Open-Ended Visual Question Answering by Multi-Modal Domain Adaptation

Yiming Xu[1]*, **Lin Chen**[2], **Zhongwei Cheng**[3], **Lixin Duan**[4], **and Jiebo Luo**[5]
[1] Northwestern University
[2] Wyze Labs Inc
[3] Amazon
[4] University of Electronic Science and Technology of China
[5] University of Rochester
x.yiming@outlook.com, {gggchenlin,emory.cheng, lxduan, jiebo.luo}@gmail.com

Abstract

We study the problem of visual question answering (VQA) in images by exploiting supervised domain adaptation, where there is a large amount of labeled data in the source domain but only limited labeled data in the target domain, with the goal to train a good target model. A straightforward solution is to fine-tune a pre-trained source model by using those limited labeled target data, but it usually cannot work well due to the considerable difference between the data distributions of the source and target domains. Moreover, the availability of multiple modalities (*i.e.*, images, questions and answers) in VQA poses further challenges in modeling the transferability between various modalities. In this paper, we address the above issues by proposing a novel supervised multi-modal domain adaptation method for VQA to learn joint feature embeddings across different domains and modalities. Specifically, we align the data distributions of the source and target domains by considering those modalities both jointly and separately. Extensive experiments on VQA 2.0 and VizWiz datasets demonstrate that our proposed method outperforms the existing state-of-the-art baselines for open-ended VQA in this challenging domain adaptation setting.

1 Introduction

The task of visual question answering (VQA) is to build a model for answering questions given an image-question pair. Recently, it has received great attention from computer vision community (Zhou et al., 2015; Kazemi and Elqursh, 2017; Tan and Bansal, 2019; Anderson et al., 2017; Kim et al., 2018; Zhang et al., 2018; Singh et al., 2019). VQA requires techniques from both image recognition and natural language processing, and most existing works use Convolutional Neural Networks (CNNs) to extract visual features from images and Recurrent Neural Networks (RNNs) to generate textual features from questions, and then combine them to generate the final answers.

However, most existing VQA datasets are created in a way that is not suitable as training data for real-world applications. For example, VQA 2.0 (Goyal et al., 2019) and Visual7W (Zhu et al., 2016), arguably two of the most popular datasets for VQA, were created by using images from MSCOCO (Lin et al., 2014) with questions asked by crowd workers. Therefore, the images are typically of high quality and the questions are less conversational. On the contrary, the recently proposed VizWiz dataset (Gurari et al., 2018) was collected from blind people taking photos and asking questions about those photos. Therefore, the images in VizWiz are often of poor quality, and questions are more conversational with some questions might even be unanswerable due to the poor quality of the images. While VizWiz dataset reflects a more realistic setting for VQA, its size is much smaller due to the difficulty of collecting such data. A straightforward solution to this problem is to first train a model on the VQA 2.0 dataset and then fine-tune it using the VizWiz data. However, this solution can only provide limited improvement with two major issues. First, the VQA datasets are constructed in a different way, making them differ significantly in visual content, textual questions and answers. (Sha et al., 2018) conducted an experiment to classify different VQA datasets with a simple multi-layer perceptron (MLP) of one hidden layer and it achieved over 98% accuracy. This is a strong indication of the significant bias across different datasets. Our experiments also validate that directly fine-tuning the model trained on VQA 2.0 results in minor improvement on VizWiz. Sec-

*This work was done during Yiming Xu's internship at Futurewei Technologies.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 367–376

ond, the two modalities (visual and textual) also pose a big challenge in the generalizability across datasets. It is a nontrivial task to consistently bridge the domain gap in a coordinated fashion, when multiple modalities are involved, due to the nature of the multi-modal heterogeneity with no common feature representations.

Domain adaptation methods, which handle the difference between two domains, have been developed to address the first issue (Hoffman et al., 2015; Koniusz et al., 2017; Tzeng et al., 2017; Ganin and Lempitsky, 2015; Shen et al., 2017; Gong et al., 2012; Guo and Xiao, 2012; Yao et al., 2015). However, most existing domain adaptation methods focus on single-modal tasks such as image classification or sentiment classification, and thus may not be directly applicable to multi-modal settings. On the other hand, these methods are usually subject to a strong assumption on the label distribution that the source domain and the target domain share the same (usually small) label space, which is usually unrealistic. (Qi et al., 2018) proposed a new framework for unsupervised multi-modal domain adaptation, but it was not designed for the VQA tasks. Recently, several VQA domain adaptation methods have been proposed to address the multi-modal challenge. However, to the best of our knowledge, all the existing VQA domain adaptation methods focus on the multiple choice setting, where several answer candidates are provided and the model only needs to select one from them. In contrast, we focus on a more challenging open-ended setting where there is no prior knowledge of answer choices.

