How to Practice VQA on a Resource-limited Target Domain

Mingda Zhang* Rebecca Hwa Adriana Kovashka
Department of Computer Science, University of Pittsburgh
{mzhang,hwa,kovashka}@cs.pitt.edu
https://cs.pitt.edu/~mzhang/practice-vqa/

Abstract

Visual question answering (VQA) is an active research area at the intersection of computer vision and natural language understanding. One major obstacle that keeps VQA models that perform well on benchmarks from being as successful on real-world applications, is the lack of annotated Image–Question–Answer triplets in the task of interest. In this work, we focus on a previously overlooked perspective, which is the disparate effectiveness of transfer learning and domain adaptation methods depending on the amount of labeled/unlabeled data available. We systematically investigated the visual domain gaps and question-defined textual gaps, and compared different knowledge transfer strategies under unsupervised, self-supervised, semi-supervised and fully-supervised adaptation scenarios. We show that different methods have varied sensitivity and requirements for data amount in the target domain. We conclude by sharing the best practice from our exploration regarding transferring VQA models to resource-limited target domains.

1. Introduction

Visual question answering (VQA) [1] aims at building algorithms to answer free-form, open-ended questions under the context depicted by an image. Tremendous efforts have been devoted to tackle the challenge, including collecting large-scale VQA datasets (*e.g.* GQA [14] contains 22 million annotated image–question–answer triplets) and developing increasingly complex and powerful deep neural networks (*e.g.* MCAN [34] and LXMERT [27] among many others [21, 7, 37]). As a result, state-of-the-art VQA models have achieved near 80% accuracy on some well-established benchmarks, alluring various attempts of migrating these powerful tools into daily life for solving real world challenges. Just imagine how convenient it might be for visually impaired (*e.g.* blind) people if a machine intelligence assistant can help answer questions for them [13, 12].

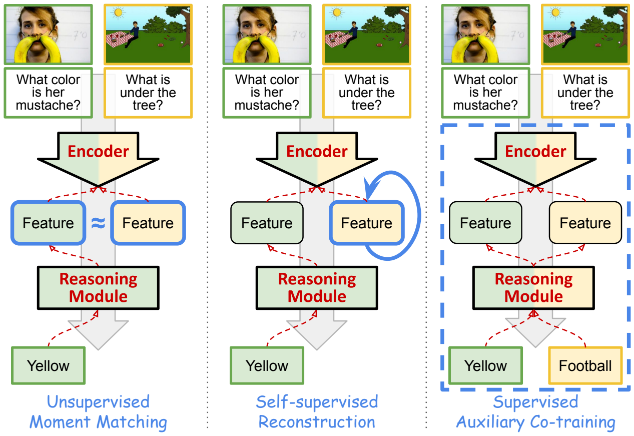

Figure 1. Illustration of different types of knowledge transfer in VQA we explored in Sec. 3.3. (Left) Unsupervised moment matching reduces domain discrepancy by enforcing feature alignment. (Middle) Self-supervised adaptation aims at reconstructing masked visual or textual features from remaining context. (Right) Supervised auxiliary co-training leverages target labels to ensure model compatibility with both source and target domains. All three can affect the *encoder*, but only the last gives supervision to the *reasoning module*. Gray arrows represent forward pass and red dashed arrows represent gradient backpropagation.

However, there is still a long way to go from benchmark excellence to real-world success. Specifically, a long-lasting challenge is that the target domain of interest may not have sufficient annotated data due to various factors (*e.g.* data privacy, high annotation cost, *etc.*), preventing these models from being directly trained for the task. According to a recent survey [2], application-specific VQA datasets usually have thousands of images and questions, one or two *magnitudes smaller* than academic general-purpose VQA datasets. A common strategy for mitigating data scarcity is *transfer learning*, which trains a model on related, large-scale, richly-annotated source domain before applying it to the target domain. Another related strategy is *domain adaptation*, which in particular addresses the data distribution change between source and target (*i.e. domain shift*).

Yet, VQA has some unique characteristics that make di-

[*] Work was done as a graduate student. Currently at Google Research.

rect transfer of knowledge non-trivial. For example, since multiple modalities are involved, the domain shift can occur in the visual, textual, or both modalities, and dataset-specific biases can exist in high-level semantic space or in low-level syntactic space [36]. With the recent emergence of Transformer-based modern neural networks, it could be even more challenging since the representation and reasoning are mingled throughout the network. Despite a few prior works attempting to build cross-dataset VQA models [5, 33, 36], some fundamental knowledge is still missing, for example, *what information about the target dataset, and how much of it, is minimally needed for a successful knowledge transfer.*

In this paper, we aim to fill this gap by focusing on knowledge transfer across domains in VQA. Specifically, we probed different adaptation strategies, *i.e.* unsupervised, self-supervised, semi- and fully-supervised (see Fig. 1), and compared their sensitivity to labeled or unlabeled data.

