SEMANTIC ENRICHMENT FOR VIDEO QUESTION ANSWERING WITH GATED GRAPH NEURAL NETWORKS

Preprint · December 2023

CITATIONS
0

5 authors, including:

Chenyang Lyu
Dublin City University
46 PUBLICATIONS **116** CITATIONS
SEE PROFILE

Wenxi Li
Shanghai Jiao Tong University
5 PUBLICATIONS **1** CITATION
SEE PROFILE

Tianbo Ji
Nantong University
27 PUBLICATIONS **65** CITATIONS
SEE PROFILE

Longyue Wang
Tencent AI Lab
120 PUBLICATIONS **1,361** CITATIONS
SEE PROFILE

SEMANTIC ENRICHMENT FOR VIDEO QUESTION ANSWERING WITH GATED GRAPH NEURAL NETWORKS

Chenyang Lyu[1,2], Wenxi Li[3], Tianbo Ji[4], Yi Yu[5] , Longyue Wang[6]

[1]Dublin City University, Dublin, Ireland
[2]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[3]Shanghai Jiao Tong University, Shanghai, China
[4]Nantong University, Nantong, China
[5]National Institute of Informatics, Tokyo, Japan
[6]Tencent AI Lab, Shenzhen, China

ABSTRACT

Video Question Answering (VideoQA) is a complex task that requires a deep understanding of a video to accurately answer questions. Existing methods often struggle to effectively integrate the visual and language-based semantic information, subsequently leading to an incomplete understanding of video content and sub-optimal performance. To address the challenge, we introduce a novel approach in this paper to enrich the semantics of video frames, questions, and answer candidates. Specifically, we parse video frames and questions into semantic graphs - visual semantic graph and question semantic graph, which captures information about objects, their attributes, and relationships. These graphs are then encoded using a Gated Graph Neural Network (GGNN), For answer candidates, we propose to verbalize them using Large Language Models (LLMs) to further inject more semantic information from visual and acoustic aspects. We evaluate our approach on benchmark VideoQA datasets: AVQA and Music-AVQA. Experimental results show that our approach outperforms competitive baseline models, achieving state-of-the-art performance on various question types.

***Index Terms*—** VideoQA, Semantic Enrichment, Semantic Graph, GGNN, LLMs

1. INTRODUCTION

Video Question Answering (VideoQA) [1, 2, 3, 4, 5] is an emerging field that has gained increasing attention from the computer vision and natural language processing communities in recent years. VideoQA aims at understanding and answering questions about the visual content of videos, which is particularly challenging since it requires the model to effectively integrate visual and linguistic information to comprehend the video content and accurately find the appropriate answers for the given questions [6, 1, 3, 7]. Despite the significant progress made in VideoQA [8, 9, 4, 10], existing approaches still have limitations in understanding the complex relationships and dynamics within videos and accurately finding answers for the given questions from a set of candidates with ambiguous or incomplete information. One of the primary challenges in VideoQA is effectively integrating visual and language-based semantic information in the data - video, question and answer. Video content is inherently multimodal, consisting of both visual and acoustic components, and effectively leveraging the semantic information in both modalities is crucial for accurate VideoQA [1, 6]. Moreover, the importance of the semantic information in questions and answers, which are usually incomplete and ambiguous in the VideoQA data, are often underestimated [8, 11]. Therefore, such issue of lacking rich semantic information data can lead to suboptimal performance in VideoQA.