In this paper, we address the aforementioned challenges by proposing a novel multi-modal domain adaptation framework, which learns a multi-modal feature embedding that simultaneously keeps each domain invariant and each individual modality discriminative, based on an adversarial loss and a classification loss. We additionally incorporate the maximum mean distance (MMD) to further reduce the domain mismatch by learning embeddings from different modalities.

Our contributions are summarized as follows:

1) We propose a novel supervised multi-modal domain adaptation framework to tackle the more challenging open-ended VQA task. To the best of our knowledge, this is the first attempt of using domain adaptation for open-ended VQA.

2) We propose a method that learns a multi-modal feature embedding that simultaneously keeps each domain invariant and each individual modality discriminative, with an adversarial loss and a classification loss. At the same time, it minimizes the difference of cross-domain feature embeddings jointly over multiple modalities.

3) We conduct extensive experiments on two popular benchmark datasets (*i.e.*, VQA 2.0 and VizWiz), and the results clearly show the effectiveness of our proposed method over the existing state-of-the-art baselines.

2 Related Work

VQA datasets: Over the past few years, several VQA datasets (Zhu et al., 2016; Goyal et al., 2019; Gurari et al., 2018; Krishna et al., 2017; Antol et al., 2015) and tasks were proposed to encourage researchers to develop algorithms that answer visual questions. One limitation of many existing datasets is that they were created either automatically or from an existing vision dataset like MSCOCO (Lin et al., 2014) with the questions either generated automatically or contrived by human annotators. This makes the images in these datasets typically of high quality and the questions less conversational, and thus might not be directly applicable to real-world applications such as (Gurari et al., 2018) which aims to answer the visual questions asked by blind people in their daily life. The main differences between (Gurari et al., 2018) and other VQA datasets are as follows: 1) Both the image and question quality of (Gurari et al., 2018) are lower as they suffer from poor lighting, out of focus and audio recording problems like clipping a question at either end or catching background audio content; 2) The questions can be unanswerable since blind people can hardly verify whether the images contain the visual content they are asking about, due to blurring, inadequate lighting, framing errors, finger covering the lens, *etc.* Our experiments also reveal that fine-tuning the model trained on the VQA 2.0 dataset provides limited improvement on VizWiz, due to the significant difference in bias between both datasets.

VQA settings: There are two main VQA settings, namely multiple choice and open-ended following (Antol et al., 2015)[1]. Under the multiple choice setting, the model is provided with multiple candidates of answers and is expected to se-

[1]Please note that these two terms are inherited from the original paper proposed a VQA dataset by (Antol et al., 2015) and are commonly used in VQA challenges.

lect the correct one. VQA models following this setting take characteristics of all answer candidates like word embeddings as the input to make a selection (Sha et al., 2018; Jabri et al., 2016). However, in the open-ended setting, there is neither prior knowledge nor answer candidates provided, and the model can respond with any free-form answers. This makes the open-ended setting more challenging and realistic (Kim et al., 2018; Kazemi and Elqursh, 2017; Singh et al., 2019; Anderson et al., 2017).

VQA models: Recently, a plethora of VQA models were proposed (Zhou et al., 2015; Kazemi and Elqursh, 2017; Anderson et al., 2017; Kim et al., 2018; Singh et al., 2019). Most of them consist of image and question encoders, and a multi-modal fusion module followed by a classification module. (Kazemi and Elqursh, 2017) used an LSTM to encode the question and a residual network (He et al., 2015) to compute the image features with a soft attention mechanism. (Anderson et al., 2017) implemented a bottom-up attention using Faster R-CNN (Ren et al., 2015) to extract features of detected image regions, and then a top-down mechanism used task-specific context to predict an attention distribution over the image regions. The final output was generated by an MLP after fusing the image and question features. (Kim et al., 2018) used a bilinear attention between two groups of input channels on top of low-rank bilinear pooling which extracted the joint representations for each pair of channels. (Singh et al., 2019) proposed an approach that takes original image features, bottom-up attention features from object detection module, question features and the optical character recognition (OCR) strings detected from the image as the input, and answers either with an answer from the fixed answer vocabulary or by selecting one of the OCR strings detected in the image. Similar to the state-of-the-art model by (Singh et al., 2019), our VQA base model also takes original image features, bottom-up attention features and question features to predict the final answer. Details of our VQA base model is described in the next section.

Domain adaptation: Domain adaptation techniques have been proposed to learn a common domain invariant latent feature space where the distributions of two domains are aligned. Recent works typically focused on transferring knowledge from a labeled source domain to a target domain where there is no or limited labeled data (Hoffman et al., 2015; Koniusz et al., 2017; Tzeng et al., 2017; Shen et al., 2017; Ganin and Lempitsky, 2015; Gong et al., 2012; Guo and Xiao, 2012). (Hoffman et al., 2015) optimized for domain invariance to facilitate domain transfer and used a soft label distribution matching loss to transfer information between tasks. (Tzeng et al., 2017) proposed a framework which combines discriminative modeling, untied weight sharing and a GAN loss to reduce the difference between domains. (Shen et al., 2017) estimated empirical Wasserstein distance between the source and the target samples and optimized the feature extractor network to minimize the estimated Wasserstein distance in an adversarial manner. (Ganin and Lempitsky, 2015) utilized gradient reversal layer (GRL) to incorporate the training process of domain classifier, label classifier and feature extractor to align domains. Similarly, (Guo and Xiao, 2012) simultaneously minimized the classification error, preserved the structure within and across domains, and restricted similarity on target samples. The major difference between our work and these works is that we propose a novel *multi-modal* domain adaptation framework, while these works assumed a single modality.

