Exploring deep learning for multimodal understanding

Lao, M.

Chapter 1

Introduction

1.1 Motivation

Over recent years, artificial intelligence [1], and especially by deep learning [2, 3] have witnessed high-speed development. Due to the significant increase in data collection and storage capabilities, researchers have made remarkable progress in various practical areas of machine intelligence. Deep learning has resulted in major advances in computer vision. For example, intelligent visual systems have outperformed humans in image classification [4, 5], instance segmentation [6, 7] and object detection [8, 9]. Moreover, with the widespread utilization of word embedding [10, 11] and large-scale pre-training models [12, 13], deep learning has also advanced the field of natural language processing [14, 15], such as text classification [16, 17], and machine reading comprehension [18, 19]. Due to the increasing amount of multimedia data (e.g. image, text, and audio) on the internet as well as on social networks, intelligent machines are expected to automatically process and understand multimodal (e.g. visual, textual and auditory) information, especially for the tasks at the intersection of computer vision and natural language processing.

Early multimodal learning tasks [20, 21] are image-text retrieval [22, 23] and image captioning[24, 25]. The former aims at retrieving the images/text associated with the textual/visual queries, while the purpose of the latter is to generate textual descriptions of given images. Another prevalent vision-language tasks includes visual grounding [26, 27] and text-based image generation [28]. However, we assume that the aforementioned tasks may fail to reveal and evaluate the authentic capacity of multimodal understanding for intelligent machine. The potential reasons are concluded from two aspects: First, The tasks only achieve the multimodal information processing by converting from one modality to the other through retrieval or generation, and may not truly understand both visual or textual information. Furthermore, The necessary abilities to solve these tasks are more about the low-level perceptions, such as the recognition of existence, properties and attributes in

image and text data. High-level understanding, however, (e.g. logical and commonsense reasoning) may be neglected. This plays a crucial role on putting multimodal learning into practice.

To address this problem and to achieve joint multimodal understanding [29], Visual Question Answering (VQA) [30, 31, 32] has become an urgently needed yet challenging task in the past few years. An automated VQA system intends to provide the correct answer for a natural language question asked about an input image. The most unique property of the VQA task is the diversity of to-be-achieved functions based on different question inputs. As illustrated in Fig. 1.1, the VQA system is expected to cope with various sub-tasks in computer vision, from low-level perception (object recognition/detection, scene/attribute classification, and counting) to high-level reasoning (commonsense and knowledge -based reasoning). It requires fine-grained recognition for both visual and textual content,and comprehensive reasoning over multimodal features to deduce the correct answers.

Figure 1.1: Diverse question-image pairs as the training samples in VQA task, which covers different sub-tasks in computer vision.

VQA is an important field of research both for theoretical and practical aspects: 1) Firstly, VQA research can be exploited as an alternative to the visual turing test [33], since a well-performed VQA machine is expected to tackle various sub-tasks in computer vision according to different textual queries from users. Secondly, the learning framework in VQA tasks follows a general multimodal understanding process: (1) uni-modal feature representations [34], (2) multimodal attention [35] and (3) fusion [36]. Therefore, its advances in model architectures and training strategies shed the light on other research vision-language tasks [37, 38]. Meanwhile, VQA covers many sub-tasks in computer vision from low-level perception to high-

level logical reasoning, and acts as an indispensable downstream task to evaluate the large-scale vision-language pretraining models [39, 40].

2) Practice: One of the most intuitive applications of VQA is to assist visually-impaired individuals by providing a VQA system [41, 42] that can analyze images and answer various questions, thereby improving their daily lives. Moreover, VQA technology can improve human-computer interaction in different application scenarios, including robot tutors [43], smart home management systems [44], automated driving systems [45], and medical diagnosis [46]. Recently, the widespread use of ChatGPT [47], has made the VQA task even more crucial as a key technology to enhance the precision and dependability of ChatGPT's question-answering abilities.

1.2 Background and Related Work

As one of the important multimodal understanding applications, Visual Question Answering (VQA) has been widely studied in the past few years. Dozens of datasets specialized for VQA tasks have been proposed, whose questions involved various functions (sub-tasks in computer vision) requiring different levels of reasoning abilities. The VQA v1/v2 dataset [30, 48] is a widely-used dataset, where the images are the realistic pictures obtained from the MSCOCO dataset [25]. Also based on high-quality real-world images, GQA [49] is considered as a more advanced dataset, which involves extensive annotative information (e.g. scene graphs [50]) as well as evaluation metrics (e.g. consistency and plausibility). Moreover, through using photos taken by visually-impaired people, Vizwiz dataset [51] is introduced to assist blind people for their proposed low-level perception questions. Other popular datasets for question answering about realistic images are COCO-VQA [52], DAQUAR [53], Visual7W [54] and Visual Genome [55]. Furthermore, the VQA-abstract [56] dataset attempts to evaluate VQA models over questions related to the abstract cartoon images, while CLEVR [57] is a diagnostic dataset that tests a range of visual reasoning abilities according to the synthetic pictures.

In this thesis, we discuss the related works of VQA tasks into five research directions:

- How to design an efficient VQA model architecture?
- How to improve robustness of a VQA algorithm against language bias?
- How to solve the VQA problem involving additional auditory information?
- How to address a continuous steam of data?
- How to train models across multiple decentralized devices over privacy-sensitive data?

1.2.1 Multimodal Attention & Fusion

The common processing pipeline of VQA models typically consists of the multimodal representation, attention, fusion, and answer prediction modules, as depicted in Fig. 1.2. Specifically, the multimodal representation module aims to extract image and question features via pretrained visual [58, 59] and textual [60, 61, 62] models. Answer prediction projects the final fusion feature into label space to select correct answers. To improve the performance of VQA models in current datasets, most related work focuses on improving the multimodal attention and fusion schemes to achieve fine-grained and efficient vision-language interaction.