In addition to the appearance-related visual domain gaps across datasets, we also studied the sub-domains defined by the question types, *e.g. color* (“what color is her mustache?”), *object* (“what is under the tree?”). Intuitively, it requires different skills like counting or spatial reasoning to tackle different types of questions [32]. By training models with selected question types, we demonstrated how effective different strategies are with respect to the varied information needed for answering the questions. We attempted to answer the following questions regarding the VQA task:

- For different adaptation strategies, how much labeled or unlabeled data is minimally needed in the target domain for successful adaptation?
- Under a fixed data annotation budget what strategies could maximally improve model performance?
- Which modality is more relevant to the domain discrepancy in VQA? Where should the adaptation strategies be applied to minimize the domain shifts?
- When the questions require different skills for answering, what adaptation strategies are most effective to facilitate knowledge transfer? For the same collection of images, how much information can be leveraged if the questions are asked differently?

We conclude by sharing some lessons we have learned with future practitioners about how to build an effective VQA system on resource-limited datasets.

2. Background

Domain adaptation aims to overcome a practical limitation in machine learning, where the models trained from certain data may need to be applied on a different distribution. This technique has been proven successful on multiple applications such as object recognition [19, 20], machine translation [31, 30], *etc*. Various methods have been proposed to improve domain robustness, and these methods can be roughly categorized into three families, *i.e.* discrepancy-based methods (*i.e.* focusing on reducing domain discrepancies between source and target with different measures), adversarial-based methods (*i.e.* relying on a domain discriminator to adversarially encourage learning domain-invariant representations), and reconstruction-based methods (*i.e.* incorporating auxiliary tasks to bring source and target domain closer). Interested readers could check [9, 38, 24] for more complete reviews. However, most previous efforts are spent on *single-modality* domain adaptation. In this work, we experimented with different strategies, compared their data amount sensitivity to other transfer learning methods, including in sub-domains, and showed their effectiveness in the *multi-modal* VQA task. We individually probed the visual and textual modality and showed that bridging both is necessary for alleviating domain discrepancy.

Knowledge transfer in VQA. Partly because VQA requires models to correctly understand questions (*in text form*) and retrieve relevant cues from *visual context* to produce a prediction, applying domain adaptation for VQA is more challenging than in the single-modality setting. As shown in previous work [5, 33], even with a fully-annotated target dataset where answers (or decoys) are available, the performance boost by typical domain adaptation is usually limited. It is even more difficult in the unsupervised setting where no answers are available in the target dataset. The only work that attempts unsupervised adaptation in VQA is [36], but they only demonstrate their model across synthetic datasets with domain shifts only occurring in visual space. In fact, “on real dataset shifts” their models “only achieve marginal gains”. In this work we perform a systematic study for both unsupervised and supervised VQA domain adaptation strategies. In addition to visual and textual discrepancies, we also analyze the subtle differences across sub-domains defined by question types. We show that semi-supervised adaptation which exploits limited labeled samples together with a larger amount of unlabeled samples is most effective when target resources are limited.

Recent work shows inaccurate object detection might prevent VQA models from transferring across datasets [17]; we also observe when visual domain discrepancies are reduced the model shows more improvements compared to minimizing the text discrepancy. [16, 17] further examine how the answer space affect transferrability, while we focus on the impact of amount and type of supervision.

3. Overview of the Investigation Framework

For fair comparison, we systematically evaluate different knowledge transfer techniques on a unified investigation framework. We choose LXMERT and MCAN as two representative architectures: one requires massive pre-training and the other does not. By feeding varied amount of target information during training, the effectiveness of knowledge transfer can be measured by the target dataset accuracy.

3.1. Formal Formulation

Consider we have one *labeled* dataset $\mathcal{D}^S = \{d_i^S\}$ and one *sparsely labeled* dataset $\hat{\mathcal{D}}^T = \{\hat{d}_j^T\} + \{d_k^T\}$, where d^S/d^T represents image–question–answer triplets $\{v, q, a\}$ coming from source S and target domains T, respectively, and $\hat{d}^T$ represents image–question pairs $\{v, q\}$ *without answers*. Our goal is to build a visual question answering model $\mathcal{M}$ to answer questions from $\hat{\mathcal{D}}^T$. It is worth noting that we also assume $|\hat{\mathcal{D}}^T| \ll |\mathcal{D}^S|$, *i.e.* $\hat{\mathcal{D}}^T$ has much fewer samples than $\mathcal{D}^S$. This is because in real-world applications the task at hand usually comes with limited amount of samples.[1] Obviously the lack of ground-truth labels in our target dataset prevents directly training the model, thus the most feasible solution is to obtain relevant knowledge from the other dataset $\mathcal{D}^S$ and transfer the skill to $\hat{\mathcal{D}}^T$ despite the potentially large *domain gaps* between the two datasets.