Domain adaptation for VQA: Although domain adaptation has been successfully applied to computer vision, its applicability to VQA has yet to be well-studied. There was one recent work investigating domain adaptation for VQA by (Sha et al., 2018). It reduces the difference in distributions by transforming the feature representation of the data in the target domain. However, one major limitation is the assumption of a multiple choice setting, where four answer candidates are provided as the input to the model. It is unrealistic because one can never guarantee that the ground truth answer is among four candidates. Moreover, it is unclear how to create answer candidates for an image-question pair. On the contrary, our model is only provided with an image-question pair and can generate any free-form answers. This makes our task more challenging and realistic.

3 The VQA Framework

In this section, we describe our base VQA framework. Given an image I and a question Q, the VQA model estimates the most likely answer $\hat{a}$ from a large vocabulary based on the content of

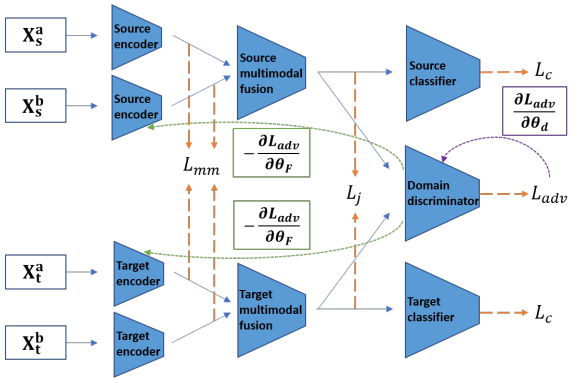

Figure 1: The proposed multi-modal domain adaptation framework. $\mathbf{X}_s^a, \mathbf{X}_s^b, \mathbf{X}_t^a, \mathbf{X}_t^b$ denote original features for two modalities. The blue arrow denotes forward propagation while the orange arrow denotes the loss calculation. The purple and green arrows denote backward propagation for discriminator loss L_{adv}.

the image, which can be written as follows:

$$\hat{a} = \text{argmax}_a \, P(a|I, Q). \quad (1)$$

Our base framework consists of four components: 1) a question encoder; 2) an image encoder; 3) a multi-modal fusion module; and 4) a classification module. We will elaborate about each component in the following subsections.

Question encoding: The question Q of length T is first tokenized and encoded using word embedding based on pre-trained GloVe (Pennington et al., 2014) as $S = \{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_T\}$. These embeddings are then fed into a GRU cell (Cho et al., 2014). The encoded question is obtained from the last hidden state at time step T denoted as $\mathbf{q} = f^q(Q; \boldsymbol{\theta}_q) \in \mathcal{R}^{d_q}$, where $f^q(Q; \boldsymbol{\theta}_q) = \mathbf{h}_T$, $\mathbf{h}_t = \text{GRU}(\mathbf{x}_t, \mathbf{h}_{t-1}; \boldsymbol{\theta}_q)$ for $1 \leq t \leq T$, and d_q is the feature dimension.

Image encoding: Similar to (Anderson et al., 2017) and (Singh et al., 2019), we first feed the input image I to an object detector (Girshick et al., 2018) pre-trained on the Visual Genome dataset (Krishna et al., 2017) based on Feature Pyramid Networks (FPN) (Lin et al., 2016) with ResNeXt (Xie et al., 2017) as the backbone. The output from $fc6$ layer is used as the region-based features, *i.e.*, $\mathbf{V}_r = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K\}$ with $\mathbf{v}_i$ as the feature for i-th object. Meanwhile, we divide the entire image into a 7×7 grid, and obtain the grid-based features $\mathbf{V}_g$ by average pooling features from the penultimate layer $5c$ of a pre-trained ResNet-101 network (He et al., 2015) on ImageNet dataset. Finally, we combine $\mathbf{V}_r$ and $\mathbf{V}_g$ as well as the question embedding $\mathbf{q}$ to obtain the joint feature embedding in a multi-modal fusion module (see next paragraph for more details).

