Domain Adaptation in Vision Question Answering

Tanmay Gejapati

Abstract

Domain adaptation in Visual Question Answering (VQA) presents significant challenges, especially when only a small number of labelled examples are available in the target domain. This work addresses the problem of adapting VQA models between the natural images of the VQA v2 dataset and the abstract scenes of the VQA abstract (abs) dataset. By analysing pretrained embeddings from DINOv2 and BERT, we uncover substantial domain gaps in both image and question spaces. To bridge this gap, we propose a model architecture utilizing criss-cross attention mechanisms to enhance the interaction between visual and textual features. We explore three training strategies: Naïve Sampling, Importance Sampling, and Domain Adversarial Networks (DANN). Our experiments demonstrate that the DANN approach significantly outperforms the others, effectively aligning feature distributions across domains and improving performance on the target dataset. This work establishes a baseline for domain adaptation in VQA tasks and highlights the potential for further research using advanced models like BEiTv3.

1 Introduction

Visual Question Answering (VQA) combines computer vision and natural language processing to answer questions based on visual content. While significant progress has been made with models trained on large-scale datasets, these models often struggle to generalize across different domains due to variations in image styles, question distributions, and answer vocabularies. This domain gap becomes particularly problematic when labeled data in the target domain are scarce, limiting the applicability of VQA systems in diverse real-world scenarios.

In this work, we focus on the challenge of domain adaptation in VQA with a limited number of labels. Specifically, we address the adaptation between the VQA v2 dataset, comprising natural images, and the VQA abstract (abs) dataset, which consists of synthetic, cartoon-like scenes. The significant differences between these datasets in terms of visual appearance and context pose a substantial domain gap that hampers cross-dataset generalization.

Our primary goal is to develop methods that effectively bridge this domain gap, enabling models trained on one dataset to perform well on the other with minimal additional labeled data. Achieving robust domain adaptation in VQA not only enhances model generalization but also reduces the reliance on extensive annotation efforts in new domains.

We make the following contributions:

- **Domain Gap Analysis**: We perform a thorough analysis of the pretrained embeddings from DINOv2 (for images) and BERT (for questions) to understand the extent of the domain gap between VQA v2 and VQA abs datasets.
- **Model Architecture**: We use a model architecture that incorporates criss-cross attention mechanisms. This design allows for enhanced interaction between image and text embeddings, facilitating better feature fusion.
- **Training Strategies**: We explore and compare three training methods for domain adaptation:
    1. **Naïve Sampling**: Combining samples from both datasets without any weighting.

2. **Importance Sampling**: Assigning weights to samples to emphasize more relevant examples and reduce domain shift.
3. **Domain Adversarial Networks (DANN)**: Utilizing adversarial training to align feature distributions across domains by making them indistinguishable to a domain classifier.

- **Experimental Validation**: Our experiments demonstrate that the DANN approach significantly improves cross-domain performance compared to the other methods. We provide quantitative results and t-SNE visualizations to illustrate the effectiveness of our approach.
- **Comparison with State-of-the-Art Models**: We compare our methods with BEiTv3, a state-of-the-art model in vision-language modelling, to contextualize our findings and highlight the challenges in domain adaptation for VQA.

The remainder of this report is organized as follows: In the next section, we review related work in domain adaptation and VQA. We then present our experimental analysis of pretrained embeddings and discuss the proposed methods to bridge the domain gap. We provide detailed experimental results and comparisons, followed by conclusions and directions for future work.

2 Related Work

Domain adaptation is a well-studied problem in machine learning, aiming to transfer knowledge from a source domain to a target domain with different data distributions. In computer vision, techniques such as feature alignment, adversarial training, and instance reweighting have been employed to mitigate domain shifts in tasks like image classification and object detection [1][2].

In the context of Visual Question Answering, most research has focused on improving model architectures within a single domain. Notable works include the use of attention mechanisms to better fuse visual and textual information [3][4], and leveraging large-scale datasets to train robust models [5]. However, the issue of domain adaptation in VQA has received limited attention.

Some studies have explored transfer learning approaches, fine-tuning models pretrained on large datasets for specific target domains [6]. While effective to some extent, these methods often require substantial labelled data in the target domain, which may not be feasible in practice. Additionally, they may not adequately address the complex interplay between visual and textual modalities inherent in VQA tasks.