To address these challenges, we propose a novel approach focusing on semantic enrichment that leverages the rich semantic information in semantic graphs of video frames and questions and verbalize answer candidates with Large Language Models (LLMs) [12, 13] to improve the performance of VideoQA systems. Specifically, we parse video frames and questions to their corresponding semantic graph: Visual Semantic Graph and Question Semantic Graph. The semantic graphs contain the information of objects, attributes, and relationships between them, which is able to capture the fine-grained semantic information in the video frames and questions. We then use a Gated Graph Neural Network (GGNN [14]) to encode the semantic graphs of video frames and questions, which leverages the graph structure to propagate information across the nodes and edges in the semantic graphs. We also employ an additional gating mechanism to control the degree of information passing between nodes. For enriching the semantics of answer candidates, we use LLMs to verbalize them in order to extend the semantic information in the answer candidates using the knowledge in language models. Our approach uses language models to expand the original answer candidates and enrich the semantic

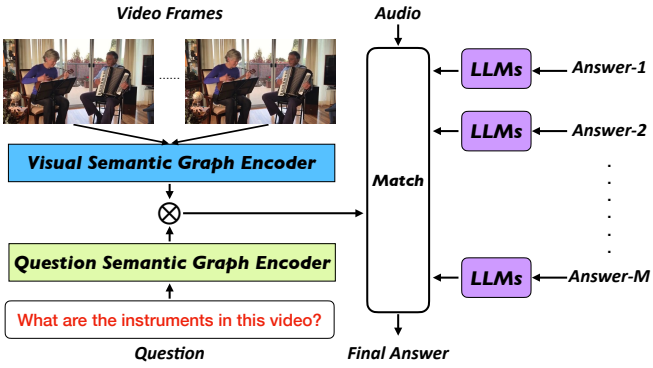

Fig. 1. An overview of our proposed approach for VideoQA.

of them by providing more detailed description.

We evaluate our proposed approach on two benchmark VideoQA datasets: AVQA [4] and Music-AVQA [15] and demonstrate that it outperforms existing state-of-the-art methods in terms of accuracy. Our ablation studies further show the effectiveness of our semantic enrichment approach in improving the overall performance of VideoQA systems.

2. METHODOLOGY

In this section, we describe our proposed approach for VideoQA with semantic enrichment for video frames, questions and answer candidates. Figure 1 shows the overall architecture of our approach. Our VideoQA data consists of video frames (V_f), audio (A), questions (Q), and answer candidates (Ans). We firstly encode video frames and audio in section 2.1. In section 2.2 and 2.3, we parse video frames and the question to corresponding semantic graphs, which are then encoded by GGNNs. We subsequently enrich the semantics of answer candidates in section 2.4 via verbalizing them to inject more detailed information into answers. Finally, in section 2.5 we fuse the semantic graph representations of video frames and the question to produce a joint representation, which is used in matching the most plausible answer candidate.

2.1. Encoding Video Frames and Audio

For each video frame V_f^i, we employ a ViT [16] to obtain its representations $H_f^i \in R^{m\times 1}$. Then we get the video-level representation via: $H^v = \sum_i U^i H_f^i$, where $H^v \in R^{m\times 1}$ and $U^i \in R^{m\times m}$ is the weight matrix for the i-th frame. We encode audio A via a transformer-based encoder [17] to obtain its representations $H^a \in R^{m\times 1}$.

2.2. Encoding Visual Semantic Graph

We parse the video frames to visual semantic graph. Specifically, we firstly use an image captioning model to obtain captions for each video frame [18] which are then converted by visual scene parser [19] to visual semantic graphs. The visual semantic graphs contain the objects and the relationships among them, which capture rich semantic information in the video frames. To encode the visual semantic graphs, we propose using a GGNN, which consists of multiple layers and each layer updates the node representations based on the representations of their neighboring nodes. The node representations are then passed through a gating mechanism to selectively combine the information from the current and previous layers.

Let $G^v = (V^v, E^v)$ be the visual semantic graph for a single video frame, where V^v is the set of nodes representing the objects and actions in the video frames, and E^v is the set of edges representing the relationships between them. Each node v_i is associated with a feature vector $h_i^0 \in R^{m\times 1}$ representing its appearance and semantic information. We use pre-trained contextualized word representations [20] to obtain the initial node features for both visual semantic graph and question semantic graph.

