Adapting ViLT for GQA: Visual Question Answering with the Vision-and-Language Transformer on the GQA Dataset

Shashwath Santhosh, sks272 [1] **Keshav Shivkumar, ks1830** [1] **Goutham Swaminathan, gs982** [1]

Abstract

In this work, we explore the adaptability and performance of the Vision-and-Language Transformer (ViLT) (Kim et al., 2021) when applied to a novel dataset and task. ViLT is a minimal VLP (Gan et al., 2022) model that simplifies the processing of visual inputs in a convolution-free manner, addressing challenges related to efficiency and expressive power found in traditional VLP models. We modify the existing ViLT model and its codebase for Visual Question Answering (Antol et al., 2015) on the GQA dataset, diverging from its original application on the VQAv2 dataset (Goyal et al., 2017). Our adjustments facilitate training and evaluation on the GQA dataset (Hudson & Manning, 2019), yielding an overall accuracy of 72.93% and a binary accuracy of 76.44%. Although direct comparison between the datasets is not appropriate, our work demonstrates the flexibility and potential of the ViLT model for various vision-and-language tasks. Our modified ViLT code, tailored for the GQA dataset, is available for further exploration and development, offering valuable insights into the model's potential in different contexts.

1. Introduction

Vision-and-Language Pre-training (VLP) has rapidly advanced the state of the art in various joint vision-and-language downstream tasks. The recent emergence of Transformer-based architectures (Vaswani et al., 2017), such as BERT (Devlin et al., 2019), GPT (Radford et al., 2019), and ViT (Dosovitskiy et al., 2021), has revolutionized the natural language processing and computer vision domains, demonstrating impressive results across a wide range of tasks. However, traditional VLP models often depend on image feature extraction processes (Kumar & Bhatia, 2014), involving region supervision and convolutional architectures such as InceptionNet (Szegedy et al., 2015), VGG (Simonyan & Zisserman, 2014), ResNets (He et al., 2016), leading to challenges in efficiency and expressive power.

In this context, the Vision-and-Language Transformer (ViLT) was introduced as a minimal VLP model, designed to address these limitations. ViLT simplifies the processing of visual inputs by employing a convolution-free approach, similar to how textual inputs are processed. This monolithic model leverages the Transformer architecture, successfully applied in various natural language processing tasks, and extends it to the visual domain. Consequently, ViLT offers improved efficiency and expressive power compared to traditional VLP models.

We chose to work with ViLT due to its potential for adaptability and generalization, as well as its promising performance in the original paper. Moreover, its streamlined architecture and the availability of pre-trained weights and code made it an attractive choice for our investigation. The original implementation of ViLT focused on Visual Question Answering (VQA) using the VQAv2 dataset, showcasing its capability to handle complex vision-and-language tasks effectively.

In this project, we explore the application of ViLT to the GQA dataset, a large-scale dataset for real-world visual reasoning and compositional question answering. The GQA dataset provides a unique challenge due to its emphasis on spatial and relational reasoning, as well as multi-step inference, making it a suitable testbed for ViLT's adaptability and performance.

Our primary goal is to modify the existing ViLT model and its codebase to perform Visual Question Answering on the GQA dataset. By making necessary adjustments to the code, we enable the model to train and evaluate on this new dataset, assessing its performance and potential improvements. In addition, we aim to provide valuable insights into the ViLT model's capabilities and limitations, and offer suggestions for future research directions.

This project contributes to the growing body of research on vision-and-language tasks, demonstrating the adaptability and potential of the ViLT model for a wide range of

[1]Department of CS, University of Rutgers, NJ, USA. Correspondence to: Keshav Shivkumar <ks1830@scarletmail.rutgers.edu>.

applications. Our work not only showcases the model's performance on the GQA dataset but also highlights the importance of developing flexible and efficient VLP models for tackling complex, real-world problems in the intersection of computer vision and natural language processing.

