Exploring deep learning for multimodal understanding

Lao, M.

English Summary

We are living in a new media age where the quantity of vision and language data increases exponentially. Therefore, it is important to design intelligent multimodal reasoning systems to satisfy our need for understanding multimedia data. A prime example of such a system, which has garnered significant attention in the context of question-answering (QA) scenarios, is an automated Visual Question Answering (VQA) system. A VQA system concentrates on answering textual questions, according to images from users. This can be applied in various practical scenarios, such as robot tutors and smart-home management systems. To improve the performance as well as robustness of multimodal QA systems, and further enable them to meet the open-world challenges, we explore five themes for efficient and practical VQA models.

The first theme focuses on efficient feature fusion for vision-language inputs, which plays a crucial role on improving the predicting accuracy of current VQA models. Motivated by the multi-stage reasoning behaviour in human beings, we attempt to achieve multiple and fine-grained multimodal interactions for enhancing fusion performance. To this end, we introduce a novel Multi-stage Hybrid Embedding Fusion (MHEF) network to improve multimodal fusion from the aspects of dual-space embedding and multi-stage learning. With multi-stage MHEF, we can not only improve the performance in each single stage, but also obtain additional accuracy improvements by integrating all prediction results from each stage.

The second theme of this thesis is to study the language bias in current multimodal reasoning algorithms, which acts as a prime impediment to model robustness and reliability in VQA task. First, to reduce the overdependence on dominant answers for different question types, we introduce a Language Prior based Focal Loss (LP-Focal Loss) by rescaling the standard cross entropy loss in VQA training, which dynamically assigns lower weights to biased answers when computing the training loss, thereby reducing the contribution of more-biased instances in imbalanced training data. Second, based on the fact that VQA samples with different levels of language bias contribute differently for answer prediction, we propose a novel Language Bias driven Curriculum Learning (LBCL) framework to overcomes the language bias via an easy-to-hard learning strategy. Specifically, in the initial training stage, the VQA

model mainly learns the superficial textual correlations between questions and answers (easy concept) from more-biased examples, and then progressively focuses on learning the multimodal reasoning (hard concept) from less-biased examples in the following stages.

In the third theme, we turn to consider and analyze the bias problem in Audio-Visual Question Answering (AVQA), which is a more sophisticated question answering task involved textual-visual-auditory information for multimodal understanding. Through detailed causal-graph analyses and careful inspections of their learning processes, we reveal that AVQA models are not only prone to over-exploit 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 mitigate this issue, we propose a COllabrative CAusal (COCA) Regularization to remedy this from two aspects. First, a novel Bias-centered Causal Regularization (BCR) is proposed to alleviate specific shortcut biases by intervening bias-irrelevant causal effects, and further introspect the predictions of AVQA models in counterfactual and factual scenarios. Second, based on the fact that the dominated bias impairing model robustness for different samples tends to be different, we introduce a Multi-shortcut Collaborative Debiasing (MCD) to measure how each sample suffers from different biases, and dynamically adjust their debiasing concentration to different shortcut correlations.

The fourth theme is to address the lifelong learning challenges that the VQA faces in an open world, because VQA systems are always supposed to extend their knowledge and meet the ever-changing demands of users. Therefore we introduce a new VQA setting called Multi-Domain Lifelong VQA (MDL-VQA), which aims to develop a VQA system that can continuously learn in order to address new domains from new data while preserving knowledge learned from previous domains. To address these challenges, we propose a novel Self-Critical Distillation (SCD) framework for MDL-VQA, which alleviates forgetting issue via transferring previous-domain knowledge from teacher to student models. Specifically, it contains logits- and feature-level distillations to promote teacher to transfer informative old knowledge, and meanwhile facilitate student to acquire helpful knowledge in current domain. Through blending such dual-level distillations, SCD enhances the VQA model's stability to anti-forgetting while keeping its plasticity to learn newly coming knowledge.

For the last theme, we focus on a practical personalized federated learning scenario, where the training of VQA models is across multiple clients or servers, while keeping data localized, without centralizing it. To facilitate the research, we propose a FedVQA setting to train personalized VQA client models for distinct visual scenes, while optimizing a generic model to generalize well on unseen scenes, through client collaboration under the privacy constraint. To address the generic knowledge forgetting issue in FedVQA, we propose a novel federated pairwise preference preserving (FedP^3) framework to improve personalized learning. Specifically, we first

design a differentiable pairwise preference (DPP) to improve knowledge preserving by formulating a flexible yet effective global knowledge. Then, we introduce a forgotten-knowledge filter (FKF) to encourage the client models to selectively consolidate easily-forgotten knowledge. By aggregating the DPP and the FKF, FedP^3 coordinates the generic and the personalized knowledge to enhance the personalized ability of clients and generalizability of the server.

We conducted extensive experiments to verify the efficacy of the proposed approaches for the five themes. The results demonstrate remarkable improvements over various baselines and state-of-the-art methods. Thus this thesis provides important novel contributions, insights, and findings for the research community and future applications in the field of multimodal understanding and VQA.