In the following sections we will consider the *unsupervised* case where $\{d^T\} = \varnothing$, *i.e.* no target answers are available in the target domain, as well as a relaxed condition where a fraction of target samples come with correct answers. This is important in real-world applications where labeling a full dataset might be infeasible but a small portion of samples can still be annotated under limited budgets. In fact, we show that even a few annotations can play important roles probably *because they provide direct supervision for the VQA reasoning module (as shown in Fig. 1) in addition to the encoders.*

3.2. Benchmark: Data and VQA Methods

For most experiments we choose VQA-v2 and VQA-Abstract [11] for $\mathcal{D}^S$ and $\hat{\mathcal{D}}^T$, respectively. First, there exists large *visual domain gaps* between these two datasets [36] as the former takes images from everyday life while the latter is built upon clipart abstract scenes. Second, the answer space and distributions are similar, which excludes the potential impact by answer space shift (out of scope for this work). Last, we leverage the question type annotations provided in these datasets to create finer-grained sub-domains within each dataset, as different question types naturally demand varied skills to answer. Statistics about the two datasets are shown in Tab. 1. For completeness we also experiment with VQA-v2 and GQA [14]; their domain gap mainly lies in the language space.

To simplify the cross-dataset evaluation, we follow traditions to formulate the task as a multi-way classification, but carefully choose the answer vocabulary. Specifically, we merge all answer candidates from involved datasets and keep the top 1000 most frequent answers as labels.

Dataset	VQA-v2		VQA-Abstract	
	Train	Val	Train	Val
Color	59,838	2,506	5,356	2,700
Count	71,445	2,954	8,493	4,173
Location	13,314	610	2,935	1,488
Object	208,655	8,422	16,044	8,035
Reason	13,466	539	1,324	598
Verify	240,936	9,903	24,461	12,313
Others	24,463	1,060	1,387	693
Total	632,117	25,994	60,000	30,000

Table 1. Number of instances in VQA-v2 and VQA-Abstract by different question types. Overall the target dataset has only about 10% of samples (60K vs. 632K) compared with source. In our experiments we use even fewer samples to probe sensitivity.

We choose two mainstream VQA models as the base architectures in our explorations. MCAN [34] is the winner of VQA Challenge 2019, and serves as the backbone in the VQA Challenge 2020 winner model [22]. It can achieve competitive performance after training purely from a VQA dataset without relying on extensive pre-training from external data. This characteristic is particularly important for our study as it can effectively isolate the dataset and facilitate cross-dataset evaluation. We also choose LXMERT [27] as a representative of the transformer families, which have been dominating various vision-and-language benchmarks. It is worth noting that due to the massive pre-training, we need to carefully reconstruct the dataset (for pre-training and for fine-tuning) to avoid potential data leakage; otherwise the target dataset may be unintentionally exposed to the model thus the evaluation can no longer reflect performance against unseen domains. Specifically, we explicitly pre-train a custom checkpoint from scratch without using any Image–Question–Answer triplets; it is trained purely with image-text pairs from Conceptual Captions [25]. In other words, our pre-trained checkpoint does not use the VQA subsidiary task, and Image–Question–Answer triplets are only available in task-specific fine-tuning.

3.3. Different Settings for Knowledge Transfer

Since our goal is to achieve high accuracy on $\hat{\mathcal{D}}^T$, we leverage the labeled dataset $\mathcal{D}^S$ to empower the model for visual question answering capability, and add an auxiliary objective to ensure its knowledge is transferable to $\hat{\mathcal{D}}^T$.

$$L(\mathcal{D}^S, \hat{\mathcal{D}}^T; \theta) = L_{ce}(\mathcal{D}^S; \theta) + \lambda L_{aux}(\mathcal{D}^S, \hat{\mathcal{D}}^T; \bar{\theta}) \quad (1)$$

In this equation θ refers to the model parameters of the VQA model, and L_{ce} represents a traditional cross-entropy

[1] For example, popular medical VQA dataset ImageCLEF-2019 [3] has 16K Q&A pairs on 4K radiology images; VizWiz [13] contains 31K images/questions for assisting the visually impaired; the advertisement understanding dataset [15] involves 65K images and 200K Q&A pairs. As a comparison, GQA [14] has 22M Q&A on 113K images.

loss applied to $\mathcal{D}^S$ as it is the only source (in several settings) with ground-truth answers. L_{aux} is the auxiliary term which has different forms depending on the training paradigm, and $\bar{\theta}$ refers to related model parameters. We will elaborate more in the following subsections. λ is a weighting hyperparameter and we chose $\lambda = 1.0$ empirically.

3.3.1 Unsupervised Moment Matching

The moments of data distributions are known as important domain-specific features, and multiple different schemes have been proposed to match the moments between source and target, such as MMD [28, 20, 29], CORAL [26], CMD [35], HoMM [6], *etc*. Using moments to bridge distributions is most popular for vision-only applications, but recent work has shown it also effective for multi-modal tasks like visual question answering [36]. Therefore, when all samples from target dataset are unlabeled *i.e.* $|\{d_j^T\}| = 0$, we follow [36] and choose a simplified version of moment matching [23] as a general domain adaptation strategy to align the feature distributions. Specifically, we minimize *moment distances* as defined below and the auxiliary objective can be seen as additional regularization.

$$L_{aux} = d_{\text{moment}}(\mathcal{D}^S, \hat{\mathcal{D}}^T; \bar{\theta}) = \sum_{k=1}^{2} \left(\|\mathbb{E}(\mathbf{X}_S^k) - \mathbb{E}(\mathbf{X}_T^k)\|_2 \right) \tag{2}$$

Here $\mathbf{X}$ stands for the features after adaptation modules, k represents the moment order (k = 2 in our experiments), and $\bar{\theta}$ are the related parameters (*e.g.* some modules are listed in Table 2). The major advantage of moment matching is its general applicability and flexibility such that it can be used, plug-and-play, with a wide variety of model architectures. In our experiments, we apply it to both MCAN and LXMERT, and show immediate performance improvements across multiple dataset pairs. However, a limitation is also obvious: group-level moments are the only statistics used, thus much instance-level information is lost.