2. Related Work

MDETR (Kamath et al., 2021) is a modulated detection system built on the DETR architecture, which is an end-to-end object detection model that uses a convolutional backbone followed by a Transformer Encoder-Decoder. The model is designed for tasks like referring expression comprehension, segmentation, visual question answering, and phrase grounding. MDETR takes advantage of the pre-trained transformer language model for text encoding and projects both image and text features into a shared embedding space. MDETR can handle multiple tasks, making it a versatile solution for visual question answering and other related tasks. The use of a pre-trained transformer language model for text encoding allows MDETR to leverage the power of language models in understanding the text inputs effectively. However, MDETR relies on a heavier convolutional backbone and a separate Transformer Encoder-Decoder, unlike ViLT, which has a more lightweight and unified approach to handling visual inputs. Moreover, MDETR's architecture might be relatively more complex, as it includes modality-dependent linear projections and a joint transformer encoder.

The LXMERT (Learning Cross-Modality Encoder Representations from Transformers) (Tan & Bansal, 2019) paper presents a model that aims to learn joint representations of images and text through a transformer-based architecture. The LXMERT model consists of separate encoders for visual and textual modalities, as well as a cross-modality encoder that combines the output of the two modality-specific encoders. Both LXMERT and ViLT are designed for tasks involving visual and textual inputs, such as visual question answering. However, their approaches to handling these inputs differ significantly. ViLT simplifies the processing of visual inputs using a minimal visual embedding pipeline, making it more lightweight and computationally efficient. In contrast, LXMERT employs separate encoders for visual and textual modalities, which can result in a more complex architecture and increased computational requirements. Yet LXMERT has demonstrated strong performance in the domain of VQA. ViLT's simplicity and efficiency make it well-suited for real-world applications, while LXMERT's rich cross-modal representations can enable it to excel in tasks that require a deep understanding of both visual and textual information.

ViLT aims to simplify the processing of visual inputs by adopting a convolution-free approach similar to the one used for processing textual inputs. This results in a more lightweight and computationally efficient model compared to other vision-and-language models. By leveraging the Vision Transformer architecture and pre-training on large-scale datasets, ViLT achieves competitive or better performance on downstream tasks, such as VQA, compared to previous VLP models. In the VQAv2 dataset, ViLT has shown strong performance in the visual question answering task. Its simplified architecture and unified approach to processing visual and textual inputs make it well-suited for real-world applications, offering a more computationally efficient alternative to more complex models. While ViLT's simplified architecture has its benefits, it may not capture certain complex cross-modal relationships as effectively as models with separate encoders for visual and textual inputs, like LXMERT. ViLT's reliance on the Vision Transformer architecture, which was initially designed for image classification tasks, might not be optimal for all vision-and-language tasks.

3. Implementation Details

To adopt the ViLT methodology into GQA, the implementation was modified in each of the steps:

- **Preprocessing**: The GQA dataset differs from VQAv2 in terms of structure and the number of unique question-answer pairs. To adapt the ViLT model for GQA, preprocessing modifications are necessary which include loading data, tokenizing text inputs, and transforming images. This also includes handling the differences in JSON structures and adjusting the model to accommodate 1878 unique question-answer pairs in GQA, compared to 3129 pairs in VQAv2. Data splitting, batching, and shuffling for different stages of the experiment (training, validation, and testing) are handled.
- **Model architecture**: The core architecture of ViLT remains unchanged, as it is based on the Vision Transformer (ViT). It unifies the processing of visual and textual inputs using a minimal visual embedding pipeline. The model takes in visual and textual embedding sequences as input and produces a contextualized feature sequence as output. The ViT's weights pretrained on ImageNet are used in the model. The final model architecture is shown in Fig 1.
- **Fine-tuning**: The model is fine-tuned on the GQA dataset using a combination of masked language modeling (MLM) and image-text matching objectives. The text tokens are masked, and the model predicts the ground truth labels from the corresponding contextualized vector. ViLT utilizes a 2-layer MLP MLM head, and the MLM loss is computed as the negative log-likelihood loss for the masked tokens, similar to BERT's MLM objective. The model is fine-tuned on

