Multimodal Commonsense Knowledge Distillation for Visual Question Answering (Student Abstract)

Shuo Yang[1], **Siwen Luo**[2], **Soyeon Caren Han**[1,2]
[1]University of Melbourne,
[2]University of Western Australia
syang4418@student.unimelb.edu.au, siwen.luo@uwa.edu.au, caren.han@unimelb.edu.au

Abstract

Existing Multimodal Large Language Models (MLLMs) and Visual Language Pretrained Models (VLPMs) have shown remarkable performances in general Visual Question Answering (VQA). However, these models struggle with VQA questions that require external commonsense knowledge due to the challenges in generating high-quality prompts and the high computational costs of fine-tuning. In this work, we propose a novel graph-based multimodal commonsense knowledge distillation framework that constructs a unified relational graph over commonsense knowledge, visual objects and questions through a Graph Convolutional Network (GCN) following a teacher-student environment. This proposed framework is flexible with any type of teacher and student models without further fine-tuning, and has achieved competitive performances on the ScienceQA dataset. The code is in https://github.com/adlnlp/MCKDVQA.

Introduction

In recent years, VQA tasks developed to be more challenging by asking questions beyond the image contents and requiring external commonsense knowledge to answer[1]. Existing works on such commonsense VQA tasks tried different methods to integrate visual, question and commonsense knowledge features (Wang, Han, and Poon 2024). For example, Ravi et al. (2023) encodes the contextualized commonsense inferences on the question phrases as additional textual features and integrates with object visual features to fine-tune the Vision-and-Language pretrained model (VLPM). VQA-GNN (Wang et al. 2023) jointly encodes the scene graph of the image and concept graph of the question context as the unified graph for training. T-SciQ (Wang et al. 2024) proposes the new chain-of-thought (CoT) prompting strategy to fine-tune the Multimodal large language model (MLLM). However, these works face problems from two aspects: 1) though incorporating CoT in MLLMs has shown remarkable performances on knowledge-based VQA, generating the high-level reasoning CoT is challenging; 2) directly fine-tuning the large VLMs can be computationally expensive. To address these issues, in this work, we propose a multimodal teacher-student knowledge distillation framework that is computationally efficient to jointly learn the features of multi-modalities (Cabral et al. 2024; Han et al. 2020; Cao et al. 2023). Specifically, in the teacher model, this framework integrates the object entities from image, question and commonsense knowledge graph together in a unified graph and explicitly learns the relationships among them through the Graph Convolutional Neural Network (GCN)(Yao, Mao, and Luo 2019), inspired by Han et al. (2022) and Long et al. (2022). The learned graph features are passed to the student model, which can be any model structure of a smaller size, for the final answer prediction. Notably, instead of fine-tuning based on one vision-and-language model structure, this framework can be flexibly plugged with any pretrained visual and textual encoder for diverse feature extractions in the teacher model. Moreover, this proposed method provides flexibility that can be adapted to environments with different computational efficacies while maintaining competitive performances compared to large VLPMs and MLLMs. We evaluated our proposed framework with the ScienceQA and achieved competitive results.

Methodology

Figure 1 depicts the overall workflow of our proposed graph-based multimodal commonsense knowledge distillation framework. We first represent inputs as graphs to capture the relationships between different modalities enriched by commonsense knowledge. We then employ a GCN to train the teacher graph model. This trained teacher then distils learnt knowledge to the student models of varying size.

Graph Construction: To capture the relationships among the multimodal inputs and enrich them with commonsense knowledge understanding, we construct a set of heterogeneous subgraphs $G = \{G_1, G_2, \dots, G_M\}$ for a dataset with M samples. Each subgraph $G_i = \{V_i, E_i\}$ represents an individual input sample comprising an image, a question, and contextual information. The node candidates V_i within each subgraph are categorised into two types: content nodes V_{sub} and commonsense nodes V_k. The content nodes V_{sub} includes four types of node representation for each input modality: a question node for the textual query, a language context node for textual context, a visual context node for image context and a V-L node for combined visual and textual context.