The node representations are updated using the following equation:

$$h_i^l = \text{GNN}^l(h_i^{l-1}, h_j^{l-1} j \in \mathcal{N}(i)), \tag{1}$$

where GNN^l is the GNN layer at level l, $\mathcal{N}(i)$ is the set of neighboring nodes of v_i, and $h_j^{l-1} j \in \mathcal{N}(i)$ is the set of representations of its neighbors.

The gating mechanism is used to selectively combine the information from the current and previous layers:

$$h_i^l = \sigma(W_l[h_i^{l-1}; h_i^l]) \odot h_i^{l-1} + (1 - \sigma(W_l[h_i^{l-1}; h_i^l]^T])) \odot h_i^l, \tag{2}$$

where σ is the sigmoid function, $W_l \in R^{m\times 2m}$ is the weight matrix at level l, $[h_i^{l-1}; h_i^l]$ is the concatenation of the current and previous representations of node v_i, and $\odot$ is the element-wise multiplication.

After L layers of GGNN updates, we obtain the final node representations $h_i^L \in R^{m\times 1}$. We firstly obtain a frame-level representation: $f = \sum_i \sum_j W_f^{i,j} h_i^j$, where $f \in R^{m\times 1}$ and $W_f^{i,j} \in R^{m\times m}$ represents the weight matrix for fusing the representations of the i-th node in the j-th layer. We the combine f to obtain a video-level representation: $H_g^v = \sum_j W_v^j f_j$, where $W_v^j \in R^{m\times m}$ and $H_g^v \in R^{m\times 1}$. The graph-based video-level representation H_g^v is then fused with the graph representations of the question to match the most probable answer candidate for the given question.

2.3. Encoding Question Semantic Graph

For enriching the semantics of questions, we employ Semantic Role Labeling [21] to parse the question to a semantic graph which captures the linguistic-based semantic structure of the question. To encode the question semantic graph, we

adopt a similar approach as the one used for the visual semantic graph. Given a question Q, we construct a semantic graph $G^q = (V^q, E^q)$, where V^q and E^q represents the set of nodes and edges. Each node v_i^q is associated with a vector representation $r_i^l \in R^{m\times 1}$ that corresponds to the representations of the i-th node in the l-th GGNN layer.

During the node update phase, we firstly follow Eq. 1 to update the node representations. Specifically, the node in question semantic graph is then updated with the gating mechanism based on on the aggregated messages from its neighbors, its current hidden state as well as the visual and audio representations:

$$r_i^l = \sigma(W_l[r_i^{l-1}; r_i^l]) \odot r_i^{l-1} + (1-\sigma(W_l[r_i^{l-1}; r_i^l])) \odot r_i^l, \quad (3)$$

where σ is the sigmoid function, $W_l \in R^{m\times m}$ is the weight matrix at level l, $[r_i^{l-1}; r_i^l]$ is the concatenation of the current and previous representations of node v_i^q, and $\odot$ is the element-wise multiplication. We combine the node representations to obtain a graph-level representations of the question: $H_q = \sum_i \sum_j W_q^{i,j} r_i^j$, where $H_q \in R^{m\times 1}$ and $W_q^{i,j} \in R^{m\times m}$.

2.4. Verbalizing Answer Candidates with LLMs

The verbalization of answer candidates using LLMs [1] is a critical aspect of our proposed methodology. This approach allows for the augmentation of original answer candidates with detailed and diverse descriptions, particularly focusing on visual and acoustic elements of the video content, while meticulously avoiding the generation of irrelevant details.

The process begins with inputting the initial answer candidates, which could be objects, actions, or a combination of both derived from the video content, into the LLM. A specially designed prompt p (*Please add more additional relevant details from perspectives such as acoustic and visual aspects...without introducing superfluous information...*) is then used to request the LLM to elaborate on these initial answer candidates. The intention behind this prompt is to stimulate the LLM to generate more detailed and diverse descriptions of the answer candidates from visual and acoustic aspects, and to prevent the generation of irrelevant details. The LLM-generated descriptions are subsequently integrated into the original answer candidates, resulting in enriched answer candidates that are ready for the next steps of our VideoQA process. We obtain the enriched representations of the i-th answer candidate via:

$$H_i^{ans} = g(LLMs([p; Ans_i])) \quad (4)$$

where $H_i^{ans} \in R^{m\times 1}$ is the enriched representations of the i-th answer candidate, g is the answer encoder [20] and $LLMs([p; Ans]$ represents that we input concatenation of the prompt p with the original answer Ans to LLMs.

