Exploring deep learning for multimodal understanding

Lao, M.

Chapter 9

Conclusions

In this thesis, we have addressed six research questions regarding the three themes: multimodal fusion, language bias, audio visual question answering, lifelong learning, and personalized federated learning. In this chapter, we conclude the main findings from our approaches and results. In addition, we discuss the limitations of our methods and possible solutions to address them. Last but not least, we point out several trends for future work.

9.1 Main Findings

In each research chapter, we have proposed a new approach to answer the corresponding research question. Next, we conclude these approaches and present the main findings inspired by empirical and theoretical analyses.

(1) We start the multimodal fusion research (**RQ1**) in Chapter 2 with proposing a novel Multi-stage Hybrid Embedding Fusion (MHEF) network to achieve fine-grained and efficient vision-language interactions for answer prediction. Specifically, we introduce a Hybrid Embedding Fusion (HEF) approach to enrich visual and textual representation via triplet-space feature projections. Then, motivated by the multi-step reasoning in human cognition, we present a novel Multi-stage Fusion Structure (MFS) to obtain diverse and better fusion features for answer prediction. Through combining HEF and MFS to train a multi-stage prediction framework, we can not only improve the performance in each single stage, but also obtain additional accuracy boosts by integrating all prediction results from each stage. Extensive experiments verify that our MHEF method yields superior results over existing multimodal fusion scheme, and further achieve promising performance on two widely-used VQA datasets.

(2) After investigating the model architecture in VQA models, we turn to address the language bias problem (**RQ2**), which severely impair the reliability and robustness in current VQA models. To alleviate the models' overdependence on data bias,

we propose a simple yet effective Language Priors based Focal Loss (LP-Focal Loss) by rescaling the standard cross entropy loss. Specifically, our method exploited language priors captured by a question-only branch, and further dynamically assigned weights for different training instances. Extensive experiments verified the effectiveness and generalizability of the LP-Focal Loss, and it achieved state-of-the-art performance on the VQA-CP v2 dataset.

(3) Apart from the rectification of loss function for bias mitigation, we seek to propose a more efficient and delicate learning framework (**RQ3**) to further improve the out-of-distribution performance for unbiased VQA models. Inspired by the easy-to-hard cognitive process of human beings, we propose a novel Language Bias driven Curriculum Learning to alleviate language bias from a multi-stage curriculum learning manner with knowledge distillation. Extensive experiments verify that our method is model-agnostic, and achieve state-of-the-art performance on widely-used out-of-distribution dataset. We believe our appraoch inspires future works related to unbiased VQA models.

(4) Going beyond VQA models, where we turn to discover the shortcut bias problem in more complicate Audio-Visual Question Answering (AVQA) task (**RQ4**). Through detailed causal-graph analyses and careful inspections of their learning processes, we reveal that AVQA models are not only prone to over-exploit the prevalent language bias, but also suffer from additional joint-modal biases caused by the shortcut relations between textual-auditory/visual co-occurrences and dominated answers. To tackle this issue, we introduce a model-agnostic COllaborative CAusal (COCA) Regularization to jointly overcome multiple shortcut biases via causal inference in both factual and counterfactual scenarios. To our best knowledge, this work is the first attempt to analyze the potential biases in AVQA task from the perspective of causal graph, and we believe that it would shed the light on the future works for rubust AVQA models.

(5) Instead of training VQA models under a stationary domain that is fixed by the choice of a given dataset, we turn to focus on the VQA learning framework under lifelong learning setting (**RQ5**). To put the VQA models into practice, we present a novel yet practical VQA task, namely Multi-Domain Lifelong VQA (MDL-VQA). To efficiently avoid forgetting problem in MDL-VQA benchmark, we further propose a Self-Critical Distillation (SCD) framework to allow the VQA model to introspect its learned knowledge and further reduce forgetting ratio while efficiently learning on new data. Extensive experiments demonstrate that our SCD can significantly improve the model's anti-forgetting ability and outperform other approaches by large margins on the MDL-VQA benchmark.

(6) After the research on lifelong VQA task over multi domains, we move our attention to the other real-world practical scenario, personalized federated learning (**RQ6**). To this end, we propose a new FedVQA task that trains VQA algorithms under the decentralized learning settings, where each local client learns from their

own private dataset represented by a specific visual scene. The purpose of the task is to train personalized models to perform well on the local data, while aggregating a generic global model to generalize well on the data in unseen visual scenes. To overcome the forgetting of global knowledge in FedVQA, we propose a new federated pairwise preference preserving approach, to flexibly yet effectively encourage the local models to review the useful global knowledge during the personalized training. To the best of our knowledge, this work is the first attempt to explore VQA algorithms under the personalized federated learning, which would facilitate more future works about privacy-based VQA research.

After discussing the findings described in different chapters, we attempt to conclude how VQA algorithms and multimodal understanding would be strengthened when comprehensively considering all research questions, and how the aforementioned contributions could come together to advance the field.