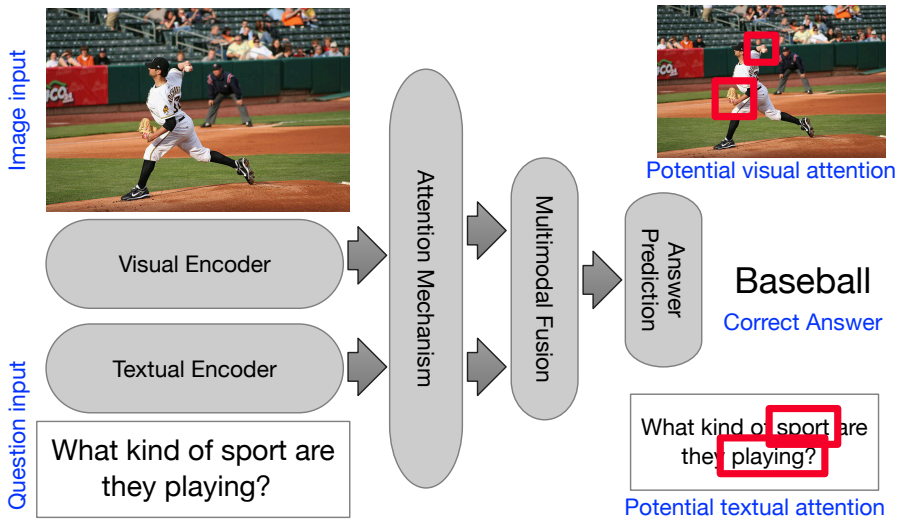

Figure 1.2: The common pipeline of VQA algorithm, which typically contains multimodal encoding, attention, fusion and answer prediction modules.

The attention mechanism [63, 64, 65] in VQA task mimics human cognition by selectively concentrating on question-related visual regions or objects. This provides informative and better visual and textual representations. Yang et. al [66] present a stacked attention network to update the attended image regions using multi-stage learning. Not limited to visual attention, Lu et. al [67] propose a co-attention model to simultaneously focus on important visual and textual information. Kim et. al introduce [68] a bilinear attention network to achieve fine-grained and high-order interactions between multimodal inputs. Li et. al [69] present a relation-aware graph attention network to encode images into graphs, and build multi-type inter-object relations via graph attention mechanism. Another state-of-the-art attention techniques include structured attention [70], Human-annotative attention [71], and multimodal relation attention [72], which significantly enhance the performance as well as visual explainability of current VQA models. With the fast development of transformer originally proposed in machine translation task, it has become the mainstream yet well-performed attention mechanism in the current state-of-the-art VQA models [73, 74, 75].

The other research hotspot on the VQA model architecture side is multimodal fusion, which has a fundamental role in the integration of both visual and textual features for accurate answer prediction. Early methods [76, 77] typically use simple linear fusion schemes to merge image and question inputs, such as concatenation, summation and multiplication. For achieving high-order interactions between multimodal features, bilinear pooling [78, 79] has been considered as an effective way to fuse information from two sources, since it can take all pairwise interactions among given features into consideration. However, due to the fact that standard bilinear pooling employs the outer-product operation and consumes a large number of parameters, advanced bilinear pooling methods (MCB [80], MLB [81], MFB [82], MUTAN [83] and MFH [84]) are introduced to reduce the utilization of computational resources, and to enhance the predictive accuracy.

In this thesis, we seek to improve the VQA model architecture by designing an efficient multimodal fusion scheme (Chapter 3). Specifically, we propose a novel Multi-stage Hybrid Embedding Fusion (MHEF) network to achieve multi-stage information fusion in hybrid feature spaces, which achieve superior performance compared to existing multimodal fusion approaches.

1.2.2 Language Bias Problem

With the rapid development of attention mechanisms and multi-modal fusion techniques, VQA models have shown high performance on in-distribution datasets [30, 48]. However, many researches [32, 85] pointed out that most models encounter the unwanted shortcut 'language bias', where they are prone to over-reliance on the superficial correlations between question patterns and frequent answers, thereby neglecting the fine-grained analysis of visual information. This undesirable behavior causes VQA models to fail to be robust against label distribution shift, and perform worse on out-of-distribution datasets [85, 86, 87].

Recently, a variety of de-biasing strategies [88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99] are proposed to alleviate this issue. We roughly divided them into three categories: 1) Anti-bias models seek to build efficient yet practical learning modules to alleviate the data bias. Some related works [85, 100, 101] rectify the question encoding process into several reasoning stage, so as to overcome inherent language prior in the textual representation. The other alternatives [102, 103] focus on improving the process of answer prediction into answer re-ranking module, thereby leveraging the semantic of answer embedding to facilitate VQA model to capture unbiased information. 2) Human-annotation methodologies tend to utilize the human-annotative information to encourage the VQA model to distinguish the unbiased information (e.g. question related visual concepts) from the biased counterparts (e.g. question patterns). The HINT [92] approach leverages human visual attention collected from VQA-HAT [104], and proposes to align the to-be-trained visual attention maps in