Adversarial training methods, such as Domain Adversarial Neural Networks (DANN), have shown promise in learning domain-invariant features [7]. By introducing a domain classifier with a gradient reversal layer, the model learns to extract features that are discriminative for the main task while being indistinguishable between domains. While DANN has been applied in various fields, its application in multimodal tasks like VQA is less explored.

Our work bridges this gap by integrating domain adversarial training into a multimodal VQA model with criss-cross attention mechanisms. By doing so, we aim to align the feature distributions of image and text embeddings across domains, even with a small number of labeled examples in the target domain. This approach differs from previous works by focusing on the multimodal nature of VQA and addressing domain adaptation without heavy reliance on target domain annotations.

3 Experiments

3.1 Analysing Pretrained Embeddings

Embeddings for the images and questions were generated from pretrained DINOv2 and BERT respectively. The following are t-SNE plots of these embeddings.

3.1.1 VQA v2

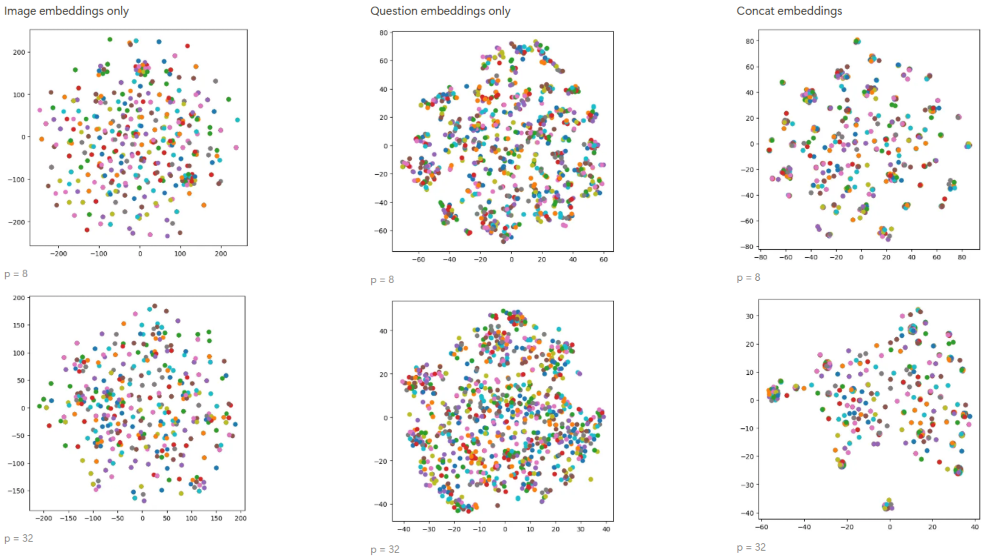

Figure 1: t-SNE plots of embeddings for VQA v2 dataset

Description:

- `p=8` and `32` are the `perplexity` parameter for the t-SNE plots, higher perplexity reflects more global comparison among the embeddings. These two numbers were chosen to show local and global comparison between the embeddings, more on this in the observations noted below.
- `Concat embeddings` is simply channel-wise concatenation of DINOv2 and BERT generated embeddings.

Observations:

- **Image space:** At both levels of t-SNE plot, it looks pretty consistent, which implies the image space is quite homogeneous. The lumps are most likely because of repeated images and t-SNE not converging them down to a single point like it ideally should be.
- **Question space:** Again, this space has pretty homogeneous embeddings, one thing to note is the scale of image space's plots and that of question space, likely due to the size of embedding vectors. But the point density is higher meaning unlike the images, the questions are indeed more unique.
- **Concatenated space:** Pretty consistent density across the space with some clusters scattered around.

