Improving Robustness of VLMs using LLMs

Kanishk Jain, Nathan Cormerais, Youry Macius
Université de Montréal
{kanishk.jain, nathan.cormerais, youry.macius}@umontreal.ca

Abstract

We tackle the problem of out-of-distribution robustness of vision-language models for the task of visual question answering. Current state-of-the-art vision-language models showcase exceptional performances on answering visual questions, closing the gap with human performance on benchmark datasets. However, these models struggle significantly when faced with questions that fall outside their training distribution. They produce incorrect answers by exploiting spurious statistical regularities of the training data. Our study aims to reduce the reliance on spurious correlations in vision-language models by leveraging large language models. Specifically, we plan to explore solutions based on fine-tuning and in-context prompting for improving robustness of the underlying vision language models.

1 Introduction

Advancements in hardware capabilities and the availability of large-scale datasets have together ushered a new era of progress in the fields of computer vision and natural language processing. The emergence of Large Language Models (LLMs) have closed the gap and in certain cases exceeded human performance in tasks like question answering, text summariztion, sentiment analysis, machine translation etc. On the other hand, Vision Language Models (VLMs) have been popularized recently for various vision-language tasks like image recognition, visual question answering, image captioning, visual grounding and image-text retrieval. However, recent works (Chen et al., 2023) have shed light on vulnerability in robustness of VLMs on changing the distribution of image and text.

We study the robustness of VLMs to the changing data distribution for the task of visual question answering. Despite achieving strong performance on benchmark VQA datasets, these models struggle in face of questions out of the training distribution. A recent work (Dancette et al., 2021) found that models for VQA leverage shortcuts in the dataset in the form of biases and answer question without utilizing the visual context in the image. We believe that large VLMs also suffer from the same problem but they come into light in face of out-of-distribution textual prompts.

In this project, we aim to improve the robustness of VLMs on out-of-distribution questions. We plan to study how data-augmentation techniques in visual and linguistic modalities affects the robustness of VLMs in task of VQA. Specifically, we leverage regularization techniques like Cut-Mix and Mix-Up in the visual domain, and the LLM-based paraphrasing of questions with modifications like replacing the subject, adding word modifiers and changing question style in the textual domain. We investigate how different combinations of data augmentation techniques (image only, text only, and both image and text) affect the fine-tuning performance of these models for the task of visual question answering. We also experiment with different fine-tuning approaches like incontext prompting (Wei et al., 2022b) and parameter-efficient finetuning using LORA (Hu et al., 2022).

To summarize, our contributions are: (1) We study effectiveness of data-augmentation in visual and linguistic modalities for improving robustness of VLMs on VQA task, (2) We analyse the robustness of different fine-tuning based approaches on out-of-distribution questions in VQA, and (3) We use LLM-based evaluation metric for the VQA to overcome the weakness of existing string-matching based evaluation criteria.

2 Related Work

Large Language Models (LLMs) are deep neural architectures with billions of parameters and are pre-trained on vast amounts of data using masked

language modelling objective. LLMs are usually stacked with deep layers of transformer based encoder-decoder blocks and powered with self-attention mechanism. Pre-trained LLMs (Wei et al., 2022a; Jiang et al., 2023) exhibit exceptional generalization capabilities on various downstream tasks in both generative and discriminative settings. In this work, we utilize LLMs for imrpoving robustness of VLMs.

Vision Language Models (VLMs) are multimodal architectures which accept both image and textual prompts as input and generates response to the query in text. These architectures are usually pre-trained on large-scale datasets (Sharma et al., 2018; Schuhmann et al., 2022), which results in strong generalization on various downstream tasks like image captioning (Xu et al., 2015), visual question answering (VQA) (Antol et al., 2015) and image-text retrieval (Luo et al., 2022). Depending on the pre-training objective, we have different VLMs, like CLIP (Radford et al., 2021) which is trained on contrastive learning objective between image-text pairs. ViLT (Kim et al., 2021) propose a vision-and-language transformer without convolution or region supervision, which represents a class of models that leverage transformer architectures without relying on conventional convolutional neural networks (CNNs) or predefined region features for processing visual inputs. Following the success of (Radford et al., 2021), BLIP (Li et al., 2022) effectively utilizes the image-caption pairs by bootstrapping the captions. Recently proposed, Llava (Liu et al., 2023) combines a vision encoder along with a Large Language Model (LLM) and performs instruction tuning on the resulting VLM.