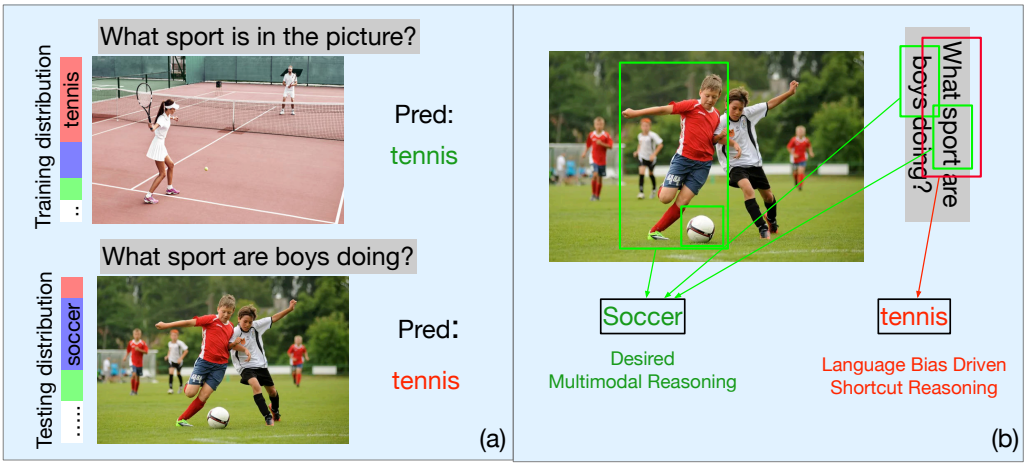

Figure 1.3: The conceptual illustrations of the language bias issue in VQA task. (a) The different label distributions between training and testing splits is the potential cause to language bias. (b) The reasoning behaviours for desired multimodal learning and question-type based shortcut learning.

VQA model and the human-annotated counterparts. SCR [105] use the explanations of ground-truth answers to equips VQA models with the self-critical ability via a matching regularization, which promotes the most attentive image regions to be consistent with the explanations of ground-truth answers. 3) Non-annotation methodologies attempt to propose some efficient debiasing strategies to reduce bias, without any extra information. Some typical works include ensemble-based methods [106, 107, 108], adversarial regularization [89], causal inference [88], balanced loss function [90, 109], and out-of-distribution sample generation [94, 110, 111]. We conclude the detailed taxonomy of bias mitigation techniques and related works involved in VQA robustness in the Chapter 2.

In the thesis, we focus on the more flexible non-annotation based approaches to overcome the language bias, and improve out-of-distribution generalization of VQA models. We propose a simple yet effective Language Prior based Focal (LP-Focal) loss, which dynamically assigns weights for samples answered by different questions. To further improve the out-of-distribution generalization of VQA models, we introduce a new Bias driven Curriculum Learning (LBCL) approach, which follows an easy-to-hard strategy to alleviate language bias progressively.

1.2.3 Audio Visual Question Answering

With the increasing amount of auditory information on the Internet, multimodal systems only specialized for visual-textual inputs may not sufficient. By involving the audio input into the VQA task, the Audio-Visual Question Answering (AVQA) [112, 113, 114] is expected to answer natural language question for the audio-visual information from a stream of video, which has been an emerging research topic

in recent years. Beyond the VQA tasks that concern vision-language understanding, AVQA requires models to comprehensively understand three modalities and perform spatial-temporal reasoning over audio-visual scenes. Particularly, for the audio-visual questions in AVQA task, without considering either visual or auditory information would fail to deduce correct answers reasonably. The illustrated learning framework in the AVQA task is depicted in the Fig. 1.4.

Figure 1.4: AVQA requires a comprehensive multimodal reasoning to answer the textual question over video-audio information.

In the past few years, many well-established datasets [112, 113, 114] designed for AVQA have been proposed. A novel benchmark named Pano-AVQA [113] has been developed as a large-scale grounded audio-visual question answering dataset on panoramic videos. Specifically, Pano-AVQA collects 5.4K 360deg online video clips, and further organizes samples into the questions about spherical spatial relations and audio-visual relation. Considering that musical performance is an important multimodal scene with auditory-visual interactions, Li et. al [114] introduce a large-scale Spatio-Temporal Music AVQA (MUSIC-AVQA) dataset. It consists of 9k videos covering 22 instruments, accompanied with diverse question types established from 33 templates. Beyond Audio-visual question answering in specific scenarios, Yang et. al [112] propose a large-scale AVQA dataset with 5.7K videos derived from daily audio-visual activities. Moreover, it contains specially-designed questions requiring both video and audio clues for answer prediction, where information contained in a single modality is insufficient or ambiguous. To improve the performance of the AVQA model, Zhuang et al. [115] conduct early fusion between visual and auditory modalities, and then introduce an attention memory unit to enrich the multimodal features. Miyanishi et al. [116] jointly consider the motion, appearance, and audio information as the input, and introduce a modulated multi-stream 3D ConvNets to achieve fine-grained multimodal reasoning. Inspired by the recent sucecess of transformer models, Yun et al. [113] propose a transformer-based model to encode multimodal features. To emphasize attention mechanism in both visual and audio inputs, Li et al. [114] introduce spatial and temporal grounding modules to reason spatio-temporal associations between audio and visual modalities under a question query.

Unlike the aforementioned works, we focus on whether AVQA suffers from a similar bias problem to impair its model robustness. To this end, we analyze the potential biases in AVQA task from the perspective of causal graph, and further reveal the multi-shortcut biases problem in this task. Furthermore, we propose a backbone-agnostic COCA approach to cooperatively alleviate different biases with an instance-aware manner, and enhance the multimodal reasoning capacity of AVQA models (Chapter 6).

1.2.4 Lifelong Learning

Lifelong learning [117, 118, 119], also known as continual learning, has been widely studied in computer vision and natural language processing tasks. Research in this area can be broadly categorized into two main settings: 1) class- or task-incremental learning [120, 121, 122] and domain-incremental learning [123, 124, 125]. In the former, models must learn to classify an increasing number of classes sequentially from a single domain, while in the latter, models learn to solve tasks that cross different domains while sharing the same label space. The main challenge in lifelong learning is catastrophic forgetting, which occurs when a model's performance on previous tasks deteriorates after training on new tasks. To alleviate this issue, the mainstream strategies can be divised into three types: 1) rehearsal methods [126, 127, 128], which retrain on a limited subset of stored samples while training on new tasks. 2) parameter isolation methods [129, 130], which assign new branches with different model parameters for new tasks while freezing previous task parameters. 3) regularization-based methods [131, 132, 133], which incorporate extra regularization into the loss function to solidify previous knowledge when learning on new data.