[1] In this work, we use GPT-4 (https://platform.openai.com/docs/model-index-for-researchers).

2.5. Training objective

After obtaining all the necessary representations, we finally fuse these features to generate a joint representation of the video and question:

$$H = \mathbf{E}(H^v, H_g^v, H^a, H^q) \quad (5)$$

where $H \in R^{m\times 1}$ is the joint representation of the video and question, $\mathbf{E}$ represents the fusing function consisting of a linear layer followed by a ReLU.

During the training process of our VideoQA system, we leverage the *Cross-Entropy* loss function:

$$J = -\frac{1}{N} \sum_{k=1}^{P} \frac{exp^{(H_k^{ans})^T H_k}}{\sum_{l=1}^{M} exp^{(H_{k,l}^{ans})^T H_k}} \quad (6)$$

Here, H_k^{ans} signifies the correct answer representation in the k-th example, whereas M represents the total number of possible answer candidates as we focus on multiple-choice VideoQA.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets We tested our method on two well-known audio-visual VideoQA datasets: 1) AVQA [4], which includes 57,015 real-life videos and 57,335 question-answer pairs. It's divided into training (34,401), validation (5,734), and testing (17,200) samples. 2) Music-AVQA [15], designed for multi-modal understanding. It has 45,867 question-answer pairs across 9,288 videos. The dataset is split into training (32,087), validation (4,595), and testing (9,185) QA pairs.

Hyperparameters In training, we use AdamW [22] to adjust the model parameters, setting epsilon at 1×10^{-8}. We base our setup on CLIP [20] to initialize our visual encoder and text encoder, we use Whisper [17] to initialize the audio encoder. Training is done for 10 epochs with a learning rate of 3×10^{-5} and a batch size of 4, we set L to 3. We uniformly sample 16 frames from each video. We use a maximum gradient norm of 5 and stop early if validation performance drops.

3.2. Results

We present our main experimental results in Table 1 and Table 2. Table 1 illustrates the comparative performance of various methods on the AVQA dataset divided by question types. It is clear from the table that our proposed approach substantially outperforms all other methods, emerging as the most accurate across all question types. In particular, it shows noteworthy improvements in complex categories such

Table 1. Experimental results of VideoQA on AVQA [4] test set divided by question types. The performance of state-of-the-art approaches are taken from [4].

Methods	Which	Come From	Happening	Where	Why	Before Next	When	Used For	Total Accuracy
HME [1]	82.2	85.9	79.3	76.6	57.0	80.0	57.1	76.5	81.8
HME+HAVF [4]	85.6	88.3	83.1	83.5	61.6	80.0	57.1	88.2	85.0
PSAC [8]	78.7	80.0	77.0	79.4	44.2	76.0	42.9	58.8	78.6
PSAC+HAVF [4]	89.0	91.1	83.2	81.7	61.6	82.0	52.4	76.5	87.4
LADNet [9]	81.1	87.1	76.6	81.8	67.4	78.0	47.6	76.5	81.9
LADNet+HAVF [4]	84.2	89.0	79.1	81.4	68.6	82.0	52.4	76.5	84.1
ACRTransformer [2]	82.5	82.8	79.4	82.5	54.7	80.0	47.6	58.8	81.7
ACRTransformer+HAVF [4]	88.5	91.7	83.9	84.9	50.0	82.0	57.1	64.7	87.8
HGA [23]	82.1	84.3	79.5	83.1	59.3	82.0	57.1	88.2	82.2
HGA+HAVF [4]	88.6	92.2	83.8	82.6	61.6	78.0	52.4	82.4	87.7
HCRN [24]	83.7	84.1	80.2	80.9	52.3	74.0	57.1	70.6	82.5
HCRN+HAVF [4]	89.8	92.8	86.0	84.4	57.0	80.0	52.4	82.4	89.0
GMGA [5]	93.7	97.3	90.4	89.5	61.8	92.0	64.9	88.2	93.0
Our approach	**94.1**	**98.5**	**91.1**	**90.5**	**63.3**	**92.0**	**67.1**	**89.2**	**94.2**