Robustness in VLMs: Growing application of VLMs has resulted in many recent works Chen et al. (2023); Dancette et al. (2021) that scrutinize the robustness of VLMs to changes in input image and text. Chen et al. (2023) indicate that real-world test samples often differ from the data used for adaptation during fine-tuning, highlighting the need for improved robustness. They find that full fine-tuning doesn't consistently provide better robustness and that adapter-based methods can sometimes outperform them. In addition, Dancette et al. (2021) identify that shortcut learning during the training phase reduces the robustness of VLMs as it enables the model to leverage shortcuts based on dataset biases while minimizing the learning objective. They introduce VQA-CounterExamples (VQA-CE) to detect shortcuts in model performance. Similarly, recent works like Gu et al. (2023); Zhou et al. (2023) propose techniques to improve robustness of VLMs. Gu et al. (2023) propose a more robust version of Prompt Learning (ProL) by integrating multiple-scale image features into the prompt. In addition, Zhou et al. (2023) propose a framework combining knowledge distillation and data augmentation to enhance the robustness of VLMs. They introduce Discrete Adversarial Distillation (DAD) as a method to transfer knowledge from larger to smaller models, improving their performance on new data. In this work, we focus on the VQA task for evaluating the robustness of VLMs.

3 Methods

Given an image I and a corresponding question Q about the image, the goal is to generate a free-form answer A in response to the question. For this task, we utilize pre-trained vision-language models (VLMs) that are capable of accepting an image and a textual query as inputs and outputting a textual response to the query. We utilize two kinds of transformer-based VLMs for this work: (1) Vision-and-Language Transformer Without Convolution or Region Supervision, ViLT, and (2) Llava. We describe each architecture in detail in the section 3.1. In this work, we want to improve the robustness of VLMs to out-of-distribution questions Q'. We leverage large-language models (LLMs) to generate out-of-distribution questions and describe the steps in detail in section 3.2. Finally in section 3.3, we describe our approach to improve the robustness of VLMs on out-of-distribution questions.

3.1 ViLT architecture

The Vision-and-Language Transformer (ViLT) (Kim et al., 2021) architecture marks a novel direction in Vision-and-Language Pre-training (VLP) by adopting a convolution-free approach for processing visual inputs. The network architecture is illustrated in figure 2. Central to the ViLT architecture is a simplified visual embedding strategy, where a linear projection method is applied to image patches in a manner similar to text tokens, eliminating the necessity for CNNs or object detection-based feature extraction in the visual embedding process. Moreover, ViLT employs a unified transformer model for processing both visual and textual inputs, treating the two modalities in a parallel fashion. This unified approach simplifies the model's architecture and enables easier

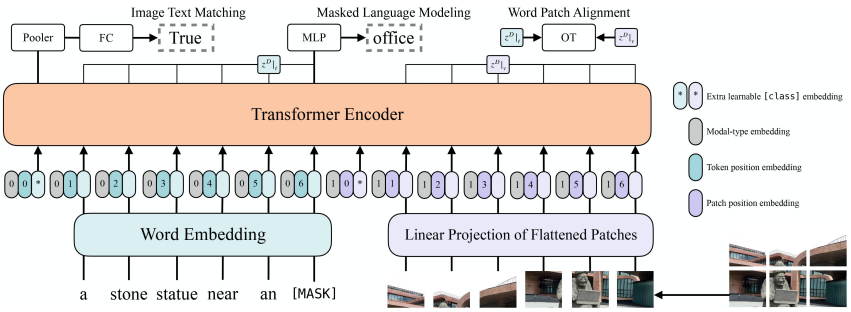

Figure 1: Model architecture for ViLT

interactions between modalities.