Inspired by the significant progress in vision-language learning, several works explore lifelong learning in the perceptual-level multimodal tasks, such as cross-modal retrieval [134] and image captioning [135, 136]. For the VQA tasks that requires both perceptual- and reasoning-level understanding, [137] is the first attempt to exploit a simple class-incremental learning setting for lifelong VQA, where samples in question types '*wh-*' and '*yes/no*' are tested under different sequence. [138] Proposes a CLOVE benchmark to establish the scene- and function-incremental learning through splitting the GQA dataset in natural visual domain, and mitigate the forgetting problem by replaying scene graphs. Moreover, Srinivasan et al. [139] introduces an attractive CLiMB benchmark, where models are expected to continually learn crossing different multi-modal reasoning tasks, including VQA.

In contrast to the existing lifelong learning in multimodal tasks, we focus on the continual tasks under multi-domain learning, and further propose a novel yet practical VQA task, namely Multi-Domain Lifelong VQA (MDL-VQA). This task requires VQA models to accumulate informative knowledge from sequentially-arrived domains, while also alleviating catastrophic forgetting problem for learned knowledge

in previously seen domains. Furthermore, we propose a novel Self-Critical Distillation (SCD) framework specifically designed for MDL-VQA to overcome the forgetting issue without data storage (Chapter 7).

1.2.5 Federated Learning

One often overlooked property of VQA systems is the human-computer interaction, where the input images and questions typically involve the private information from users. Therefore, in practice, we are usually not allowed to merge all the private samples together to form a large-scale dataset for the training of VQA machines. Federated Learning (FL) is a machine learning approach that enables training models across multiple decentralized devices or servers while keeping the data locally stored and not transferring it to a central server [140, 141]. The most widely adopted FL algorithm is FedAvg [142], which averages weight parameters across local models trained on private client datasets to learn a global model. Recent research efforts have focused on improving FedAvg from various perspectives, including model convergence [143, 144], robustness [145], communication [146], and non-IID clients [147, 148].

To further handle the heterogeneity of data and models, personalized FL (PFL) has been introduced [149]. In contrast to traditional FL, PFL aims to learn a customized model for each client, tailored to their specific objectives. This method acknowledges the diversity of data among clients by constructing a "personalized" model that fits each client's needs. One group of techniques [148, 150] has leveraged multi-task learning (MTL) methods to incorporate clients' task objectives into the FL framework. The other group contains post-processing techniques [151, 152]. [152] uses meta-learning to learn an initial model that can be adapted to each client through local fine-tuning. [152] Indicates that fine-tuning can achieve comparable results to other personalized methods. In our work about federated learning in VQA task (Chapter 8), we use an MTL-based approach that can optimize generic and personalized VQA models simultaneously. While the benchmarks for conventional FL are well-established, few studies have focused on federated VQA. The most closely related work [153] proposes a vision-and-language FL framework with shareable networks, but only considers the scenario where clients learn different tasks (e.g., VQA and image captioning) rather than personalized federated VQA across different scenes.

We argue that proposing a Federated VQA task (FedVQA) is a practical and challenging task for two reasons. First, FedVQA not only aims to improve individual personalized models through collaborative training, but also considers the model's ability to directly deploy on unseen scenes. Second, since the heterogeneous data collected from different scenes include scene-specific characteristics (e.g., distinct high-frequency words), the model trained on our FedVQA has a high risk of failing

to converge. To the best of our knowledge, this work is the first attempt to explore VQA tasks in personalized federated learning (Chapter 8).

1.3 Thesis Outline and Research Questions

In Section 1.2, we have briefly reviewed the recent progress from four research directions *Robust Visual Question Answering* has become an increasingly prevalent and significant topic in the VQA research. This field is also strongly related to the main contributions of the thesis. For a better understanding of robust VQA, we first present a comprehensive review for existing works related to robust VQA. Our review includes the main challenges, benchmarks and strategies specialized for model robustness in VQA task. Chapter 2 is based on the submitted manuscript:

> **Lao, M.**, Pu, N., and Lew, M. S. "Robust Visual Question Answering: Challenges, Benchmarks and Strategies." Submitted to ACM Transactions on Information Systems, 2023.

Although deep learning is leading the state-of-the-art performance on numerous VQA benchmarks, we should notice its limitations and challenges, such as multimodal fusion, language bias, and training with limited data or stream data, etc. There is still considerable space for promoting the developments of deep learning. In the following research chapters (Chapters 3, 4, 5, 6, 7, 8), we propose new approaches to address research questions (**RQ**) and challenges in terms of the four research directions describe in section. In Chapter 9, we discuss our main findings, limitations, possible solutions and future research themes.

- In Chapter 3, we focus on the multimodal fusion techniques in the VQA learning frameworks. Their purpose is to incorporate image and question features and to generate integrated visual-textual features for answer prediction. A well-designed multimodal fusion strategy could effectively facilitate the joint understanding and reasoning over vison-language information, and improve the performance of answer prediction. We describe two properties for existing multimodal fusion methods in current VQA research: 1) *Latent Space Embedding*: most methods rely on latent embedding with two branch networks in which the visual and textual features are embedded into a common latent space, and then using operations such as multiplication summation to fuse them. 2) *Single interaction*: these methods typically merge extracted image and question features via single-step fusion, and employ the merged feature to predict answers. We assume current fusion approaches are not sufficient to achieve fine-grained and effective multimodal joint reasoning, due to the restrictions derived from by two aforementioned properties. Thus, the condition

gives rise to the first research question **RQ 1: How can we build an efficient multimodal fusion method for vision-language understanding?** In this Chapter, we introduce a novel Multi-stage Hybrid Embedding Fusion (MHEF) network to enhance the multimodal fusion process from two aspects: First, we introduce a Dual Embedding Fusion (DEF) approach that transforms one modal input into the reciprocal embedding space before integration. The DEF is further incorporated with the Latent Embedding Fusion to form a novel Hybrid Embedding Fusion (HEF). Second, we design a Multi-stage Fusion Structure (MFS) for the HEF to form the MHEF network, so as to obtain diverse and better fusion features for answer prediction. By jointly training the multi-stage framework, 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. This work is based on the published journal paper:

> **Lao, M.**, Guo, Y., Pu, N., Chen, W., Liu, Y. and Lew, M. S. "Multi-Stage Hybrid Embedding Fusion Network for Visual Question Answering." Neurocomputing, 2021.

- In addtion to pursuing state-of-the-art performance via designing efficient model architecture, the robustness problem has become a more crucial research topic in the VQA community. One unavoidable problem related to model robustness for VQA task is *Language Bias*. Specifically, current VQA models are prone to over-reliance on the fallacious correlations between question patterns and frequent answers, thereby neglecting the fine-grained analysis of visual information. This undesirable behavior severely impairs the reliable and interpretability of model accuracy, and further limits their applications in real-world scenarios. To alleviate this problem, an improved loss function tailored to language bias may be an intuitive yet effective solution. Thus, we come to the second research question **RQ 2: How to find a model-agnostic loss function to reduce language bias for current VQA learning frameworks?** Inspired by the focal loss exploited in imbalanced object detection, we propose a novel Language Prior based Focal Loss (LP-Focal Loss) by rescaling the standard cross entropy loss. Specifically, we employ a question-only branch to capture the language biases for each answer candidate based on the corresponding question input. Then, the LP-Focal Loss dynamically assigns lower weights to biased answers when computing the training loss, thereby reducing the contribution of more-biased instances in the train split. In Chapter 4 is based on the published conference paper:

Lao, M., Guo, Y., Liu, Y. and Lew, M. S. "A Language Prior based Focal Loss for Visual Question Answering." IEEE International Conference on Multimedia and Expo, 2021.

- In Chapter 5, we conduct an in-depth investigation of the language bias problem exited in current VQA models. To examine the effectiveness of debiasing strategies, VQA models are typically required to be evaluated in the out-of-distribution (OOD) setting, where the the train and test splits possess entirely different label distribution. To achieve state-of-the-art performance on the OOD generalization for bias mitigation, we assume the loss-function based strategies discussed in Chapter 2 may not be sufficient to tackle this issue. Therefore, we move to the third research question **RQ 3: How to design a VQA learning architecture specialized for language bias to improve out-of-distribution performance?** Based on the fact that VQA samples with different levels of language bias contribute differently for answer prediction, in Chapter 5, we overcome the language bias problem by proposing a novel Language Bias driven Curriculum Learning (LBCL) approach, which employs an easy-to-hard learning strategy with a novel difficulty metric Visual Sensitive Coefficient (VSC). Furthermore, to avoid the catastrophic forgetting of the learned concept during the multi-stage learning procedure, we propose to integrate knowledge distillation into the curriculum learning framework. This work is according to the published conference paper:

  Lao, M., Guo, Y., Liu, Y., Chen, W., Pu, N., and Lew, M. S. "From Superficial to Deep: Language Bias driven Curriculum Learning for Visual Question Answering." ACM International Conference on Multimedia, 2021.

- In Chapter 6, we aim to investigate the more complicated Audio-Visual Question Answering (AVQA) task. Here, the purpose is to answer textual questions over given video-audio pairs with comprehensive multimodal resoning. Compared with the VQA task, AVQA additionally involves auditory information from the corresponding video. With highly-developed deep learning models, current AVQA models are capable to achieve remarkable accuracy on existing datasets. We turn to focus on the robustness problem of AVQA models, especially the data bias problem. The research question in this chapter is **RQ 4: Do AVQA models suffer from data bias problem similar to VQA models? If so, how can we mitigate the corresponding biases to improve model robustness?** To this end, we first conduct detailed causal-graph analyses and modality-removal experiments, 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. In

this chapter, we propose a COllabrative CAusal (COCA) regularization to remedy this more challenging issue of data biases. Specifically, a novel Bias-centered Causal Regularization (BCR) is proposed to alleviate specific shortcut biases by conducting factual and counterfactual regularization. Moreover, a Multi-shortcut Collaborative Debiasing (MCD) is introduced to measure how each sample suffers from different biases, and dynamically adjust their debiasing concentration to different shortcut correlations. This chapter is based on the published conference paper:

> **Lao, M.**, Pu, N., Liu, Y., He, K., Bakker E.M., and Lew, M. S. "COCA: Collaborative Causal Regularization for Visual Question Answering" Association for the Advancement of Artificial Intelligence, 2023

- In Chapter 7, we move our attention to put current VQA research into more practical scenarios. Apart from the tremendous success of current VQA modes, their training process always learns through a stationary domain that is fixed by the choice of a given dataset. However, this limitation violates many practical scenarios where the data is continuously increasing from different domains. In real-world applications, the VQA systems are always expected to constantly acquire and update their knowledge, thereby catering to the evolving demands from users. Hence, we further propose the fifth question **RQ 5: How can we build a lifelong learning VQA system?** To address this question, we introduce a new and challenging multi-domain lifelong VQA task, dubbed MDL- VQA, which enables the VQA model to continuously learn across multiple domains while mitigating the forgetting on previously-learned domains. Furthermore, we propose a novel replay-free Self-Critical Distillation (SCD) framework tailor-made for MDL-VQA, which enhances the VQA model's stability to anti-forgetting while keeping its flexibity to learn newly-coming knowledge. This chapter is based on the published conference paper:

  > **Lao, M.**, Pu, N., Liu, Y., Zhong, Z., Bakker E.M., Sebe, N. and Lew, M. S. "Multi-Domain Lifelong Visual Question Answering via Self-Critical Distillation." ACM International Conference on Multimedia, 2023.