Multi-modal fusion and classification: The question embedding $\mathbf{q}$ is used to obtain the attention weights on region-based image features $\mathbf{V}_r$. Then, the region-based features $\mathbf{V}_r$ are averaged based on the attention weights to obtain the weighted region-based image features. Similarly, grid-based features $\mathbf{V}_g$ are fused with question embedding $\mathbf{q}$ by concatenation. The fused grid-based features and the weighted region-based image features are concatenated to obtain the final image features $\mathbf{v}$. We have also tried other combination schemes such as (Ben-younes et al., 2017; Yu et al., 2018, 2017), but they fail to outperform concatenation and are much slower. Since our focus is on domain adaptation instead of the base VQA model, we use concatenation in our work.

We denote the final image feature embedding as $\mathbf{v} = f^v(\mathbf{q}, I; \boldsymbol{\theta}_v)$. The final joint embedding $\mathbf{e} = f^j(\mathbf{q}, \mathbf{v})$ is calculated by taking the Hadamard product of $\mathbf{q}$ and $\mathbf{v}$, and then is fed to an MLP $f^c(\mathbf{e}; \boldsymbol{\theta}_c)$ for classification, *i.e.*, $a = f^c(\mathbf{e}; \boldsymbol{\theta}_c)$. The final answer is determined by $\hat{a} = \text{argmax}_a \, f^c(\mathbf{e}; \boldsymbol{\theta}_c)$.

4 Multi-Modal Domain Adaptation

In this section, we present our framework for supervised multi-modal domain adaptation. We assume there are two modalities[2] of source samples $\mathbf{X}_s = [\mathbf{X}_s^a, \mathbf{X}_s^b]$, which would be vision and language in the context of VQA, where a, b denote the two modalities, and labels $\mathbf{Y}_s$ drawn from a source domain joint distribution $P_s(x, y)$, as well as the two modalities of target samples $\mathbf{X}_t = [\mathbf{X}_t^a, \mathbf{X}_t^b]$ and labels $\mathbf{Y}_t$ drawn from a target joint distribution $P_t(x, y)$. We also assume there are sufficient source data so that a good pre-trained source model can be built, but the amount of target data is limited so that learning on only the target data leads to poor performance. Our goal is to learn the target representations for two modalities f_t^a, f_t^b, the multi-modal fusion f_t^j and the target classifier f_t^c with the help of the pre-trained source representations f_s^a, f_s^b, f_s^j and the source classifier f_s^c. For the VQA task in our work, a, b denote visual and textual modalities, respectively.

A typical approach to achieve this goal

[2]For simplicity, we assume the data has two modalities, but it can be easily generalized to more modalities.

is to regularize the learning of the source and target joint representations by minimizing the distance of empirical distributions between the source and target domains, *i.e.*, between $f_s^j\left(f_s^a(\mathbf{X}_s^a;\boldsymbol{\theta}_s^a), f_s^b(\mathbf{X}_s^b;\boldsymbol{\theta}_s^b);\boldsymbol{\theta}_s^j\right)$ and $f_t^j\left(f_t^a(\mathbf{X}_t^a;\boldsymbol{\theta}_t^a), f_t^b(\mathbf{X}_t^b;\boldsymbol{\theta}_t^b);\boldsymbol{\theta}_t^j\right)$. In this way, the data from the source domain and the target domain are projected onto a similar latent space, such that well-performing source model can lead to well-performing target model. Following this idea, we propose a novel multi-modal domain adaptation framework as shown in Figure 3.

4.1 Joint Embedding Alignment

We propose to reduce the difference of joint embeddings between the source and the target domains by minimizing the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012). The intuition is that two distributions are identical if and only if all of their moments coincide.

Empirically, we can minimize the following object function

$$\mathrm{MMD}(\mathbf{X}_s,\mathbf{X}_t)=\left\|\frac{1}{n_s}\sum_{i=1}^{n_s}\varphi(\mathbf{x}_i^s)-\frac{1}{n_t}\sum_{i=1}^{n_t}\varphi(\mathbf{x}_i^t)\right\|_{\mathcal{H}}. \quad (2)$$

We then define the loss function as

$$L_j = E_{\mathbf{X}_s\sim p_s,\mathbf{X}_t\sim p_t}\left[\mathrm{MMD}^2(\mathbf{e}_s,\mathbf{e}_t)\right], \quad (3)$$

where $\mathbf{e}_s = f_s^j\left(f_s^a\left(\mathbf{X}_s^a;\boldsymbol{\theta}_s^a\right), f_s^b(\mathbf{X}_s^b;\boldsymbol{\theta}_s^b);\boldsymbol{\theta}_s^j\right)$ and $\mathbf{e}_t = f_t^j\left(f_t^a(\mathbf{X}_t^a;\boldsymbol{\theta}_t^a), f_t^b(\mathbf{X}_t^b;\boldsymbol{\theta}_t^b);\boldsymbol{\theta}_t^j\right)$. By minimizing the difference between source and target joint embeddings, we enforce that the joint embeddings of both source domain and target domain will be projected onto a similar latent space.