3.1.2 VQA abs

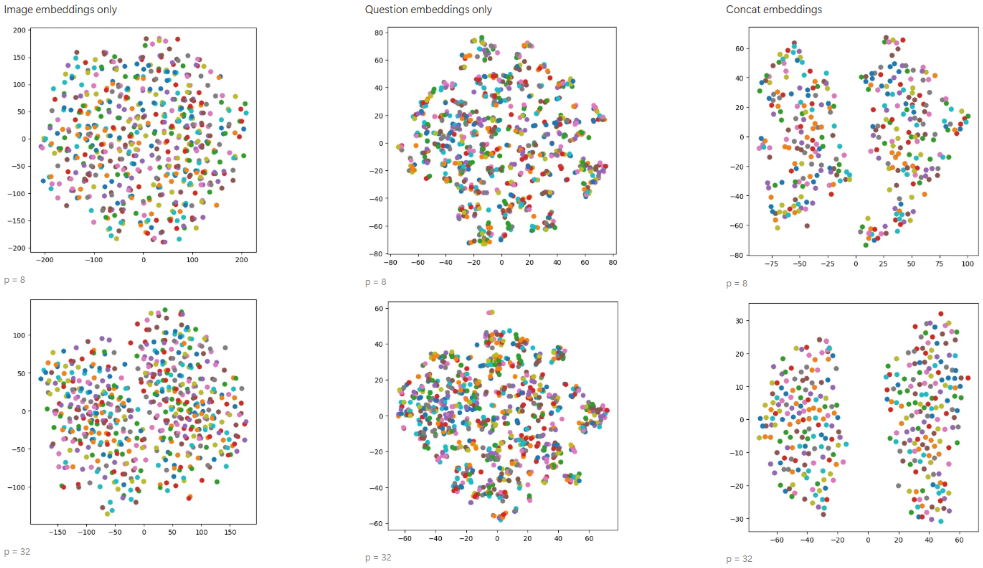

Figure 2: t-SNE plots of embeddings for VQA abs dataset

Observations:

- **Image space:** p=32 shows a faint boundary between two major clusters while within the cluster, points seem to be evenly spread out.
- **Question space:** Just like VQA v2, this is mostly uniform too but with a higher visible point density compared to the corresponding image space.
- **Concatenated space:** There is a very distinct boundary between two clusters with uniform density inside each of them. This is likely due to the fact that image space has a visible but faint boundary that has been enhanced by concatenation with question space. And the major difference in abstract's images is the "indoor" and "outdoor" setting having very obvious traits like walls being almost consistently brown and outdoor being a bright blue sky.

3.1.3 Combined plots

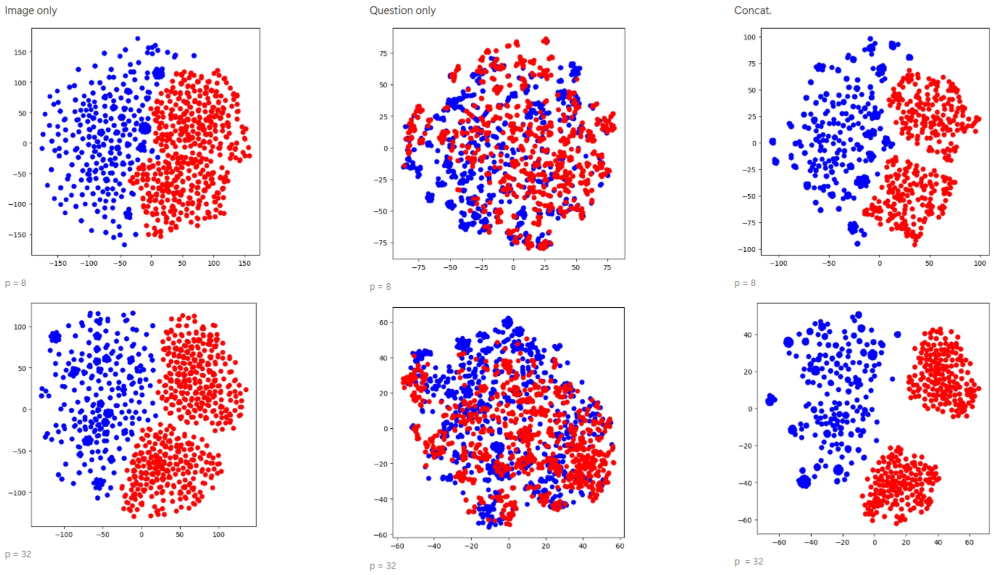

Figure 3: Combined t-SNE plots of embeddings for VQA v2 and VQA abs datasets

Description: The blue and red points correspond to VQA v2 and abs. respectively.

Observations:

- **Image space:** Clear distinction between v2 and abs. and also two major clusters in that of abs. implying pretrained DINOv2 is overly focusing on ”syntactic” over ”semantic” details.
- **Question space:** No clear distinction, which implies the questions are pretty similar across these two datasets.
- **Concatenated space:** Just like image space, this too has clear distinction between the datasets and the indoor-outdoor settings in VQA abs.