- In Chapter 8, we turn to concentrate on an other practical open-world problem: personalized federated learning. Even though current VQA models achieve significant progress on centralized training with state-of-the-art model architectures, the use of such training paradigms poses a significant challenge to privacy constraints in practical VQA applications. Therefore, our last research point seeking to answer **RQ6: How to put VQA algorithms into the**

decentralized learning scenarios with privacy constraints? First, to facilitate such types of research, we present a challenging yet practical Personalized Federated VQA task, named FedVQA. The goal of FedVQA is to train personalized VQA client models for distinct visual scenes, while optimizing a generic model to generalize well to unseen scenes, through client collaboration under the privacy constraint. Second, to address the challenge, we propose a novel Federated Pairwise Preference Preserving (FedP^3) framework to improve both generic (central server) and personalized (local clients) learning via preserving generic knowledge under FedVQA constraints. This chapter is based on the published conference paper:

> **Lao, M.**, Pu, N., Zhong, Z., Sebe, N. and Lew, M. S. "FedVQA: Personalized Federated Visual Question Answering over Heterogeneous Scenes." ACM International Conference on Multimedia, 2023.

- Finally, Chapter 9 summarizes the main findings from the research of this thesis. Also, we discuss the limitations and potential solutions, and point out directions for future research.

Additionally, this thesis draws on insights and experiences from related work in other publications during my PhD studies:

- **Lao, M.**, Guo, Y., Chen, W., Pu, N., Lew, M. S., "VQA-BC: Robust Visual Question Answering via Bidirectional Chaining.", International Conference on Acoustics, Speech and Signal Processing, 2022.
- He, K., Pu, N., **Lao, M.**, Bakker, E. M., and Lew, M. S. "Dual Selective Knowledge Transfer for Few-Shot Classification." Applied Intelligence, 2023.
- He, K., Pu, N., **Lao, M.**, and Lew, M. S. "Few-shot and meta-learning methods for image understanding: a survey." International Journal of Multimedia Information Retrieval, 2023.
- Chen, W., Xu, H., Pu, N., and Liu, Y., **Lao, M.,** and Liu, L., Wang, W., Lew, M. S., "Lifelong Fine-grained Image Retrieval." IEEE Transactions on Multimedia, 2022.

1.4 Main Contributions

The contributions of this thesis can be summarized and characterized as: 1) Task Contribution; 2) Technical Contribution;

Figure 1.5: Our MDL-VQA benchmark involves five VQA tasks in the visual domains of composite, realistic, blurry-objects, artistic and abstract scenes. Unlike standard domain-incremental learning, the label spaces for different domains are inconsistent, where words in red shading denote some general answers coexist in several domains, and in yellow shading are domain specific.

1.4.1 Task Contribution

A new and challenging VQA task in lifelong learning. As illustrated in Fig. 1.5, we propose a new and challenging VQA task, namely Multi-Domain Lifelong Visual Question Answering (MDL-VQA) This task requires VQA models to accumulate informative knowledge from sequentially-arrived domains, while also alleviating the catastrophic forgetting problem for learned knowledge in previously seen domains. The comparison among MDL-VQA benchmark and various VQA settings are illustrated in Tab. 1.1. On the one hand, beyond the stationary settings in Tab. 1.1 (e.g., fully-supervised, unsupervised domain adaptation), MDL-VQA requires VQA models to learn from sequentially-arrived tasks, which not only considers the adaptation performance on the new domain, but also attempts to mitigate the forgetting from previous domains. Thus, MDL-VQA is a more challenging yet more practical setup towards satisfying the demands of real-world VQA than existing relevant VQA setups. On the other hand, in contrast to conventional lifelong learning

Table 1.1: The comparison of existing VQA setups, including fully-supervised learning (FSL), unsupervised domain adaption (UDA), class-incremental learning (CIL), domain-incremental learning generalization (DIL), and our multi-domain lifelong learning (MDL-VQA). $\mathcal{S}$ and $\mathcal{T}$ are the source domain and the target domain, respectively. The superscripts "tr" and "te" indicate the train split and the test split, respectively.

VQA Setup	Number of Train Steps	Train Data	Same Visual Domain	Same Label Space
FSL	One	$\mathcal{S}^{tr}$	✓	✓
UDA	One or Two	$\mathcal{S}^{tr}$ and $\mathcal{T}^{tr}$	×	✓
CIL	Multiple	$\mathcal{S}_i^{tr}, 1 \leq i \leq t$	✓	×
DIL	Multiple	$\mathcal{S}_i^{tr}, 1 \leq i \leq t$	×	✓
MDL-VQA	Multiple	$\mathcal{S}_i^{tr}, 1 \leq i \leq t$	×	×

tasks (e.g., class-incremental learning) that focus on the performance on seen classes

in a single domain, the training samples in different tasks are represented in different visual and textual domains. Moreover, it may not be considered as a standard domain-incremental learning setting, since the label spaces are not consistent across different domains. We find that the challenges of MDL-VQA are three-folds:

1) severe domain shift: as depicted in Fig. 1.5 as well as the MMD-analysis [154] in Section III, MDL-VQA embraces five datasets with vastly different domains in visual inputs, accompanied with non-negligible domain shift in textual representations; 2) label-space variations: the label spaces (i.e., answer candiates) in different domains are inconsistent, where some general answers (e.g. *yes*, *one* and *red*) typically coexists in several or even all domains, meanwhile a certain number of answers are only involved in one specific domain; 3) data privacy: we highlight the data privacy issue in lifelong learning, where the training process use the data in only the current domain, without storing and replaying any instances from previous domains.