It is worth noting that we also experiment with domain adversarial loss as an alternative to encourage domain-agnostic feature representation. However, we observed similar trends with some previous work [36] that incorporating domain adversarial loss makes training highly unstable, making it difficult to serve as a probe.

3.3.2 Self-supervised Reconstruction

Self-supervised learning also learns from unlabeled samples, but its supervisory signals come from its input. By learning to predict some masked portion of the input based on the remaining portion of the input, a model can obtain knowledge, which is arguably even more effective than unsupervised training. This concept is popularized especially by the success of large pre-training language models like BERT [10] and GPT-3 [4], and recently some vision-and-language research also find it effective on tackling cross-modal tasks [21, 27].

This setting also applies to $\{d_j^T\} = \varnothing$ where no VQA answers are needed in the target dataset, but the auxiliary loss is formulated differently. Inspired by pre-training [27, 8], we added four auxiliary objectives for self-supervision *in fine-tuning*: masked language model (MLM), masked visual feature regression (MVFR), masked object classification (MOC) and masked attribute classification (MAC). MLM is for the textual domain where randomly masked question words need to be reconstructed based on the remaining input; the other three are for the visual domain (Masked Visual Learning, MVL), where MVFR requires masked visual features be reconstructed while MOC and MAC expect 1600-way object/400-way attribute classification from Faster R-CNN.

Note the model needs to predict answers for questions from the source dataset (enforced by L_{ce}) while simultaneously recovering the manipulated target information, thus the two objectives encourage the model to find a balance between adapting to the target and preserving question-answering capability. Although technically we simply moved the self-supervised objectives to the fine-tuning phase, it actually solves a problem in that the "pre-trained" checkpoint can effectively migrate to new domains without the need to incorporate the target domain *during pre-training*, but *during fine-tuning*. This makes it versatile because one cannot know the target domain during pre-training but a pre-trained checkpoint can always adapt to arbitrary new domains.

3.3.3 Supervised Auxiliary Co-training

In a relaxed condition where limited answers from the target dataset are available for training, a straightforward approach is to add an auxiliary training objective to enforce the model to predict answers for both source and target domains. The auxiliary loss can be defined as follows.

$$L_{aux} = L_{ce}(\{d_j^T\}) \tag{3}$$

Some prior works treat this as an "upper bound" for domain adaptation [5], as it leverages the ground-truth labels.

3.3.4 Semi-supervised Knowledge Transfer

Semi-supervised learning refers to combining a small amount of labeled samples with a large amount of unlabeled samples during training. We refer to applying either the moment matching approach or self-supervised learning approach together with the supervised approach, *e.g.*

$$L_{aux} = d_{\text{moment}}(\mathcal{D}^S, \hat{\mathcal{D}}^T; \bar{\theta}) + L_{ce}(\{d_j^T\}) \tag{4}$$

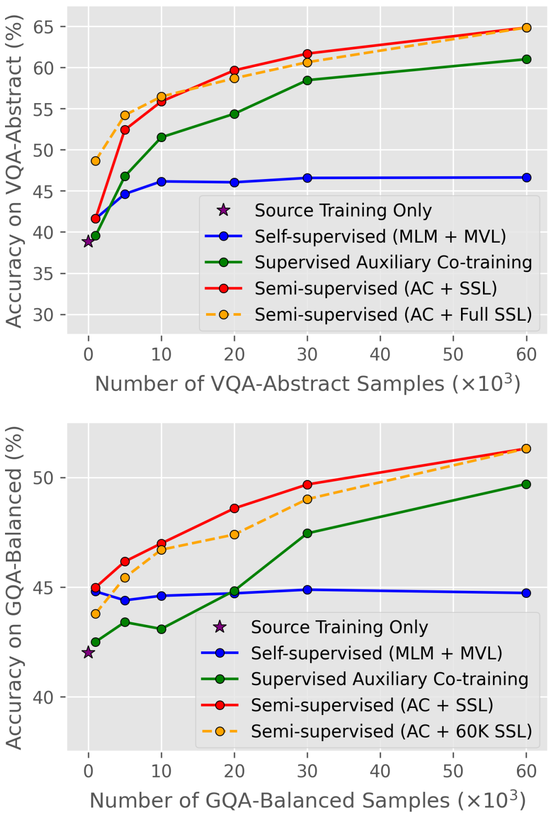

Figure 2. Accuracy for transfering LXMERT model to VQA-Abstract (top) and GQA-Balanced (bottom). We sampled 1K, 5K, 10K, 20K, 30K, 60K from the target dataset (d^T or $\hat{d}^T$) and incorporated these instances during training with different objectives (details in Sec. 3.3).