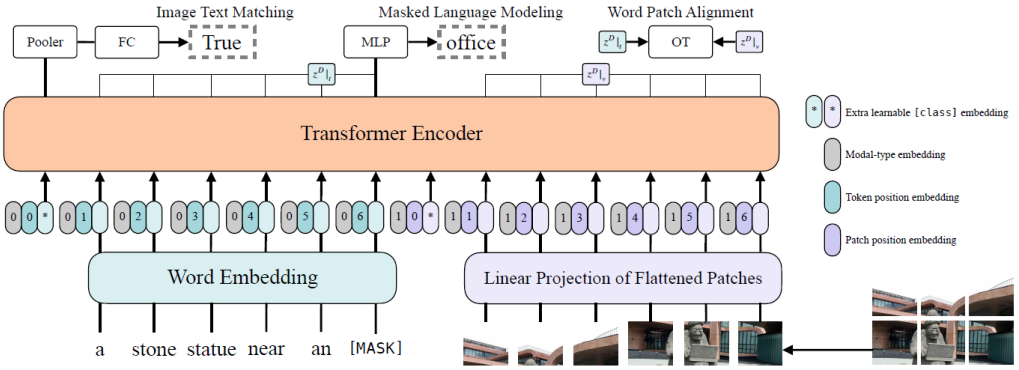

Figure 1. ViLT Architecture as proposed in the original paper

the GQA dataset for several epochs (e.g., 10 epochs), with each epoch taking around 3 hours. Validation checks are performed every 10% of an epoch to monitor the model's performance and avoid overfitting.

- **Hyperparameters**: The hyperparameters are adopted from the VQAv2 implementation. This includes a learning rate of 1e-4, binary cross-entropy loss, 10 epochs, and a batch size of 32. Validation checks are performed every 10
- **Evaluation**: The model's performance is evaluated on the GQA dataset, focusing on its effectiveness in the visual question answering task. Accuracy measurements are taken for overall accuracy and binary accuracy (for yes/no questions).

Figure 2. ViLT Evaluation Example, with the test question and model prediction for the provided image.

4. Results

Upon fine-tuning the ViLT model with the GQA dataset, the results exhibited the model's effectiveness and versatility in addressing visual question answering tasks across different datasets. With an overall accuracy of **72.93%** on the GQA dataset, the model demonstrated its proficiency in interpreting and responding to questions related to images. This performance is notable, as can be seen in Figure 2, considering the more intricate structure of the GQA dataset and its distinct set of unique question-answer pairs compared to VQAv2.

Furthermore, the model achieved a binary accuracy of **76.44%** for yes/no questions, emphasizing its ability to accurately discern and respond to binary queries within the GQA dataset. The successful adaptation of the ViLT model to the GQA dataset highlights its flexibility, adaptability, and potential applicability to a wide range of visual question answering tasks on various datasets. These outcomes also underscore the advantages of employing a unified, convolution-free method for processing both visual and textual inputs, contributing to the model's efficiency and generalizability.

4.1. Limitations

While the ViLT implementation on the GQA dataset has shown promising results, it is not without limitations. One significant constraint is that the model is unable to fully exploit the more impressive features of the GQA dataset, such as long answers and contextual information. This may lead to sub-optimal performance in certain situations where deeper understanding and reasoning are required. Furthermore, the current implementation may not fully leverage the rich structure of the GQA dataset, which contains various question types and answer choices. This suggests that there is room for improvement in terms of accuracy and model robustness. Additionally, the preprocessing and fine-tuning methods could be further optimized to better adapt to the unique characteristics of the GQA dataset. Finally, the computational and time requirements for training and fine-tuning on GQA remain a challenge, potentially limiting

its applicability to real-world scenarios and its ability to scale efficiently.

4.2. Code

The unofficial implementation of the project is available at: https://github.com/keshavshivkumar/ViLT.

Conclusion

ViLT demonstrates its adaptability and potency as a model for vision-and-language tasks, showcasing impressive results on standard datasets such as VQAv2 and GQA. By capitalizing on the advantages of transformer architectures and fusing visual and textual information, ViLT is capable of handling an extensive array of tasks with remarkable precision. As the domain of vision-and-language research progresses, models like ViLT will be instrumental in creating intelligent systems that can comprehend and deduce the intricate connections between images and text.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., and Gao, J. Vision-language pre-training: Basics, recent advances, and future trends, 2022.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019. doi: 10.1109/CVPR.2019.00686.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kim21k.html.

Kumar, G. and Bhatia, P. K. A detailed review of feature extraction in image processing systems. In *2014 Fourth International Conference on Advanced Computing Communication Technologies*, pp. 5–12, 2014. doi: 10.1109/ACCT.2014.74.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL http://arxiv.org/abs/1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015. doi: 10.1109/CVPR.2015.7298594.

Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information*

Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.