For the details in the MDL-VQA task, we re-organize five prevalent datasets: CLEVR, GQA, Vizwiz, AQUA as the sub-tasks for continual learning, covering the visual domains of composite, realistic, blurry-object, artistic and abstract scenes as depicted in Fig. 1.5. The statistical data for the utilized datasets can be seen in Tab. 1.2.

Table 1.2: The statistics of VQA datasets involved in the MDL-VQA benchmark. '*' denotes the modification of random sampling.

	Train	Test	Label	Frequent Answers
GQA*	46893	6473	1657	no, yes, left, right, man, white
CLEVR*	34926	5000	28	no, yes, 1, 0, small, rubber, metal
VQA-AB	29537	14738	426	yes, no , 2, 1, red, 3, white, blue
AQUA	14784	754	453	person, people, building, church
Vizwiz	10262	2160	3648	unanswerable, unsuitable, no, yes

In summary, we make an attempt to endow VQA models with a lifelong learning ability, and take a step towards the more practical and advanced future AI systems. We contribute a new lifelong VQA task, which is closer to real-world ReID applications and provide a research direction to explore more important but under-investigated VQA problems. The related details are elaborated in Chapter 7.

A novel yet practical VQA task under federated learning: As depicted in Fig. 1.6, we propose a challenging VQA task under the setting of personalized federated learning, named FedVQA. This task focuses on the decentralized learning under privacy constraints, where each client trains a local model to perform well on its private dataset, and meanwhile the central server optimizes a generic model to generalize well on unseen data. In addition, compared with the conventional FL on identically distributed (iid) data, the VQA samples collected from different local clients typically involves heterogeneous feature and label distributions, including

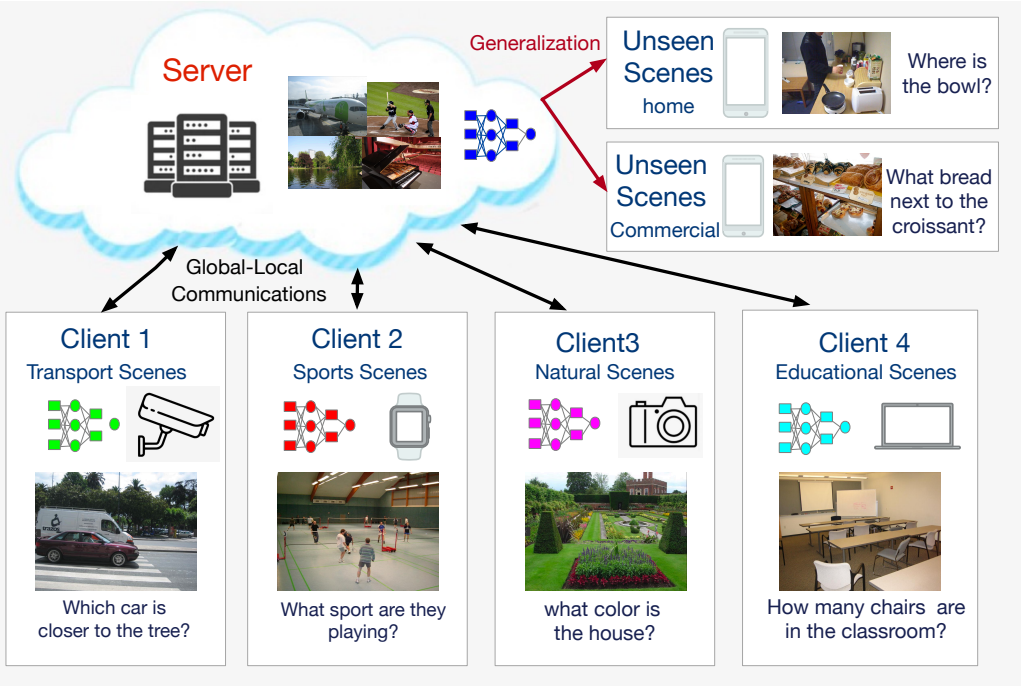

Figure 1.6: The personalized federated setting for VQA model over heterogeneous visual scenes. Given a pre-trained VQA model, we require each participated clients to train personalized models to perform well on their local data (e.g., transports, sports, natural and educational scenes). Meanwhile, the central server is expected to aggregate a generic global model to generalize on the testing data in unseen scenes (e.g., shopping and home).

diverse visual content captured from various realistic scenes (e.g. Fig. 1.6), as well as inconsistent answer distributions caused by different scene-specific questions. As a result, we construct a multi-scene FedVQA (MS-FedVQA) benchmark to evaluate models' performance on both seen personalized and unseen scenes.

In our proposed MS-FedVQA, we establish six clients' datasets for personalized training via selectively sampling from the large-scale GQA datasets [49], whose images are represented in diverse real-world scenes. Specifically, to introduce the scene-based distribution skews among different local datasets, we leverage the taxonomy of the Place 365 database [155], as well as the pretrained ResNet152-places365 model [156] to select the samples in commercial, educational, transport, natural sports and home scenes. The detailed sub-categories in each local datasets are in Tab. 1.3. It is noteworthy that, we only add the samples whose Top-1 predictive probability are higher than 0.7 into the corresponding local datasets. The label distributions over training samples in six local datasets are depicted in Fig. 1.7. We can see that the mainstream correct answers across different scenes are inconsistent in our FedVQA setting, which poses more challenges for federated solutions to mitigate label distribution skews in FedVQA task.

Table 1.3: The representative sub-categories of scenes in the six clients datasets specialized for answering questions to commercial, educational, transport, natural, sports, home scenes.