4. Experiments

Implementation Details. We modified MCAN and LXMERT[2] in PyTorch to support training with multiple datasets and various adaptation strategies. We trained all models for 100K steps with batch size 128 (20 epochs on source). We chose AdamW (w/ weight decay) as optimizer, and used linear schedule warm-up (10K steps) on learning rates with peak LR at 1e-4 and 1e-5 respectively. We ran all experiments on 8 Nvidia Quadro RTX 5000 GPUs.

4.1. Self-supervised Adaptation Complements Ground-truth Annotations

We first leverage the self-supervision signals and choose LXMERT on two different target datasets, VQA-Abstract and GQA-Balanced for experimental validations. To study the resource-limited scenario and investigate the sensitivity to data availability, we randomly sample differently-sized subsets from the target dataset (ranging from 1K to 60K instances, no more than 10% of the source dataset of size 632K) and add them to the training of VQA models. Note that for self-supervised adaptation, only image and questions from the target are used during training (but no answers), while for supervised and semi-supervised adaptation, we presume the answers are also available for training.

In Fig. 2 we show the overall accuracy comparison with LXMERT for transferring from VQA-v2 to VQA-Abstract (top) and GQA-Balanced (bottom). The x-axis represents the number of samples from the target, and y-axis the accuracy. As a naive baseline, we show the performance of a model trained solely on source dataset (*Source Training Only*) and directly evaluating on the target dataset at x=0 (purple star). In each figure we show self-supervision only, Sec. 3.3.2, blue lines), fully supervised adaptation (auxiliary co-training, Sec. 3.3.3, later abbreviated "AC," green line) and the semi-supervised adaptation (combination of the two, red line). We also show a different semi-supervised setting where the portion denoted on the x-axis are *labeled* samples used for answer supervision, and the majority (60K unlabeled target instances) only helps with self-supervised adaptation, shown by the orange dashed line.

Even 0.1% Unlabeled Samples Make a Difference. From both figures we observe that even 1K unlabeled samples from the target domain (only 0.16% of training data) could boost performance significantly (blue line, left-most point) compared to the source-only model. We also note that the improvement is very insensitive to target dataset size. In fact, even though it is promising to see that few unlabeled samples can be effective in facilitating transfer, the negative side is that the performance soon plateaus. Specifically for VQA-Abstract, models can achieve 46.2% accuracy with only 10K image–question pairs from the target dataset, but after 10K samples the model will not improve further. In other words, unsupervised adaptations without utilizing answer information has limited room for improvement, which may severely limit its application since the ultimate accuracy is still far from satisfactory. From our experiments this applies not only to self-supervision but also unsupervised adaptation like moment matching (figure not shown).

Supervised adaptations help given sufficient data. For supervised adaptation with auxiliary co-training, the models demonstrate slower (at the beginning) but more sustainable improvement with respect to the number of samples from the target dataset. With few thousands instances the performance is comparable to or even worse than unsupervised adaptation, but afterwards with more annotated data available the models keep improving steadily.

Which modules? One plausible explanation about the gap between the fully-supervised approach with unsupervised ones is the *modules under supervision*. As illustrated in Fig. 1 all methods could improve the domain robustness on the *encoder* but only the supervised training can improve the compatibility of *reasoning module*. However, one op-

[2] github.com/MILVLG/mcan-vqa, github.com/airsplay/lxmert

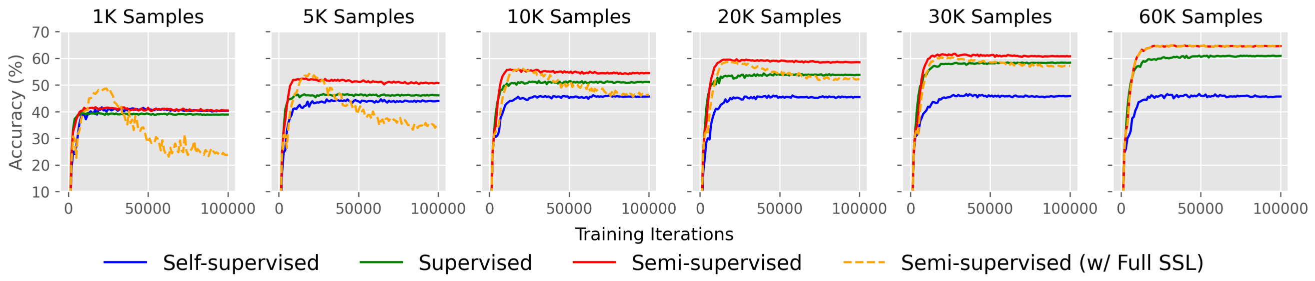

Figure 3. Training progress for LXMERT with different objectives, with 1K to 60K samples from various target datasets in training.

portunity exists: compared to encoders the reasoning module is usually lightweight especially in Transformers, thus we may not need as much as annotated data for adaptation.