3.2 Methods to Bridge the Domain Gap

From the analysis done above, it's clear that there exists a domain gap and it's unreasonable to expect good performance on VQA abs. when trained on VQA v2. To bridge this domain gap, the following methods are presented.

3.2.1 Model Architecture

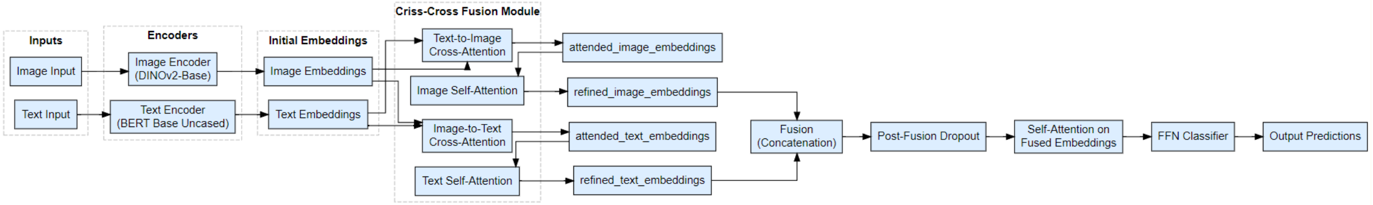

Figure 4: Model architecture with criss-cross attention mechanisms

Description: Frozen image and text encoders are first passed through cross-attention layers individually and are concatenated channel-wise after self-attention refinement. These fused embeddings are then further refined by self-attention before passing through a classification head.

Motivation for criss-cross attention: Just concatenating the initial embeddings doesn't capture the complex relationship mapping required for specific tasks asked by the question. Hence, to enhance relevant aspects of the image embeddings, such a mechanism was crucial.

3.2.2 Training

Three kinds of training methods were used to optimally perform domain adaptation:

1. **Naïve sampling**: Using both the datasets' samples for training directly.
2. **Importance sampling**: Selectively weighs samples from both datasets to reduce domain shift by emphasizing more relevant examples.
3. **Domain adversarial network**: Uses adversarial training to align feature distributions across domains by making them indistinguishable to a domain classifier.

Note: These methods were the starting point to baseline domain adaptation for the VQA task. Domain adaptation turns out to be less explored in the literature of VQA, our further work would involve finding methods that outperform this baseline.

Note: The following experiments were done for the top 5 overlapped answers as a proof of concept. Results for larger answer space are presented later into the report.

3.2.3 Naïve Sampling

Training Dataset	Training Accuracy (%)	Validation Accuracy on Self (%)	Validation Accuracy on Other (%)
v2	33.50	37.00	35.00 (abs)
abs	48.25	44.00	53.00 (v2)

Table 1: Results using Naïve Sampling

The above results are statistically significant because at random, 20% accuracy can be expected. Interestingly, having v2 as source domain hurts the performance which is less expected as general intuition suggests v2 is a harder problem than abs. Hence, solving v2 should make it easier to solve abs. but it's quite the opposite.

3.2.4 Importance Sampling

Training Dataset	Training Accuracy (%)	Validation Accuracy on Self (%)	Validation Accuracy on Other (%)
v2	74.00	70.00	75.00 (abs)
abs	20.00	20.00	20.00 (v2)

Table 2: Results using Importance Sampling

Clearly, by this method, the model didn't converge likely due to unstable training as loss is directly influenced by bad priors of prediction over two distinct domains. Due to its unstable nature, further efforts weren't taken to enhance its accuracy as something so simple as naïve sampling did better than this.

3.2.5 Domain Adversarial Network

Training Dataset	Training Accuracy (%)	Validation Accuracy on Self (%)	Validation Accuracy on Other (%)
v2	80.75	71.00	74.00 (abs)
abs	73.25	72.00	68.00 (v2)

Table 3: Results using Domain Adversarial Network

This method clearly outperformed the rest and will be used as a baseline for further domain adaptation analysis.

Inference: This method gives a more logically consistent result that training on v2 gives better performance on abs. as compared to the other way around.

Training Dataset	Training Accuracy (%)	Validation Accuracy on Self (%)	Validation Accuracy on Other (%)
v2	49.23	23.38	27.08 (abs)

Table 4: DANN results with 65 answer classes

More Classes DANN was used to train the model over top 65 overlapping answers. This will act as our baseline for further experiments. The following t-SNE plot describes the effectiveness of this approach.

