *Journal of Machine Learning Research* 9, 8 (2008).
- [4] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2612–2620.
- [5] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. 2019. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1989–1998.
- [6] Pilin Dai, Jinna Lv, and Bin Wu. 2019. Two-Stage Model for Social Relationship Understanding from Videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1132–1137.
- [7] 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).
- [8] 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*. 6576–6585.
- [9] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. KnowIT VQA: Answering knowledge-based questions about videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10826–10834.
- [10] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.
- [11] Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. 2011. Multiple classifier systems for the classification of audio-visual emotional states. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 359–368.
- [12] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. 2018. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3588–3597.
- [13] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11021–11028.
- [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*. 2758–2766.
- [15] Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11109–11116.
- [16] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 1571–1581.
- [17] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [18] Thao Minh Le, Vuong Le, Svetla Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9972–9981.
- [19] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1369–1379.
- [20] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. TVQA+: Spatio-Temporal Grounding for Video Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8211–8225.
- [21] Fangtao Li, Wenzhe Wang, Zihe Liu, Haoran Wang, Chenghao Yan, and Bin Wu. 2021. Frame Aggregation and Multi-Modal Fusion Framework for Video-Based Person Recognition. In *International Conference on Multimedia Modeling*. Springer, 75–86.
- [22] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10313–10322.
- [23] Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving pairwise ranking for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3617–3625.
- [24] Yuanliu Liu, Bo Peng, Peipei Shi, He Yan, Yong Zhou, Bing Han, Yi Zheng, Chao Lin, Jianbin Jiang, Yin Fan, et al. 2018. iqiyi-vid: A large dataset for multi-modal person identification. *arXiv preprint arXiv:1811.07548* (2018).
- [25] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2247–2256.
- [26] Flávio LC Pádua, Anisio Lacerda, Adriano C Machado, Daniel H Dalip, et al. 2019. Multimodal data fusion framework based on autoencoders for top-N recommender systems. *Applied Intelligence* 49, 9 (2019), 3267–3282.
- [27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
- [28] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. 2019. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6966–6975.
- [29] Geovany A Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency. 2011. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 396–406.
- [30] Shaoqing Ren, Kaiming He, Ross B Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*.
- [31] Adam Santoro, David Raposo, David GT Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 4974–4983.
- [32] Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4613–4621.
- [33] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8992–8999.
- [34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.
- [35] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
- [36] Wenzhe Wang, Bin Wu, Fangtao Li, and Zihe Liu. 2020. Multi-Cue and Temporal Attention for Person Recognition in Videos. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 369–380.
- [37] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*. 684–699.
- [38] 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*. 3165–3173.
- [39] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *EMNLP*.
- [40] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [41] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging video descriptions to learn video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.