Table 2. Experimental results of different models on the test set of Music-AVQA[15]. We compare our proposed method with state-of-the-art approaches on Music-AVQA, of which the results are taken from [15].

Methods	Audio Question			Visual Question			Audio-Visual Question						All
	Counting	Comparative	Avg.	Counting	Location	Avg.	Existential	Location	Counting	Comparative	Temporal	Avg.	Avg.
FCNLSTM [3]	70.45	66.22	68.88	63.89	46.74	55.21	82.01	46.28	59.34	62.15	47.33	60.06	60.34
CONVLSTM [3]	74.07	68.89	72.15	67.47	54.56	60.94	82.91	50.81	63.03	60.27	51.58	62.24	63.65
BiLSTM Attn [25]	70.35	47.92	62.05	64.64	64.33	64.48	78.39	45.85	56.91	53.09	49.76	57.10	59.92
HCAttn [26]	70.25	54.91	64.57	64.05	66.37	65.22	79.10	49.51	59.97	55.25	56.43	60.19	62.30
MCAN [27]	77.50	55.24	69.25	71.56	70.93	71.24	80.40	54.48	64.91	57.22	47.57	61.58	65.49
PSAC [8]	75.64	66.06	72.09	68.64	69.79	69.22	77.59	55.02	63.42	61.17	59.47	63.52	66.54
HME [1]	74.76	63.56	70.61	67.97	69.46	68.76	80.30	53.18	63.19	62.69	59.83	64.05	66.45
HCRN [24]	68.59	50.92	62.05	64.39	61.81	63.08	54.47	41.53	53.38	52.11	47.69	50.26	55.73
AVSD [6]	72.41	61.90	68.52	67.39	74.19	70.83	81.61	58.79	63.89	61.52	61.41	65.49	67.44
Pano-AVQA [11]	74.36	64.56	70.73	69.39	75.65	72.56	81.21	59.33	64.91	64.22	63.23	66.64	68.93
Music-AVQA [15]	78.18	67.05	74.06	71.56	76.38	74.00	81.81	64.51	70.80	66.01	63.23	69.54	71.52
GMGA [5]	85.97	74.43	81.72	75.46	81.71	78.63	86.20	71.13	77.94	73.79	72.26	76.49	77.87
Our approach	**87.32**	**75.86**	**83.12**	**76.46**	**83.06**	**79.89**	**87.31**	**72.15**	**79.21**	**74.97**	**73.56**	**77.83**	**79.61**

Table 3. Ablation study results on Music-AVQA.

Methods	Accuracy
Our approach	79.61
Our approach w/o Visual Semantic Graph	78.32
Our approach w/o Question Semantic Graph	77.96
Our approach w/o Answer verbalization	76.49

as *Which, Come From*, and *Happening*. The overall accuracy of our approach reaches 94.2, exceeding the previous state-of-the-art method by 1.2 points. Table 2 showcases the performance of various methods on the Music-AVQA dataset across three major types of questions: Audio, Visual, and Audio-Visual as well as sub-types under each category. Our proposed approach outperforms all other methods in every question category, achieving the highest accuracy. Specifically, for Audio Questions, Visual Questions, and Audio-Visual Questions, our method achieved average accuracies of 83.12, 79.89, and 77.83 respectively. The overall average accuracy of our approach reached 79.61, surpassing all other compared methodologies. The experimental results underscore the effectiveness of our method and its superior ability on VideoQA task.