To further inject the model with augmented common-

[1]Existing VLMs' error cases can be found in Appendix

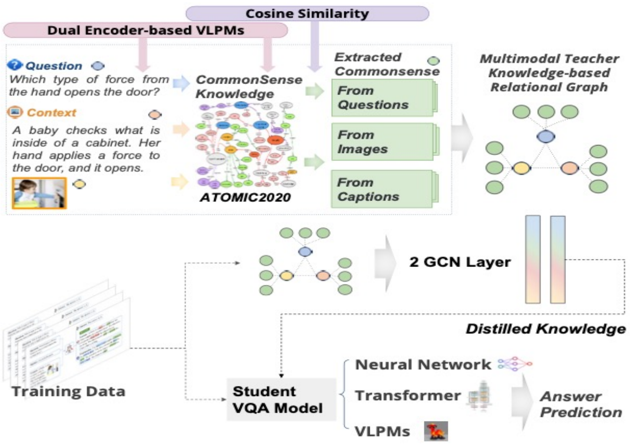

Figure 1: Overall Framework Design

sense knowledge, we integrate commonsense nodes V_k into each subgraph. Initially, each content node V_{sub} is projected into a shared single-modal embedding space using a dual-encoder-based Vision-Language Pretrained Model. We then retrieve relevant commonsense knowledge triplets from the ATOMIC2020 dataset (Hwang et al. 2021). Specifically, we compute the cosine similarity between the embedding vector $\mathbf{v}_u$ of each content node V_u and the embeddings $\mathbf{v}_k$ of all triplets in the ATOMIC2020 dataset as illustrated as:

$$\mathrm{sim}(V_u, k) = \frac{\mathbf{v}_u \cdot \mathbf{v}_k}{\|\mathbf{v}_u\| \|\mathbf{v}_k\|}. \quad (1)$$

The triplets are pre-embedded into the same shared space using the VLPM. We select the top K triplets with the highest similarity scores for each content node V_u (we set $K = 3$ in our experiments). These selected triplets are considered the most semantically relevant and are added to the subgraph as commonsense nodes V_k.

Edges for any pair of nodes $V_x, V_y \in V_{sub}$ as well as V_u with their retrieved commonsense nodes V_k are defined by either Cosine Similarity and Pointwise Mutual Information (PMI). These metrics are chosen to capture semantic relationships and statistical dependencies among the nodes.

Graph Learning: We leverage a standard two-layer Graph Convolutional Network (GCN) to capture the multimodal information and injected commonsense knowledge within the constructed graph. It is illustrated in Equation 2:

$$f(V)^{(l+1)} = \sigma^{(l)} \left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} f(V)^{(l)} W^{(l)} \right) \quad (2)$$

where: $f(V)^{(l)} \in \mathbb{R}^{N \times T^{(l)}}$ represents the node feature at layer l, $\tilde{A} = A + I_N \in \mathbb{R}^{N \times N}$ is the adjacency matrix of the graph with added self-connections; N is the number of nodes within each subgraph and $T^{(l)}$ is the dimension of feature space at layer l. We then apply an average pooling $f_{\text{pooling}}(\cdot) : \mathbb{R}^{N \times T^{(l)}} \rightarrow \mathbb{R}^{1 \times T^{(l)}}$ over each subgraph and feed the pooled embedding over each sub-graph to a multi-layer perception (MLP) $f_{\text{MLP}}(\cdot)$: $\mathbb{R}^{M \times T^{(L)}} \rightarrow \mathbb{R}^{M \times T^{(O)}}$, where $T^{(O)}$ denotes the number of the unique labels. We use the cross-entropy loss to optimize the model.

Model	NAT	SOC	LAN	AVG
Small-sized Baseline and Our Result				
MLP	41.21	42.33	34.11	42.71
Teacher + MLP (ours)	**54.38**	**49.23**	**39.28**	**53.92**
Medium-sized Baseline and Our Result				
Transformer	48.44	47.15	42.72	48.35
Teacher + Transformer (ours)	**57.74**	**55.35**	**48.46**	**56.79**
Large-sized Baselines and Our Result				
ViLT	60.48	63.89	60.27	61.14
Teacher + ViLT (ours)	**64.12**	**66.55**	**63.83**	**65.41**
VisualBERT	59.33	69.18	61.18	61.87
Teacher + VisualBERT (ours)	**61.30**	**72.84**	**64.33**	**65.69**
UnifiedQA$_{base}$	68.16	69.18	74.91	70.12
Teacher + UnifiedQA$_{base}$ (ours)	**71.41**	**73.22**	71.58	**72.33**

Table 1: Overall Performance on ScienceQA. Question classes: NAT = natural science, SOC = social science, LAN = language science. (Lu et al. 2022).