4.2 Multi-Modal Embedding Alignment

It is more challenging to reduce multi-modal domain shift than conventional single-modal domain shift. The previous loss L_j in Eq. (3) does not explicitly consider the multi-modal property. Aligning only the joint feature embedding is insufficient to adapt the source domain to the target domain. This is because the feature extractor for each modality has its own complexity of domain shift, which often differs from each other (*e.g.*, visual vs. textual). Aligning only the fused features cannot fully reduce domain differences.

Therefore, we introduce the following term to minimize the maximum mean discrepancy between every single modality, *i.e.*, $\mathrm{MMD}\left(f_s^a(\mathbf{X}_s^a;\boldsymbol{\theta}_s^a), f_t^a(\mathbf{X}_t^a;\boldsymbol{\theta}_t^a)\right)$ and $\mathrm{MMD}\left(f_s^b(\mathbf{X}_s^b;\boldsymbol{\theta}_s^b), f_t^b(\mathbf{X}_t^b;\boldsymbol{\theta}_t^b)\right)$. Then, the loss function to minimize can be written as

$$L_{mm} = \underset{\mathbf{X}_s\sim p_s,\mathbf{X}_t\sim p_t}{E}\Big[\gamma_a\,\mathrm{MMD}^2\left(f_s^a(\mathbf{X}_s^a;\boldsymbol{\theta}_s^a), f_t^a(\mathbf{X}_t^a;\boldsymbol{\theta}_t^a)\right) + \gamma_b\,\mathrm{MMD}^2\left(f_s^b(\mathbf{X}_s^b;\boldsymbol{\theta}_s^b), f_t^b(\mathbf{X}_t^b;\boldsymbol{\theta}_t^b)\right)\Big], \quad (4)$$

where γ_a and γ_b are trade-off parameters.

Figure 2: Sample image-question pairs and valid answers for VQA 2.0 and VizWiz datasets. For each image-question pair, there are 10 answers provided by 10 different crowd workers.

4.3 Classification

While minimizing the distance between source and target embeddings, we also want to maintain the classification performance on both the source domain and the target domain. Similarly as in a standard supervised learning setting, we employ the cross entropy loss for classification:

$$L_c = E_{(\mathbf{X}_t,\mathbf{Y}_t)\sim p_t}\left[\mathrm{CE}(f_t^c(\mathbf{e}_t;\boldsymbol{\theta}_t^c),\mathbf{Y}_t)\right] + \gamma_c E_{(\mathbf{X}_s,\mathbf{Y}_s)\sim p_s}\left[\mathrm{CE}(f_s^c(\mathbf{e}_s;\boldsymbol{\theta}_s^c),\mathbf{Y}_s)\right], \quad (5)$$

where CE denotes the cross entropy loss with γ_c as the trade-off parameter between the two domains.

4.4 Domain Discriminator

We also propose to use a domain classifier f^d to reduce the mismatch between the source domain and target domain by confusing the domain classifier from correctly distinguishing a sample from source domain or target domain. The domain classifier f^d has a similar structure to f_t^c or f_s^c except the last layer outputs a scalar in $[0,1]$ with the value indicating how likely the sample comes from the source domain. Thus, f^d can be optimized according to a standard cross-entropy loss. To make the features domain-invariant, the source and target mappings are optimized according to a constrained adversarial objective. The domain classifier minimizes this objective while the encoding

model maximizes this objective. The generic formulation for domain adversarial technique is:

$$L_{adv} = - E_{\mathbf{X}_s \sim p_s} \left[\log f^d(\mathbf{e}_s; \boldsymbol{\theta}_d) \right] - E_{\mathbf{X}_t \sim p_t} \left[\log(1 - f^d(\mathbf{e}_t; \boldsymbol{\theta}_d)) \right]. \quad (6)$$

For simplicity, we denote $\boldsymbol{\theta}^F = \left(\boldsymbol{\theta}_s^a, \boldsymbol{\theta}_t^a, \boldsymbol{\theta}_s^b, \boldsymbol{\theta}_t^b, \boldsymbol{\theta}_s^j, \boldsymbol{\theta}_t^j \right)$ as the parameters of all feature mappings and $\boldsymbol{\theta}^C = (\boldsymbol{\theta}_s^c, \boldsymbol{\theta}_t^c)$ as the parameters of all label predictors. The final objective function to minimize then becomes:

$$L_{\boldsymbol{\theta}^F, \boldsymbol{\theta}^C, \boldsymbol{\theta}^d} = L_c + \lambda_j L_j + \lambda_{mm} L_{mm} - \lambda_{adv} L_{adv}. \quad (7)$$