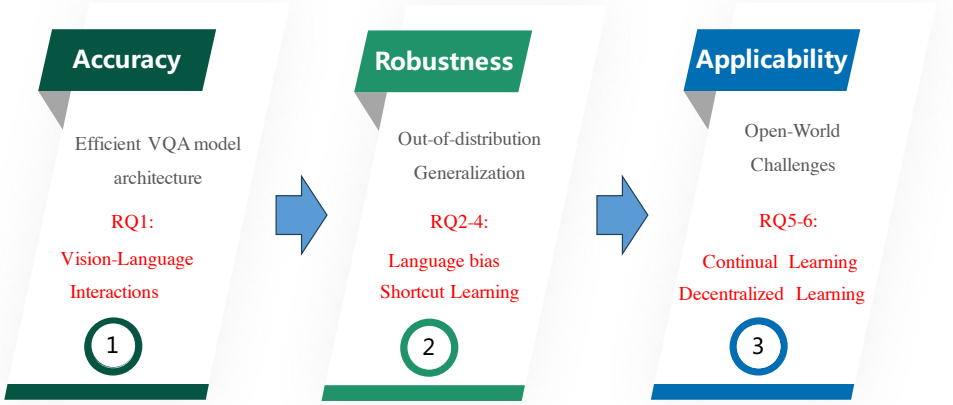

Figure 9.1: The Overview of how our contributions to different research questions can come together to advance VQA systems.

As illustrated in Fig. 9.1, all research questions in the thesis follow a three-step development strategy to improve VQA and multimodal understanding, which also reveals the research interests in the past 10 years. How to boost the model accuracy is the main focus in VQA research from the year of 2013 to 2019. Our proposed MHEF in **RQ1** focuses on the vision-language interactions, which contribute to build the efficient VQA model architecture. From 2019 to 2023, how to build robust VQA algorithms has become the crucial research content, where plenty of works pointed out that the language bias issue is severe in current VQA models. Then, our next 3 works for **RQ2-4** seek to alleviate data bias for multimodal inputs, thereby improving the out-of-distribution generalizability. Finally, our last two contributions (**RQ5-6**) concentrate on the applicability of VQA models from the aspects of continual and decentralized learning, which potentially turns to be the future research directions for open-world VQA systems.

9.2 Limitations

Our methods in this thesis have addressed seven research questions and achieved promising results in terms of three research themes. However, they still have some limitations which can be discussed from the following two perspectives.

9.2.1 Algorithmic perspective

In Chapter 3, we have discussed our motivation and analyzed insights about our proposed multi-stage fusion scheme MHEF. Nevertheless, we should realize that it still lacks of theoretical explanation to interpret the model's behaviors in different reasoning steps. If we can figure out some conclusions or proofs from the theoretical perspective, it will benefit designing concise and effective integration strategy to achieve fine-grained vision-language reasoning.

In Chapter 4 for the proposed LP-Focal loss, we introduce to use the fixed focusing parameter γ to smoothly adapts the rate at which the prior answer candidates are down weighted. Even though we conduct the detailed performance analysis to select the γ, the optimal γ settings for different questions or question categories should still be different. As a result, an instance-level focusing parameters are required to be proposed, and we assume the metric of information entropy may be the crucial to build such parameters.

In Chapter 4 and 5, both our proposed LP-Focal loss and LBCL approaches utilize the question-only branch to model the biased training distribution, and seek to decrease the influences from more-biased samples. One potential problem caused by this branch is the accuracy drop on in-distribution datasets, as the training of samples labeled by frequent answers may turn to be not sufficient. Recently, counterfactual sample augmentation has become popular in VQA research, and we assume this technique could be seamlessly integrated in our methods to further boost the in-distribution performance for unbiased VQA models.

In Chapter 6, the COCA strategy is specifically-designed for AVQA task involving multimodal reasoning over auditory, visual, and textual information. This may not be theoretically scheduled into other multimedia task that jointly understanding over two modalities (e.g. vision-language and audio-vision). In the future, we seek to reformulate our proposed COCA approach into a more general algorithm, which can be applied into any multimodal reasoning tasks under various shortcut biases.

In Chapter 7, our proposed Self-Critical Distillation (SCD) effectively alleviate the forgetting problem through logits- and feature-level knowledge distillation. However, the trade-off factors in the loss objective for dual-level distillation are still defined manually. Hence, there is still theoretical question about how to set the optimal weights to review logits- and feature-level knowledge. One potential solution is to quantify the difficulty for two anti-forgetting manners, and further propose

an instance-level trade-off factor that concentrates on recalling the hard knowledge.

9.2.2 Practical perspective

In Chapter 2, even though our MHEF obtain significant improvements over other competitive multimodal fusion strategies, the method may utilized relatively more computational resources, due to multi-stage learning and multi-space projection, especially compared with the simple linear fusion strategies. To solve it, a potential solution is to decrease the dimensions of embedding spaces when merging multimodal features.