Multimodal Graph-based Knowledge Distillation: After training the teacher graph using GCN, we distil soft labels to the student model, where it is optimised by the Kullback-Leibler Divergence (KD) loss as in Equation 3:

$$\begin{aligned} \mathcal{L}_{KD} &= \text{KLDivLoss}\left(\frac{1}{n_T} \sum_{j=1}^{n_T} P_j, P_s \right) \\ P_j = \text{softmax}&(T_j(X)) \quad P_s = \text{softmax}(S(X)) \end{aligned} \quad (3)$$

where $T(X)$, $S(X)$ represents the teacher model and student model. We formulate overall loss by adding up the student cross-entropy loss and KD loss as $\mathcal{L} = \mathcal{L}_{SCE} + \mathcal{L}_{KD}$.

Experiments and Results

We compare the micro F1-score against three types of baseline models of varying size: (1) Small-sized **MLP**; (2) Medium-sized **Transformer**; (3) Three Large-sized **VLPMs** that has been applied in the ScienceQA dataset (Lu et al. 2022): (a) VisualBERT (Li et al. 2019): integrates RoI-based visual feature and token-based textual feature through BERT-style architecture. (b) ViLT (Kim, Son, and Kim 2021): processes visual and textual tokens using a unified fusion encoder directly. (c) UnifiedQA (Khashabi et al. 2020): unifies various QA format throughout a textual-only model. We evaluate the proposed framework on the ScienceQA (Lu et al. 2022). Each group is tested with or without integrating our proposed graph-based knowledge distillation framework. From the overall performance covered in Table 1, we can see a significant improvement in their average score with 11.21% and 8.44% increase separately with our proposed framework for both MLP and Transformer baselines. For large VLPMs, despite their sophistication, we also find a non-trivial increment in their performance. This suggests the robustness and effectiveness of our method.

Conclusion

We proposed a multimodal graph-based commonsense knowledge distillation framework that addresses the limitations of existing VLMs in VQA tasks by integrating object, question, and commonsense knowledge into a unified graph structure and leveraging a GCN for relational learning. Our results on ScienceQA validate the effectiveness of this approach, showing notable performance improvements.

Acknowledgements

This study was supported by funding from the Google Award for Inclusion Research Program (G222897). We thank Dr. Josiah Poon (The University of Sydney) and Prof. Eduard Hovy (The University of Melbourne) for fruitful suggestions and discussions.

References

Cabral, R. C.; Luo, S.; Poon, J.; and Han, S. C. 2024. 3M-Health: Multimodal Multi-Teacher Knowledge Distillation for Mental Health Detection. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 152–162.

Cao, F.; Luo, S.; Nunez, F.; Wen, Z.; Poon, J.; and Han, S. C. 2023. Scenegate: Scene-graph based co-attention networks for text visual question answering. *Robotics*, 12(4): 114.

Han, C.; Long, S.; Luo, S.; Wang, K.; and Poon, J. 2020. VICTR: Visual Information Captured Text Representation for Text-to-Vision Multimodal Tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, 3107–3117.

Han, S. C.; Yuan, Z.; Wang, K.; Long, S.; and Poon, J. 2022. Understanding graph convolutional networks for text classification. *arXiv preprint arXiv:2203.16060*.

Hwang, J. D.; Bhagavatula, C.; Le Bras, R.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 6384–6392.

Khashabi, D.; Min, S.; Khot, T.; Sabharwal, A.; Tafjord, O.; Clark, P.; and Hajishirzi, H. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1896–1907.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, 5583–5594. PMLR.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Long, S.; Han, S. C.; Wan, X.; and Poon, J. 2022. Gradual: Graph-based dual-modal representation for image-text matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3459–3468.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Ravi, S.; Chinchure, A.; Sigal, L.; Liao, R.; and Shwartz, V. 2023. Vlc-bert: Visual question answering with contextualized commonsense knowledge. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1155–1165.

Wang, E.; Han, C.; and Poon, J. 2024. SCO-VIST: Social Interaction Commonsense Knowledge-based Visual Storytelling. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1602–1616.

Wang, L.; Hu, Y.; He, J.; Xu, X.; Liu, N.; Liu, H.; and Shen, H. T. 2024. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19162–19170.

Wang, Y.; Yasunaga, M.; Ren, H.; Wada, S.; and Leskovec, J. 2023. Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21582–21592.

Yao, L.; Mao, C.; and Luo, Y. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 7370–7377.