3.3. Ablation Study

Table 3 presents an ablation study aiming to understand the individual contribution of each component: *Visual Semantic Graph*, *Question Semantic Graph*, and *Answer Verbalization*, to the overall performance. From the results in Table 3, when the *Visual Semantic Graph* was excluded, accuracy dropped to 78.32 from 79.61. Similarly, excluding *Question Semantic Graph* decreased accuracy to 77.96. The most significant drop was observed when *Answer Verbalization* was removed, with accuracy falling to 76.49. In conclusion, all components are crucial for achieving the best performance for VideoQA.

4. CONCLUSION AND FUTURE WORK

We proposed a novel approach for VideoQA that significantly improves performance by leveraging semantic graphs and LLMs. Our method effectively extracts fine-grained semantic information from video frames and questions, and enriches the semantic context of answer candidates. The approach outperformed state-of-the-art methods on AVQA and Music-AVQA datasets. Future work will focus on enhancing our semantic graphs to capture deeper information levels and refining our answer verbalization method with advanced language models.

5. REFERENCES

[1] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *CVPR*, 2019.

[2] Jipeng Zhang, Jie Shao, Rui Cao, Lianli Gao, Xing Xu, and Heng Tao Shen, "Action-centric relation transformer network for video question answering," *IEEE TCSVT*, vol. 32, no. 1, pp. 63–74, 2020.

[3] Haytham M Fayek and Justin Johnson, "Temporal reasoning via audio question answering," *IEEE TASLP*, vol. 28, pp. 2283–2294, 2020.

[4] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu, "Avqa: A dataset for audio-visual question answering on videos," in *ACMMM*, 2022.

[5] Chenyang Lyu, Wenxi Li, Tianbo Ji, Longyue Wang, Liting Zhou, Cathal Gurrin, Linyi Yang, Yi Yu, Yvette Graham, and Jennifer Foster, "Graph-based video-language learning with multi-grained audio-visual alignment," in *ACMMM*, 2023.

[6] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh, "Audio-visual scene-aware dialog," in *CVPR*, 2019.

[7] Chenyang Lyu, Wenxi Li, Tianbo Ji, Liting Zhou, and Cathal Gurrin, "Gated multi-modal fusion with cross-modal contrastive learning for video question answering," in *ICANN*, 2023.

[8] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan, "Beyond rnns: Positional self-attention with co-attention for video question answering," in *AAAI*, 2019.

[9] Xiangpeng Li, Lianli Gao, Xuanhan Wang, Wu Liu, Xing Xu, Heng Tao Shen, and Jingkuan Song, "Learnable aggregating net with diversity learning for video question answering," in *ACMMM*, 2019.

[10] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu, "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration," *arXiv preprint arXiv:2306.09093*, 2023.

[11] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim, "Pano-avqa: Grounded audio-visual question answering on 360deg videos," in *ICCV*, 2021.

[12] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2303.

[13] Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu, "Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation," 2023.

[14] Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, Bo Long, et al., "Graph neural networks for natural language processing: A survey," *FTML*, vol. 16, no. 2, pp. 119–328, 2023.

[15] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *CVPR*, 2022.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.

[17] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.

[19] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *EMNLPW*, 2015.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[21] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer, "Deep semantic role labeling: What works and what's next," in *ACL*, 2017.

[22] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[23] Pin Jiang and Yahong Han, "Reasoning with heterogeneous graph alignment for video question answering," in *AAAI*, 2020.

[24] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran, "Hierarchical conditional relation networks for video question answering," in *CVPR*, 2020.

[25] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao, "Attention-based lstm network for cross-lingual sentiment classification," in *EMNLP*, 2016.

[26] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, "Hierarchical question-image co-attention for visual question answering," *arXiv preprint arXiv:1606.00061*, 2016.

[27] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian, "Deep modular co-attention networks for visual question answering," in *CVPR*, 2019.