3.2 OOD Question Generation

Despite achieving impressive performance on benchmark Visual Question Answering (VQA) datasets, current Visual Language Models (VLMs) struggle to answer questions that fall outside their training distribution. However, given these models are trained on extensive datasets, generating questions that are truly out-of-distribution can be challenging. To tackle this issue, we employ large language models for generating out-of-distribution questions. Specifically, we utilize a variant of Mistral-7B, known as OpenHermes 2.5, for this purpose. We employ zero-shot and few-shot prompting techniques to generate out-of-distribution questions by paraphrasing existing ones for enhanced clarity and diversity. This method leverages the sophisticated understanding and generation capabilities of large language models to produce questions that differ significantly from those seen during training.

3.3 Improving the Robustness of VLMs

To enhance the robustness of Visual Language Models (VLMs), we explore various data augmentation techniques within both visual and linguistic modalities. In the visual domain, we employ image regularization techniques such as CutMix and MixUp. These methods, originally developed for the image classification literature, are utilized to improve the robustness of image classification models. Specifically, by introducing variations in the visual data through these techniques VLMs are expected to learn more general visual representations which should help the model in their image understanding and as a result reduce the error rate on out-of-distribution questions. In the linguistic domain, we enhance the training set by incorporating questions generated by a Large Language Model (LLM) and subsequently fine-tune the VLM using these augmented questions. This approach not only diversifies the linguistic input but also prepares the VLM to better handle a wide array of question formulations, further contributing to its overall robustness in performing VQA task.

4 Experiments

4.1 Dataset

We utilized the dataset introduced by Goyal et al. (2017), which comprises over 214K (question, image) pairs for the validation split. This dataset is selected to evaluate the performance of our VLM because, as described in the paper, it represents a *balanced* version of the original VQA dataset. Notably, each question is linked to a pair of similar images that yield two distinct answers to the same question. On average, each image is associated with 5.4 questions.

4.2 Baselines

Our primary baseline is the ViLT model, pretrained and sourced from Hugging Face. In addition to this baseline, we conduct experiments with various data augmentation techniques in both visual and linguistic domains, training the model from scratch with these enhancements. Furthermore, we explore the effects of fine-tuning the model exclusively with the questions generated during our experiments.

4.3 Evaluation Methods

To evaluate the robustness of the models, we use the answer accuracy as the evaluation metric. However, relying solely on string matching to determine if a predicted answer aligns with the ground truth can lead to inaccuracies. This is because string matching might not recognize correct answers that are semantically similar but not identical in form. For instance, if the model responds with "one" to