Semi-supervised adaptation combines the best of both worlds. When combining the unsupervised strategy together with the supervised auxiliary co-training, we can exploit the information about target domain to its maximum. Even though supervised co-training is usually seen as an upper bound [5], we show that on LXMERT *combining both objectives demonstrates even stronger performance* than supervised adaptation, with a consistent boost in both dataset pairs. This implies that the *reconstruction-based self-supervision* may leverage a different signal which complements information from ground-truth VQA answers.

Model improves with more unlabeled data but also more easily overfits. We next analyze another practically meaningful scenario where large amount of unlabeled data might be easily obtained. Note that in research-oriented benchmarks usually both questions and answers are crowdsourced by dedicated annotators [11]. However, in real life the cost for collecting image–question pairs could be much cheaper than the image–question–answer triplets, as the pairs can be directly from real applications. For example in VizWiz [13, 12], questions are submitted by blind users together with the image, while the answers are crowdsourced later with human annotators. Learning from a large amount of unlabeled data together with very few annotated samples provides an opportunity to improve efficiency.

In Fig. 3 we show the entire training procedure with different number of target samples involved. The color of the lines corresponds to Fig 2. We mainly focus on two different semi-supervised adaptations, *i.e*. one relies on the same amount of samples for supervised and unsupervised adaptation (red line), and the other has access to a larger pool of unlabeled image–question pairs for unsupervised adaptation (yellow dashed line). The figure shows that when the ratio between unlabeled and labeled sample size is large (*i.e*. very limited labeled data), adding additional unlabeled data makes the model rapidly improve its performance at the beginning but later drop back due to overfitting. For example, in the left-most sub-figure in Fig. 3 the model takes only 1K labeled samples with 59K unlabeled ones, and the accuracy reaches 48.7% within 25K iterations before it falls back to 20%. With a held-out split, we can track the accuracy and apply early termination to avoid the performance loss caused by overfitting.

Model	Adapted Modules	Acc. (%)
MCAN	None	36.7
	Single-Modality Textual	36.6 (-0.1)
	Single-Modality Visual	35.8 (-0.9)
	Single-Modality V&T	37.8 (+1.1)
	Cross-Modal Textual	37.6 (+0.9)
	Cross-Modal Visual	38.9 (+2.2)
	Cross-Modal V&T	40.3 (+3.6)
LXMERT	None	38.8
	Single-Modality Textual	38.4 (-0.4)
	Single-Modality Visual	38.3 (-0.5)
	Single-Modality V&T	37.5 (-1.3)
	Cross-Modal Textual	38.8
	Cross-Modal Visual	38.7 (-0.1)
	Cross-Modal V&T	40.1 (+1.3)

Table 2. Applying moment matching to MCAN and LXMERT at different positions in the pipeline leads to varied performance. Refer to Fig. 4 for single-modality and cross-modal features.

4.2. Which Modules Should Be Aligned?

In most VQA models, visual and textual inputs are separately represented by corresponding encoders, which are fed to reasoning modules to merge information across modalities for final prediction (see Fig. 1). In Fig. 4 we highlight the major components used by different models such as LXMERT [27] and MCAN [34]. While in single-modality applications the feature discrepancy can be easily defined, in the multi-modal settings multi-level discrepancies can occur, and it is unclear which most affects performance. Therefore we aim to explore how domain discrepancies affect the knowledge transfer capability of VQA models.

We choose unsupervised moment matching as a general approach to probe the system at different positions. Specifically we attach the moment matching module after the vi-

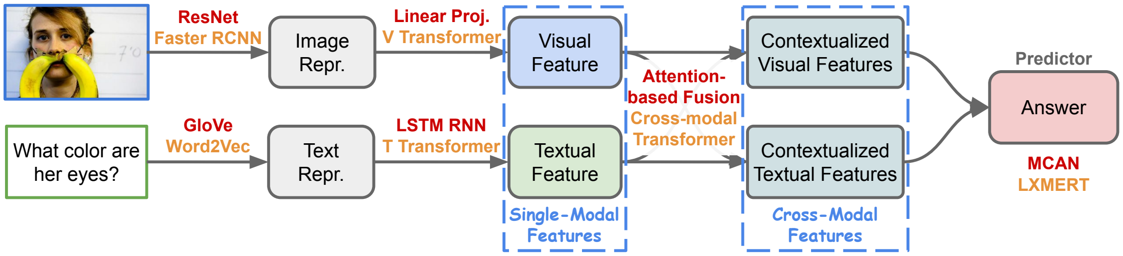

Figure 4. General illustration of VQA models. Image and text are separately encoded, then fed to a reasoning module to generate an answer for the given question under the visual scene. We experiment with aligning features at various positions to reduce domain discrepancies.