We seek a saddle point $\hat{\boldsymbol{\theta}}^F, \hat{\boldsymbol{\theta}}^C, \hat{\boldsymbol{\theta}}^d$ of L which satisfies the following conditions:

$$\begin{aligned} \left(\hat{\boldsymbol{\theta}}^F, \hat{\boldsymbol{\theta}}^C \right) &= \operatorname{argmin}_{\boldsymbol{\theta}^F, \boldsymbol{\theta}^C} L(\boldsymbol{\theta}^F, \boldsymbol{\theta}^C, \hat{\boldsymbol{\theta}}^d) \\ \hat{\boldsymbol{\theta}}^d &= \operatorname{argmax}_{\boldsymbol{\theta}^d} L(\hat{\boldsymbol{\theta}}^F, \hat{\boldsymbol{\theta}}^C, \boldsymbol{\theta}^d). \end{aligned} \quad (8)$$

At the saddle point, the parameters $\boldsymbol{\theta}^d$ of the domain classifier minimize the domain classification loss $\mathcal{L}_{adv}$ (since we maximize $-\mathcal{L}_{adv}$) while the parameters $\boldsymbol{\theta}^C$ of the label predictor minimize the label prediction loss $\mathcal{L}_c$. The feature mapping parameters $\boldsymbol{\theta}^F$ minimize the label prediction loss such that the features are discriminative, while maximizing the domain classification loss such that the features are domain-invariant. In addition to MMD which explicitly aligns the distributions, domain discriminator implicitly aligns the distributions, leading to stronger regularization in the non-convex optimization problem.

5 Experiments

In this section, we validate our proposed method for the challenging open-ended VQA task, by comparing with a few state-of-the-art baselines.

5.1 Datasets

Two popular VQA benchmarks are used in our experiments, VQA 2.0 (Goyal et al., 2019) and VizWiz (Gurari et al., 2018). A comparison of the statistics for both datasets are listed in Table 1, which shows that the scale of VizWiz is much smaller in terms of the numbers of images and questions. Although VizWiz has more unique answers, only 824 out of its top 3,000 answers overlap with the top 3,000 answers in VQA 2.0. This explains why models trained on VQA 2.0 perform poorly on VizWiz, and their limited transferability. There are 28.63% of questions in VizWiz are even not answerable due to reasons mentioned before, making the domain gap even more significant. Figure 2 shows some examples from both VQA 2.0 and VizWiz datasets. The difficulty of the task can also be seen from the VizWiz samples: images are blurry, viewpoints are unusual, some questions are unanswerable, and ground truth answers are highly inconsistent (*e.g.*, "soda", "coca cola 0", "coke 0").

Table 1: The statistics of VQA 2.0 and VizWiz dataset. Numbers are in train/validation/test order, and "# unique" denotes the number of unique answers.

	VQA 2.0	**VizWiz**
# images	83K / 41K / 81K	20K / 3K / 8K
# questions	443K / 214K / 448K	20K / 3K / 8K
# answers	4.4M / 2.1M / NA	0.2M / 0.03M / NA
# unique	3,126	58,789

5.2 Evaluation Metrics

In VQA, each question is usually associated with 10 valid answers from 10 distinct annotators. We follow the conventional evaluation metric on the open-ended VQA setting to compute the accuracy using the following formula:

$$\text{Acc}(\text{ans}) = \min \left(\frac{\text{\# humans said ans}}{3}, 1 \right). \quad (9)$$

Namely, an answer is considered correct if at least three annotators agree on the answer. Note that the true answers in VizWiz test set are not publicly available. In order to obtain the performance on the test set, results need to be uploaded to the official online submission system at https://evalai.cloudcv.org/web/challenges/challenge-page/102.

5.3 Implementation Details

In all experiments, we extract $K = 100$ objects for each image to construct the region-based features $\mathbf{V}_r$ and set the visual feature dimension to 2048. We also set the hidden dimension of GRU to 1024 and hidden dimension after fusion to 4096. The question length is truncated at 24. During training, we apply a warm-up strategy by gradually increasing the learning rate η from 0.001 to 0.01 in the first 2000 iterations. It is then multiplied by 0.15 after every 4000 iterations. We use a batch size of 128.

For domain adaptation, we let the source and target networks share the same parameters up to the penultimate layer, *i.e.*, $\boldsymbol{\theta}^v = \boldsymbol{\theta}_s^v = \boldsymbol{\theta}_t^v$ and

$\boldsymbol{\theta}^q = \boldsymbol{\theta}^q_s = \boldsymbol{\theta}^q_t$. In multi- or single-modal alignment, we use Gaussian kernel $k(x, y) = \exp^{-\frac{||x-y||^2}{2\sigma^2}}$ to compute MMD, because the Gaussian kernel can approximate functions under mild assumptions (continuous, bounded) fairly well, while other kernels such as the polynomial kernel do not have such properties. The trade-off parameters are set as $\lambda_j = 0.025$, $\lambda_{mm} = 0.008$, $\gamma_v = 0.8$, $\gamma_q = 1$, $\gamma_c = 0.001$, and $\lambda_{adv} = 0.003$.