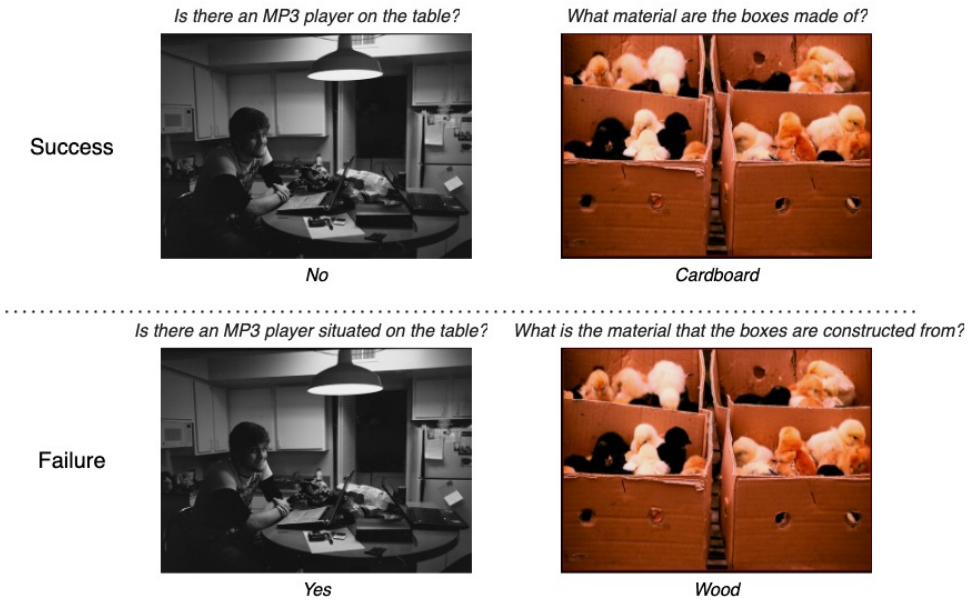

Figure 2: Qualitative results for in-distribution (top row) and out-of-distribution (bottom row) examples

a counting question, but the ground truth is represented by the numeral "1," string matching would unfairly mark a correct prediction as incorrect. To address these limitations, we turn to LLMs known for their advanced language comprehension. By leveraging LLMs, we can more accurately assess whether a predicted answer is semantically equivalent to the ground truth, thus providing a fairer and more reliable evaluation of model performance.

4.4 Experimental Details

For the experiments, we used a VLP (Vision-and-Language Pre-training) model named Vision-and-Language Transformer (ViLP). This model introduced in Kim et al. (2021) has shown a lot of potential as it is faster than previous VLP models. We used a ViLT model that was finetuned on the VQAv2 train and validation sets, however, the authors of the paper introducing ViLT mentioned that it doesn't perform as well as other VLP models with a heavy visual embedded on the VQA dataset. The authors explain that it could be because the questions from the VQA dataset are usually about objects.

4.5 Results

Using the ViLT model on the VQA-v2 dataset has demonstrated promising results, as evidenced by the high accuracy rate of 72.21% achieved on the VQA-v2 evaluation set. However, as detailed in the evaluation method section, reliance on simple string matching has a significant drawback: it overlooks the semantics of the answers. To address this, we also calculated the ViLT model's accuracy for VQA using an LLM-based evaluation metric. This approach yielded an average accuracy of 96% across a randomly selected subset of approximately 35,000 questions from the evaluation set. These findings indicate that the simple string matching-based metric significantly underestimates the model's true performance. However, as illustrated in Figure 2, the model's response can vary depending on how the question is posed, which might lead to incorrect answers. This variability demonstrates that the models encounter challenges with out-of-distribution questions, particularly noticeable in different types of questions, such as *yes/no* queries and others, where the model generates diverse answers based on the question's format.

4.6 Future work

For the midway report, we have evaluated the ViLT model on the VQA-v2 dataset and implemented an LLM-based evaluation metric to compute answer accuracy. Additionally, we utilized an LLM for paraphrasing questions using the zero-shot prompting technique. For the final presentation, we plan to implement visual data augmentation techniques such as CutMix and MixUp, and conduct comprehensive training of the ViLT model using both visual and linguistic augmentations. This will help us verify whether these augmentations improve the robustness of VLMs to out-of-distribution questions generated by the LLM.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. 2023. Benchmarking robustness of adaptation methods on pre-trained vision-language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1574–1583.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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)*.

Jindong Gu, Ahmad Beirami, Xuezhi Wang, Alex Beutel, Philip Torr, and Yao Qin. 2023. Towards robust prompts on vision-language models. *arXiv preprint arXiv:2304.08479*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Ziyang Luo, Yadong Xi, Rongsheng Zhang, GongZheng Li, Zeng Zhao, and Jing Ma. 2022. Conditioned masked language and image modeling for image-text dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 130–140, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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 (Volume 1: Long Papers)*, pages 2556–2565.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Andy Zhou, Jindong Wang, Yu-Xiong Wang, and Haohan Wang. 2023. Distilling out-of-distribution robustness from vision-language foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems*.