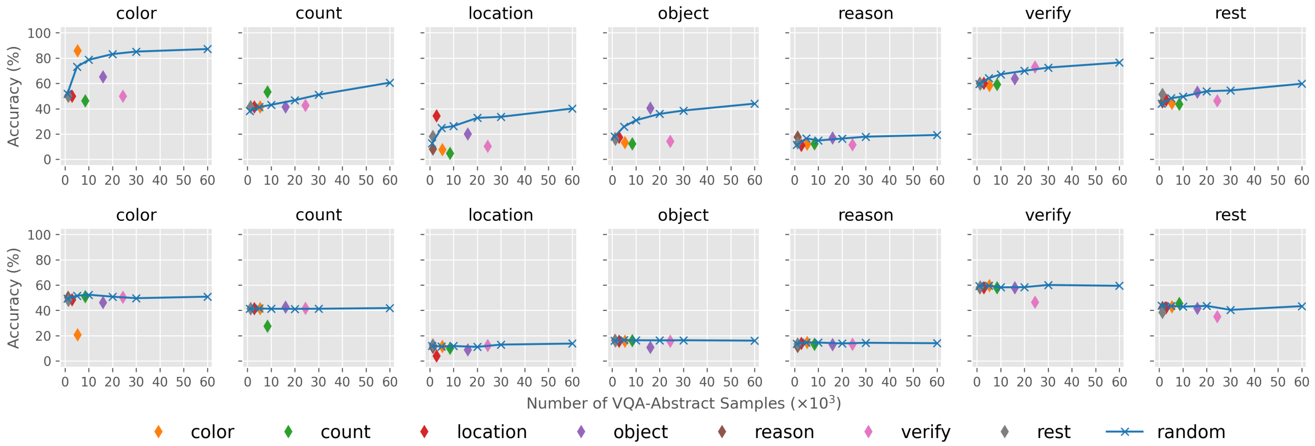

Figure 5. Supervised auxiliary co-training (top) and unsupervised moment matching (bottom) on MCAN models by different question types. When a matching sub-domain is selected during co-training the performance improvement is most significant. However, for moment matching the correspondent questions during training may even harm the performance.

sual and textual encoder to reduce feature discrepancy in the single modalities. Besides, we also investigate how things change after the cross-modal interaction by attaching the moment matching modules after it. In Tab. 2 we show the results (positions refer to Fig. 4). For cross-modal features, we see that matching encoded visual features shows more promises than textual features, *but when both modalities are aligned, the performance can be further improved.* This aligns with previous observations by [36]. Another important observation is that reducing discrepancies between the cross-modal features are generally more important than single-modality features. In fact, we even observe negative effects when single-modality features are aligned. One explanation is that even though the intermediate representations are aligned, it can still lead to domain-specific shifts in the remaining part of the pipeline, which is the arguably more important cross-modal module, *since it takes care of reasoning between two modalities*. Our suggestion for practitioners is to place the discrepancy-related domain adapters close to the prediction head to avoid degradation.

Again we note that the overall performance improvement from moment matching is less significant compared to self-supervised reconstruction, but it serves as a convenient tool to investigate the relationship between effective knowledge transfer and domain discrepancy.

4.3. Sub-domains Characterized by Question Types

If we consider the visual domain gaps between VQA-v2 and VQA-Abstract are relatively low-level (since they are mainly caused by the appearance distinctions of natural images and clipart abstract scenes), then the gap across question-defined sub-domains might be higher-level as they involve shifts in semantic meanings, even including the needs for varied type of information (intuitively for human beings the skill needed to answer *how many* would be different from *what color*). Therefore we also experiment with different adaptations on sub-domains each with questions relating to a single topic, e.g. *object*, *verify*, *color*, *location*.

In Fig. 5, each sub-figure represents the validation accuracy on a specific question type, and the blue line shows the baseline performance of models trained with randomly-sampled target instances. We compare to training various models with target data *all coming from the same question type*, and the performance is shown by different colored diamonds. Since the dataset contains different amount of questions in each type (see Tab. 1) we need to compare the accuracy with the random-selection reference (blue curve).

Matching Questions Facilitate Auxiliary Co-training. In

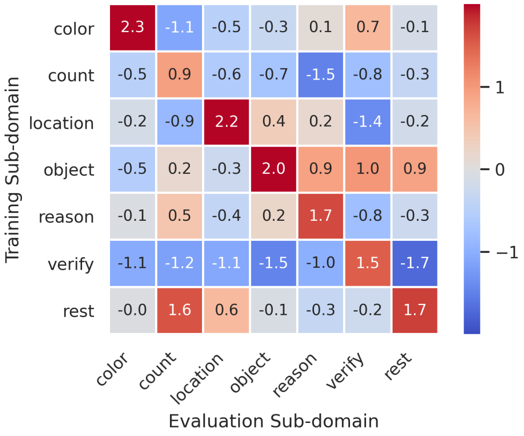

Figure 6. Training with question sub-domains. The number in each cell represents relative accuracy change when training and evaluating on different sub-domains. We observe the transferability differed greatly.

Fig. 5 (top) we show the supervised auxiliary co-training results, and clearly in all question types training with the corresponding samples give the most efficient performance gain. For example, for *color* questions *it requires only 5K color-related questions to achieve* 85% *accuracy, which is* 10 *percent more accurate than a model trained with 5K randomly-selected questions.* We also see other interesting patterns. For example, *location* is a highly specialized skill that requires dedicated samples to properly train a model. With only 2K location-related questions the model can achieve 30% accuracy on this task but if other types of questions are fed (*e.g. object*), the model can only get 15% accuracy even with 5 times more samples.