5.4 Experimental Setup

First, we conduct experiments using VQA 2.0 as the source domain and VizWiz as the target domain, to evaluate the effectiveness of our proposed method for multi-modal domain adaptation. We also conduct experiments in the opposite way, *i.e.*, using VizWiz as the source domain and VQA 2.0 as the target domain, to further demonstrate the effectiveness of our approach.

We need to emphasize that we choose not to use an overly strong base model (*i.e.*, question embedding from FastText, complex fusion techniques, OCR tokens *etc.*), as our focus is on multi-modal adaptation instead of the base model itself. Despite that, we will show that our proposed domain adaptation method with a weaker base model still outperforms the fine-tuned state-of-the-art model.

5.5 Results and Analysis

Adaptation from VQA 2.0 to VizWiz: As discussed in previous sections, we first pre-train a source model on the VQA 2.0 dataset, and then adapt it to the target dataset VizWiz. The results of our proposed method and other leading methods are shown in Table 2.

We first compare our method with the original VizWiz baseline proposed by (Gurari et al., 2018), the previous state-of-the-art VQA model BAN by (Kim et al., 2018) and the current state-of-the-art VQA model Pythia by (Singh et al., 2019). It is clear that our method outperforms the state-of-the-art models by a significant margin from Table 2.[3]

In order to validate that the better performance of our method is not from a strong base model, we additionally report the results of our method in Table 3, with 1) training our single base model from scratch using only the VizWiz dataset (**Target only**), 2) fine-tuning from the model pre-trained on the VQA 2.0 dataset (**Fine-tune**), and 3) our proposed domain adaptation method (**DA**). From

[3]The results are averaged over five runs with a standard deviation of 0.11 for our model.

Table 2: Accuracy (in %) comparisons on VizWiz.

Method	Accuracy
VizWiz baseline (Gurari et al., 2018)	47.50
BAN (Kim et al., 2018)	51.40
Pythia[4] (Singh et al., 2019)	54.72
Ours	**55.87**

Table 3: Accuracy (in %) comparison for our base model. **Target only** denotes training from scratch, **Fine-tune** means fine-tuning and **DA** presents our domain adaptation method.

Target only	Fine-tune	DA
53.11	53.97	55.87

Table 3, it shows that our model fine-tuned from VQA 2.0 is about 0.75 percent worse than Pythia fine-tuned from VQA 2.0 (53.97% vs. 54.72%), indicating that the better performance of our final model than the state-of-the-art is not from a strong base model. Moreover, the accuracy of our base model trained from scratch is 53.11%, falling behind 0.6 percent to Pythia trained from scratch, which is consistent with our observation that our method even with a weaker base model can achieve better final results.

Results breakdown into answer categories: Table 4 shows the accuracy breakdown into different answer categories. The results show that our model achieves new state-of-the-art performance on "Number" and "Other" categories as well as overall accuracy. Note that the overall accuracy for Pythia in this table is 54.22% instead of 54.72% which we were unable to reproduce using the released code and there are no breakdown numbers reported associated with it. The best we can achieve with Pythia (after fine-tuning from VQA 2.0) is 54.22% and the corresponding breakdown numbers are reported in the table.

Ablation study: We conduct an ablation study to show the contributions of different components of our method. The results show that the multi-modal MMD brings the most significant performance gain, which validates that aligning on every single modality is beneficial to the transferability of multi-modal tasks. Comparing two single modalities, MMD alignment on textual features is more helpful for model performance than MMD alignment on visual features, which we postulate is because the VizWiz dataset contains a large number of blurry images and thus those images are unhelpful for adaptation. In addition, MMD on joint embedding and discriminator is also crucial to bring further performance gain of 0.41%. Not surprisingly, an ensemble of three models pushes

Table 4: Results breakdown into different categories of different methods for domain adaptation from VQA 2.0 to VizWiz. Breakdown numbers are performance on VizWiz *test-dev* split.

(Accuracy in %)	Overall	Yes/No	Number	Unanswerable	Other
VizWiz baseline (Gurari et al., 2018)	47.50	66.90	22.00	77.00	29.40
BAN (Kim et al., 2018)	51.40	68.10	17.90	**85.30**	31.50
Pythia (Singh et al., 2019)	54.22	**74.83**	31.11	84.08	35.03
Ours	**55.87**	74.33	**32.00**	83.32	**38.53**

Table 5: Ablation study of our proposed method.

Method	Accuracy	Improved
Target only	53.11	-
(+ Fine-tune)	53.97	+ 0.86
+ MMD on V	54.61	+ 0.64
+ MMD on Q	55.46	+ 0.85
+ MMD on joint	55.69	+ 0.23
+ GRL	55.87	+ 0.18
+ Ensemble of 3 models	56.20	+ 0.33

our performance even higher to 56.20%, which is the state-of-the-art performance to date.