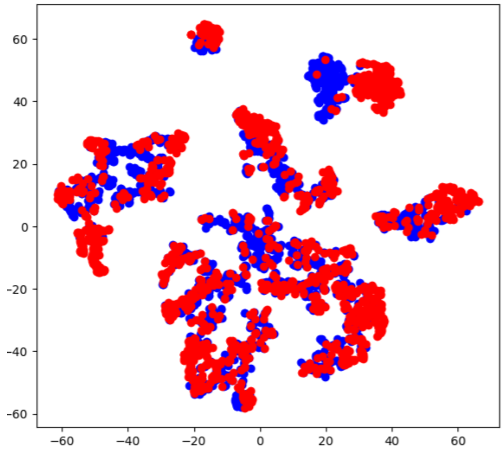

Figure 5: t-SNE plot showing embedding overlap after DANN training

Description: Red and blue are validation embeddings on v2 and abs. respectively.

Inference: These embeddings have much more overlap compared to the analysis done previously on pretrained embeddings. This suggests our DANN method does indeed achieve its objective of being robust against domain variations.

3.3 Comparison with BEiTv3

BEiTv3 is one of the state-of-the-art models in vision language modelling.

Source and Target	Description	Validation Accuracy (%)	Comments
v2 → v2	Base model open weights	77.65	
abs → abs	Fine-tuning	79.16	
v2 → abs	Direct validation on abs (after one epoch probing of classification head)	71.85	Really strong domain adaptation capability
v2 → v2	Training from scratch (4 epochs)	39.03	Pretty hard problem since over 3000 answers are possible
abs → abs	Training from scratch (4 epochs)	39.12	Achieves comparable results to v2 with fewer iterations (abs is a smaller dataset) meaning it's a relatively easier problem

Table 5: Comparison of results using BEiTv3

Overall, BEiTv3 is a much better model at naturally being domain robust hence we aim to further test all our approaches on this model to cancel the effects of model capacity and focus more towards rapid domain adaptation.

4 Further Work

Due to time constraints, this RnD effort has not yet matured into a complete research paper. However, several promising avenues remain to be explored to enhance the domain adaptation capabilities of our Visual Question Answering (VQA) model. The following points outline our planned future work:

1. **Integration of BEiTv3 and Expansion to Additional Datasets**
   - **Adopting BEiTv3 as the Primary VQA Model**: Building upon our current architecture, we aim to integrate BEiTv3, a state-of-the-art model in vision-language modelling, as the backbone of our VQA system. This integration is expected to leverage BEiTv3's robust feature extraction and multimodal understanding capabilities.
   - **Incorporating Diverse Datasets**: To validate and generalize our domain adaptation techniques, we plan to extend our experiments to additional VQA datasets such as VQA Viz and COCO-QA. This expansion will allow us to assess the model's performance across varied domains and image styles, ensuring that our approach is not limited to the VQA v2 and VQA abstract (abs) datasets alone.
2. **Implementation of Consistency Regularization Techniques**
   - **Visual Domain Consistency**: We intend to apply consistency regularization over the visual domain by utilizing advanced prompt-based models. Specifically, we will explore methods to transfer the style of images from the VQA v2 dataset to resemble those in the VQA abs dataset. This style transfer aims to create augmented images that maintain the semantic content while varying stylistic attributes.

- **Advanced Prompt-Based Augmentations**: Leveraging the capabilities of models like ChatGPT-4o, we plan to develop sophisticated prompt-based approaches. By providing the model with access to image, question, and answer triplets, we can generate variations of images that preserve the correctness of answers. This strategy is expected to enforce consistency in the model's responses despite variations in image styles, thereby enhancing its robustness to domain shifts.

3. **Exploration of Advanced Domain Adaptation Methods with Limited Labels**

   - **Few-Shot Domain Adaptation**: While our current study focuses on baseline domain adaptation methods, future work will delve into more complex techniques tailored for scenarios with a limited number of labelled examples in the target domain. Approaches such as meta-learning, adversarial feature alignment, and self-supervised learning will be investigated to improve the model's adaptability with minimal supervision.
   - **Semi-Supervised and Unsupervised Adaptation**: In addition to supervised methods, we aim to explore semi-supervised and unsupervised domain adaptation strategies. These methods can further reduce the dependency on labelled target data by leveraging unlabelled samples and intrinsic data structures, thereby enhancing the model's generalization capabilities across domains.