Dataset	Representative sub-categories of scenes
Commercial	restaurant, market, pharmacy, bakery, ticket booth, discotheque, beauty salon
Educational	campus, art gallery, music studio, church, museum, temple, lecture room
Transport	airport, subway , crosswalk, galley, bus, train station, airfield, boat deck, bridge
Natural	forest, mountain, marsh, underwater, fishpond, waterfall, ocean, lake, iceberg, desert
Sports	ballroom, arena, gymnasium, ski slope, basketball court, bowlling alley, locker room
Home	kitchen, bedroom, bathroom, closet, utility room, shower, living room, child's room

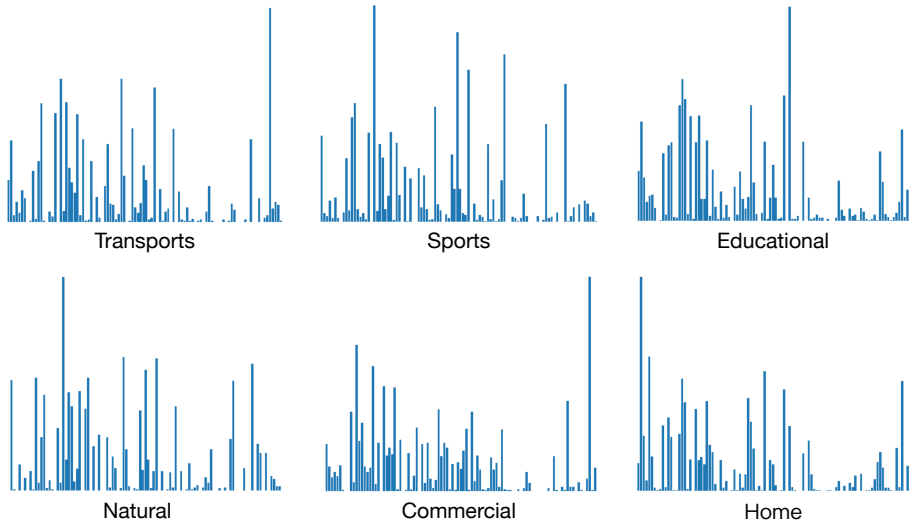

Figure 1.7: Label distributions of six scene-specific datasets over the first 100 answer candidates (overally high-frequency labels).

1.4.2 Technical Contributions

From Chapter 3 to Chapter 7, we develop new approaches based on deep learning to address the aforementioned research questions. The key contributions in these approaches are listed below.

A novel multimodal fusion scheme for VQA model architecture. Inspired by the multi-step reasoning behaviour in human cognitive processes, we present a novel Multi-stage Hybrid Embedding Fusion (MHEF) network to achieve multi-space and multi-stage interactions in a unified framework. The experimental results show that our proposed MHEF remarkably outperforms the dominant multimodal fusion approaches, and yields promising performance on the two widely-utilized VQA datasets.

A new loss function to overcome language bias problem. One potential cause to language bias problem in VQA models is the class imbalance in training data, where the samples grounded by frequent answer candidates make VQA models biased towards language modality. To tackle this issue, we propose a simple yet

effective Language Prior based Focal Loss, which decreases the gradients from more-biased samples via rescaling the standard cross entropy loss. Extensive experiments show that the our method can be generally applied to common baseline VQA models, and achieves significantly better performance on out-of-distribution dataset.

A new curriculum learning based de-biasing framework to improve out-of-distribution performance. Motivated by the fact that VQA samples with different levels of language bias contribute differently to answer prediction, we overcome the language prior problem by proposing a novel Language Bias driven Curriculum Learning (LBCL) approach, which employs an easy-to-hard learning strategy with a novel difficulty metric Visual Sensitive Coefficient (VSC). Besides, to avoid the catastrophic forgetting of the learned concept during the multi-stage learning procedure, we propose to integrate knowledge distillation into the curriculum learning framework. Extensive ablation studies demonstrate the effectiveness of each component in our method, and our method significantly improves model performance on the out-of-distribution dataset.

An effective collaborative causal regularization for unbiased Audio-Visual Question Answering models. To alleviate the multiple shortcut bias problem discovered in AVQA models, we propose a COllaborative CAusal (COCA) regularization to cooperatively overcome the complex biases and improve model robustness from two aspects: 1) a new Bias-centered Causal Regularization (BCR) to mitigate the bias caused by a specific shortcut correlation through regularization from counterfactual and factual views. 2) 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. We validated the effectiveness of COCA through extensive comparative and ablative studies.

A novel replay-free regularization based method to overcome catastrophic forgetting in lifelong VQA. Specialized for the challenge of severe domain shift as well as label-space variation in the MDL-VQA benchmark, we propose an effective Self-Critical Distillation (SCD) approach to avoid forgetting via both feature- and logits-level knowledge distillation. Such dual-level distillations in the SCD framework benefit each other mutually, which enhances the VQA model's stability to anti-forgetting while keeping its plasticity to learn newly-coming knowledge. Extensive experiments demonstrate that our SCD framework outperforms state-of-the-art competitors on the proposed MDL-VQA benchmark with different training orders.

An efficient federated pairwise preference preserving strategy to consolidate global generic knowledge in personalized federated VQA. Concentrating on the forgetting issue of global knowledge during personalized learning for local models, we introduce a flexible yet effective differentiable pairwise preference (DPP) method that formulates the global knowledge as the pairwise binary comparisons

among significance of answer prediction. Furthermore, we introduce a forgotten-knowledge filter (FKF) that seeks to generate a forgotten-knowledge driven label distribution to capture the easily-forgotten classes during local training, and then adaptively filters a significant answer subset involved in pairwise preference. Benefiting from FKF in DPP, our method not only enhances the performance of both local and global models, but also remarkably reduces the computational complexity in terms of knowledge preserving.