Comparisons on other domain adaptation methods: We compare our multi-modal domain adaptation method with some popular domain adaptation methods, including DANN (Ganin and Lempitsky, 2015), ADDA (Tzeng et al., 2017), WDGRL (Shen et al., 2017), and SDT (Hoffman et al., 2015). Note that DANN, ADDA and WDGRL were originally designed for unsupervised domain adaptation. For fair comparison, we fine-tune the model using target labels after unsupervised adaptation (hence they are indicated by a suffix '+'), and we also compare with a popular and effective supervised domain adaptation method SDT. The results shown in Table 6 illustrate that compared to direct fine-tuning, the existing domain adaptation methods do not help much (DANN performs even worse) in the multi-modal task, while our method outperforms both direct fine-tuning and existing domain adaptation methods by a notable margin.

Table 6: Accuracy (in %) comparisons of our method with state-of-the-art domain adaptation methods.

VizWiz	Accuracy
Fine-tune	53.97
DANN+ (Ganin and Lempitsky, 2015)	53.65
ADDA+ (Tzeng et al., 2017)	54.06
WDGRL+ (Shen et al., 2017)	54.28
SDT (Hoffman et al., 2015)	54.56
Ours	**55.87**

Adaptation with fewer target training samples: We also validate the robustness of our framework by reducing the target training dataset size. We experiment with various target sizes of 12.5% (2,500), 25% (5,000), 50% (10,000) and all data (20,000). The results are shown in Table 7. We can observe that with the increase of the amount of training data, the performance gain over fine-tuning is decreasing. We conjecture that this is because when we have limited amount of target data, having more prior knowledge is beneficial to model performance, while having more target data will make prior knowledge less helpful. However, our method can stably improve the performance because it sufficiently makes use of target data and source data. It is more promising that our domain adaptation method using fewer samples can achieve comparable or even better performance compared with training from scratch using doubled amount of data (especially when target data is scarce), *e.g.*, our method using 25% data (48.93%) outperforms training from scratch using 50% data (47.48%).

Table 7: Accuracy (in %) comparison using less data.

% target data	Target only	Fine-tune	DA
12.5%	39.51	43.39	45.02
25%	43.75	47.71	48.93
50%	47.48	50.12	52.32
All data	53.11	53.97	55.87

Table 8: Accuracy (in %) comparison for our single base model adapted from VizWiz to VQA 2.0.

Target only	Fine-tune	DA
68.89	69.25	70.06

Adaptation from VizWiz to VQA 2.0: In order to further validate the robustness of our method, we reverse the source domain and the target domain and perform adaptation. We pre-train the source model on VizWiz and adapt the source model to VQA 2.0. The results are shown in Table 8, from which we still can observe a significant improvement for our method against fine-tuning. In comparison, the performance of MFH (Yu et al., 2018), BAN and Pythia is 67.7%, 69.08% and 69.21%, respectively, all under-performing our proposed method. Our DA model achieves comparable performance to the state-of-the-art on VQA 2.0.

6 Conclusion

We have presented a novel supervised multi-modal domain adaptation framework for open-ended visual question answering. Under the proposed framework, we have developed a new method for VQA which can learn a multi-modal feature em-

bedding that simultaneously keeps each domain invariant and each individual modality discriminative. We validate our proposed method on two popular VQA benchmark datasets, VQA 2.0 and VizWiz, in both directions of adaptation. The experimental results show our method outperforms the state-of-the-art methods.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *ICCV*.

Hedi Ben-younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.

Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. 2018. Detectron. https://github.com/facebookresearch/detectron.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. *IJCV*, pages 398–414.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *JMLR*, pages 723–773.

Yuhong Guo and Min Xiao. 2012. Cross language text classification via subspace co-regularized multi-view learning. In *ICML*.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. In *CVPR*.

Judy Hoffman, Eric Tzeng, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *ICCV*.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *ECCV*.

Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. In *CVPR*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *NeurIPS*.

Piotr Koniusz, Yusuf Tas, and Fatih Porikli. 2017. Domain adaptation by mixture of alignments of second- or higher-order scatter tensors. In *CVPR*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, pages 32–73.

Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. 2016. Feature pyramid networks for object detection. In *CVPR*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Fan Qi, Xiaoshan Yang, and Changsheng Xu. 2018. A unified framework for multimodal domain adaptation. In *ACM Multimedia*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*.

Fei Sha, Hexiang Hu, and Wei-Lun Chao. 2018. Cross-dataset adaptation for visual question answering. In *CVPR*.

Jian Shen, Yanru Qu, Weinan Zhang, and Yingrui Yu. 2017. Wasserstein distance guided representation learning for domain adaptation. In *AAAI*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *CVPR*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *CVPR*.

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*.

Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. 2015. Semi-supervised domain adaptation with subspace learning for visual recognition. In *CVPR*.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, (12):5947–5959.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. 2018. Learning to count objects in natural images for visual question answering. In *ICML*.

Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2015. Simple baseline for visual question answering. In *arXiv:1512.02167*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Fei Fei Li. 2016. Visual7w: Grounded question answering in images. In *CVPR*.