Degraded Performance with Moment Matching, Need for Diversity. In Fig. 5 (bottom) we show when unsupervised moment matching is applied the results are vastly different. In almost all question types, feeding corresponding samples during training has *negative* effects on the overall performance, *e.g.* feeding related samples in the target deteriorates models' capability on answering similar questions. One hypothesis is that matching moment statistics may require diverse data. This suggests that when the task at hand requires specialized skills, it might be worthwhile to collect corresponding answers for these questions and use them for supervised adaptation.

Skills Have Varied Difficulty for Transfer. In Fig. 6 we fixed the amount of data to sample (1000 in our experiments) from each sub-domain for supervised auxiliary co-training, and evaluated the performance on other sub-domains. To mitigate the effects that different sub-domains have varied difficulty by nature (*i.e. color* questions are easier than *reason* for training a model to answer), we normalized the table with mean and standard deviation within each sub-domain. We see *verify* questions provides very weak signals probably because most of these answers are either yes or no. A model trained with *verify* questions performs very poorly on all other question types, indicating that the model can extract only limited knowledge from *verify* questions. On the contrary, training with *object* questions is generally helpful to the model and in addition to the same category, the model also gets improvements on a few other sub-domains such as *verify*, *reason* and *count*. We also note that some knowledge is highly specialized that can hardly transfer without corresponding instances, such as *location* and *color*. Recall that models always have access to all type of questions from the source domain, but without the proper labeled target data they have troubles on the target domain. These observations give practical hints for future work when data collection can be guided, for example, *more concrete question and answers are preferred rather than asking for verification.*

5. Take-away Messages and Future Work

Based on our explorations, we want to share with future practitioners our lessons about how to perform better knowledge transfer for visual question answering tasks.

- Collecting more answer annotations is helpful as it provides the most straightforward supervision to the model. However, when the resource is constrained, it may not be necessary to expand the sample size in the target dataset, in particular if answers are not available.
- For discrepancy-based domain adaptation on VQA tasks, it is important that the matched features should be close to the final classification head which is directly responsible for generating the VQA predictions.
- If the expected target has a specific application or desired skill, such as counting, or spatial reasoning (*e.g.* location), then collecting questions with answers from the same category is most useful. On the contrary, if answer collection is not possible and unsupervised adaptation *e.g.* moment matching is the only viable option, then one should pay extra attention to diversity in the target domain.
- When free-form questions are accepted, *verify* type of questions seem most easy for human annotators to provide but unfortunately bring the least help to the model.

In future work, question rephrasing could be an important data augmentation strategy for existing datasets, as how the questions are raised [18] may affect knowledge transfer effectiveness. We will also consider alternatives to testing the importance of training different modules.

Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Nos. 2006885 and 1718262, and by University of Pittsburgh Momentum Funds.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[2] Silvio Barra, Carmen Bisogni, Maria De Marsico, and Stefano Ricciardi. Visual question answering: Which investigated applications? *Pattern Recognition Letters*, 151:325–331, 2021. 1

[3] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. VQA-Med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF2019 Working Notes*, CEUR Workshop Proceedings, 2019. 3

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4

[5] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4, 6

[6] Chao Chen, Zhihang Fu, Zhihong Chen, Sheng Jin, Zhaowei Cheng, Xinyu Jin, and Xian-Sheng Hua. Homm: Higher-order moment matching for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 4

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and JJ (Jingjing) Liu. Uniter: Universal image-text representation learning. In *16th European Conference Computer Vision (ECCV)*, 2020. 1

[8] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 4

[9] Chenhui Chu and Rui Wang. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, 2018. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 4

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 6

[12] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 6

[13] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 6

[14] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[15] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[16] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[17] Corentin Kervadec, Theo Jaunet, Grigory Antipov, Moez Baccouche, Romain Vuillemot, and Christian Wolf. How transferable are reasoning patterns in vqa? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[18] Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2042–2051, 2021. 8

[19] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015. 2

[20] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 2, 4

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 4

[22] Duy Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. Movie: Revisiting modulated convolutions for visual counting and beyond. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[23] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source

domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4

[24] Danielle Saunders. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75:351–424, 2022. 2

[25] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL, Volume 1: Long Papers)*, pages 2556–2565, 2018. 3

[26] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2016. 4

[27] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 1, 3, 4, 6

[28] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 4

[29] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S. Yu. Visual domain adaptation with manifold embedded distribution alignment. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, 2018. 4

[30] Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017. 2

[31] Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 2

[32] Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. Separating skills and concepts for novel visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[33] Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. Open-ended visual question answering by multimodal domain adaptation. In *Findings of the Association for Computational Linguistics: EMNLP*, 2020. 2

[34] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3, 6

[35] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (CMD) for domain-invariant representation learning. In *International Conference on Learning Representations (ICLR)*, 2017. 4

[36] Mingda Zhang, Tristan Maidment, Ahmad Diab, Adriana Kovashka, and Rebecca Hwa. Domain-robust vqa with diverse datasets and methods but no target labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 4, 7

[37] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[38] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, et al. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2020. 2