In Chapter 5, we propose a novel Visual Sensitive Coefficient (VSC) metric to quantify the difficulty for the VQA model to exploit visual information for multimodal reasoning, without using visual annotations. However, the visual annotations such as the question based visual objects may be labeled by annotators in practice. As a result, there are still enough space to discover how to combine the VSC metric with the visual annotations to establish a efficient and robust curriculum metric, thereby facilitating the curriculum learning framework to alleviate language bias progressively.

In Chapter 6, our COCA regularization is established based on causal graph analysis on the advanced AVQA models, which assumes that the models are equipped with effective visual-spatial and audio-temporal attention mechanisms. As a results, COCA approach may not be suitable utilized into the simple-fusion AVQA models without attention mechanisms. Nevertheless, we think that it would not be a serious limitation to put COCA into practical scenarios, as the transformer-based attention model has become a indispensable components in current multimodal systems.

In Chapter 7, one possible drawback in our proposed Multi-Domain Lifelong VQA (MDL-VQA) benchmark is that the diversity of textual domains may not be sufficient. Even though MDL-VQA covers five different visual domains (e.g. realistic, abstract and artistic scenes), the questions across different datasets are still some general queries from low-level perception to high-level logical reasoning. In practice, the VQA systems may only focus on several specific functions in computer vision, such as visual counting and color identification. Moving forward, we attempt to introduce function-incremental setting into the MDL-VQA, so as to promote the setting to be closer to the real-world scenarios.

In Chapter 8, our proposed FedVQA task is based on the assumption that all the clients share the label space in the decentralized learning. However, in the practical scenario, it is sophisticated for different local clients to unify the answer candidates, since some answers existed in one specific scene (e.g. zebra, and tiger in natural scene) may not exist in the other irrelevant scenes (e.g. home and educational scenes). This poses a challenge to optimize a global classifier for the model in the

central server, and hinders the validation of model generalizability to the unseen visual scenes. One potential solution for the issue is to change the classification-based VQA algorithms into answer-generation based counterparts, instead of answer prediction according to classifier.

9.3 Future Research Directions

In the previous eight chapters, we have presented many methods to address the research questions regarding three research themes. A wide variety of future research is also encouraged to advance these themes. In this section, we briefly discuss future research directions regarding each theme.

9.3.1 VQA Benchmark beyond Accuracy

Current VQA benchmark typically evaluates the performance of VQA algorithm through the testing accuracy. However, due to the brittleness of VQA models toward robustness problem, such as language bias, multimodal inputs variations and sub-question consistency, the metric of predictive accuracy is not sufficient to demonstrate its authentic QA capacity. For instance, the VQA model suffers from language bias, and could easily yield remarkable performance on in-distribution (ID) dataset, which severely violates its desired multimodal reasoning behaviour. Even though on the out-of-distribution (OOD) dataset specialized for bias mitigation, the debiasing strategies can also leverage the 'inverse distribution' trick to significantly boost OOD performance at the cost of sacrificing the ID performance. As a result, we assume more comprehensive metrics beyond accuracy are needed to evaluate the authentic effectiveness of VQA models, especially from the aspect of model robustness. In the future, we plan to combine the accuracy and visual grounding annotations to improve the reliability of the VQA evaluation system.

9.3.2 Bias Mitigation in the Open-Set Learning Settings

How to mitigate data bias in multimodal question answering systems (e.g. VQA and AVQA) is a crucial topic in the thesis. Existing related work only focuses on the bias mitigation in the centralized training process, limited by a fixed and stationary dataset, while the similar issue in the open-set learning setting (e.g. lifelong learning, and federated learning) in VQA task is almost ignored. First, we assume the challenge of data bias problem for VQA models in sequential learning or decentralized training setting would be more complicate yet practical. Second, the debiasing strategies under this setting should simultaneously and collaboratively consider the tasks of anti-forgetting and anti-bias. We believe these approaches would provide important scientific contribution to the both theoretical and multimedia research community, and also promote the development of multimodal intelligent systems.

9.3.3 Federated Lifelong Visual Question Answering

One often overlooked property of VQA systems is the human-computer interaction, where the training samples are usually available sequentially due to users' response. The input images and questions typically involve the private information from users, and the training data. Therefore, in practice, we are not allowed to merge all the private samples together to form a large-scale dataset for the centralized training of VQA machines. This is in contrast to what is assumed in theoretical research. Federated Learning [372, 373] is a distributed learning approach that allows multiple devices to collaboratively train machine learning models without sharing their data, while Lifelong Learning [119, 317, 318] allows VQA models to accumulate knowledge from sequentially-arrived tasks. Research on the combination of federated and lifelong learning in VQA task, also known as Federated Lifelong Visual Question Answering, can have significant applications for privacy, scalability, and generalization. However, to our best knowledge, there is still no benchmarks specialized for such types of VQA task. Moving forward, we plan to build a federated lifelong VQA task to fill the vacancy, and enable VQA systems to achieve decentralized training over continually learned tasks over different multimodal content.