Appendix

Training Config

Five Answer Classification

```
Labels: ['yes', 'no', 'man', 'bench', 'sunny']
V2:  Train size = 400 | Val size = 100 | Total = 500
Abs: Train size = 400 | Val size = 100 | Total = 500
Training Epochs: 30
```

65 Answer Classification

```
Labels shown under the section [DANN Approach Details]
V2:  Train size = 131351 | Val size = 1625 | Total = 132976
Abs: Train size = 43837 | Val size = 1625 | Total = 45462
Training Epochs: 30
```

Methods Used to Stabilize Training

- **Regularizers** used to help with *overfitting*:
  - Batch normalization, dropout, weight decay
  - Image augmentations: masked patches
- Methods used to manage *underfitting*:
  - Label type classification, gives an additional mask over output classes to the final classification head making it easier for the model to narrow down to the precise answer.

DANN Approach Details

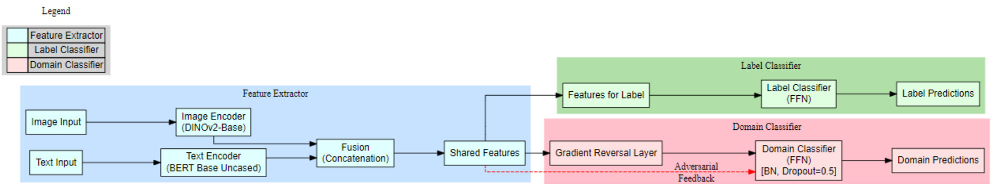

Figure 6: Domain Adversarial Neural Network (DANN) Architecture

Label types used:

```
label_type_mapping = {
    'colors':         ['beige', 'black', 'blue', 'brown', 'gray', 'green', 'orange', 'pink',
                         'red', 'red and white', 'tan', 'white', 'yellow'],
    'objects':        ['bench', 'chair', 'couch', 'floor', 'table', 'tv', 'blanket', 'book',
                         'frisbee', 'skateboard', 'soccer'],
    'living-things': ['baby', 'bird', 'boy', 'cat', 'dog', 'fish', 'flowers', 'girl', 'man',
                         'mouse', 'tree', 'woman'],
    'actions':        ['eating', 'playing', 'sitting', 'sleeping', 'standing', 'walking'],
    'locations':      ['park', 'sidewalk', 'living room', 'on table', 'sky'],
    'foods':          ['apple', 'pizza', 'sandwich', 'wine', 'food'],
    'numbers':        ['0', '1', '2', '3', '4', '5', '6'],
    'responses':      ['no', 'no one', 'nothing', 'yes'],
    'directions':     ['left', 'right'],
}
```

Domain Classifier:

- **Gradient Reversal Layer**: The shared features pass through a gradient reversal layer that inverts gradients during backpropagation to enable adversarial training.
- **Domain Classifier (FFN)**:
  - This classifier predicts the domain (source or target) of the input.
  - Includes **batch normalization** and **dropout (0.5)**, with repeated layers for enhanced performance.
- **Output**: The domain classifier produces domain predictions.

Adversarial Feedback:

- The dashed red arrow represents adversarial training:
  - The domain classifier provides feedback to the feature extractor.
  - The feature extractor is optimized to learn **domain-invariant features** by minimizing the label loss and maximizing the domain confusion.

Working Repository: `https://github.com/tanmay4269/Few-Shot-VQA` (currently private, will be public right after completion of work)

References

[1] Gopalan, R., Li, R., & Chellappa, R. (2011). Domain adaptation for object recognition: An unsupervised approach. *2011 International Conference on Computer Vision.*

[2] Long, M., Cao, Y., Wang, J., & Jordan, M. I. (2015). Learning transferable features with deep adaptation networks. *International Conference on Machine Learning.*

[3] Anderson, P., He, X., Buehler, C., et al. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[4] Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. *Advances in Neural Information Processing Systems.*

[5] Antol, S., Agrawal, A., Lu, J., et al. (2015). VQA: Visual Question Answering. *Proceedings of the IEEE International Conference on Computer Vision.*

[6] Shah, M., & Koltun, V. (2018). Robust domain adaptation in visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[7] Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. *International Conference on Machine Learning.*