Sentinels: Visual Question Answering using Transformers

Jashwanth Sajja, Rohit Dhaipule and Rutwik Segireddy

Introduction

Visual Question Answering (VQA) is a transformative task that integrates visual and textual data interpretation to answer questions about images. This technology significantly enhances accessibility for visually impaired individuals, enriches educational tools, and improves customer service experiences. It stands as a crucial development in the realm of natural language processing (NLP), targeting the efficient integration of diverse data types.

The Vision-and-Language Transformer (ViLT) is pivotal in advancing VQA by processing visual inputs directly and similarly to textual data, eliminating the need for complex and resource-heavy feature extraction. Focusing primarily on yes/no and numeric questions, ViLT overcomes the substantial computational inefficiencies inherent in traditional methods. This approach not only simplifies the integration process but also enhances the effectiveness and applicability of VQA systems.

Background

Visual Question Answering (VQA) is a challenging task that requires the integration of computer vision and natural language processing to answer questions about an image. Transformers have been influential in advancing VQA research in recent years.

Transformers utilize self-attention mechanisms to capture long-range dependencies in the input data, which is particularly useful for VQA where the question and image need to be jointly understood. ViT, for example, has been shown to outperform convolutional neural networks on various visual tasks by directly applying the Transformer architecture to images. BERT, on the other hand, has been widely used for language understanding and can be effectively combined with vision models for VQA. The advantages of these Transformer-based models include their ability to handle complex reasoning and their strong generalization capabilities. However, training Transformer models can be computationally expensive and time-consuming. We have explored techniques like freezing the lower layers of the Transformer during training, which can significantly reduce training time without compromising performance.

Data

We employed the VQA dataset from "Making the V in VQA Matter" by Goyal et al. (2017), designed for tasks involving yes/no and numeric questions. The dataset contains over 3 million tokens and totals more than 18GB, segmented into training and validation/test sets to facilitate a thorough assessment of the model's capabilities.

The training set includes 166k instances for yes/no questions and 46k for numeric questions. The validation and test sets contain 80k and 22k instances respectively, each labeled for supervised learning. This organization ensures a comprehensive training and testing environment.

Description	Yes/No	Numeric
Training Set	166,878	46,406
Validation/Test Set	80,537	22,131
Total Tokens	2,237,396	831,241
Data Size	14.3 GB	5 GB

This structured dataset supports the development and evaluation of models that efficiently integrate visual and textual data to answer diverse question types.

Methods.

Our study aimed to investigate the impact of freezing multiple layers in the ViLT (Vision-and-Language Transformer) model on training time and performance. The ViLT model, renowned for its ability to seamlessly integrate visual and textual information, was fine-tuned using the COCO dataset, which features images paired with corresponding questions and

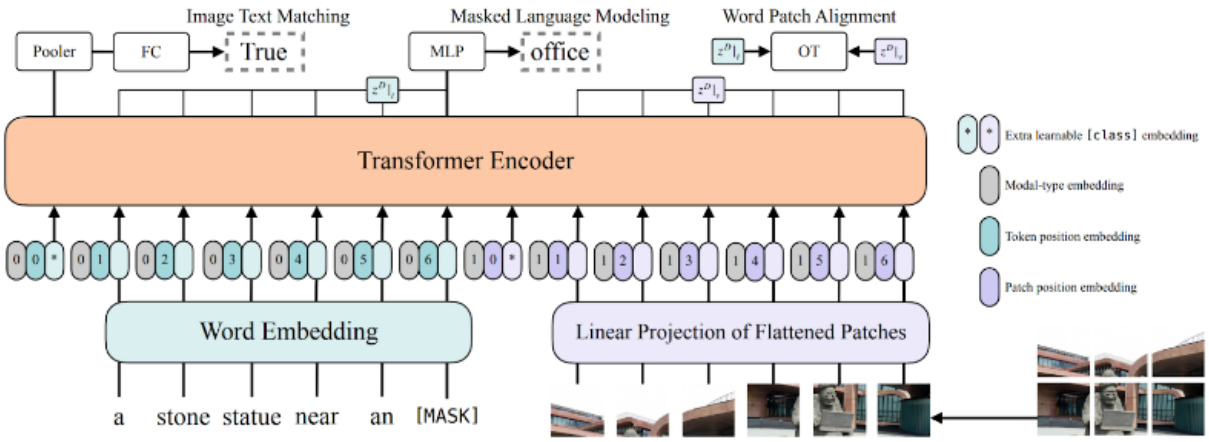

Figure 1 - ViLT Architecture

annotations. This dataset served as the bedrock for training two distinct subsets: one tailored for binary (yes/no) classification and another for numerical predictions. For binary classification tasks, we augmented the ViLT architecture with an additional classification layer positioned just before the output layer. Conversely, for numerical predictions, we filtered output values to be less than 10 and employed regression techniques for model training and validation.

Model Architecture and Training

Harnessing the Hugging Face Transformers library, we instantiated the pre-trained ViLT model and initialized its weights using pre-existing parameters. Our experimental design centered on systematically freezing layers within the ViLT architecture, beginning from the top and gradually moving towards the bottom. This approach allowed us to evaluate the impact of restricting updates to specific layers on overall model performance. By iteratively fine-tuning the model with varying degrees of layer freezing, we aimed to discern patterns in training time and accuracy, providing insights into the optimal balance between computational efficiency and predictive power.

Framework and Hyperparameters

PyTorch served as our primary framework for implementing experiments, offering robust support for deep learning tasks and facilitating seamless integration with the ViLT model. Key hyperparameters such as learning rate, batch size, number of epochs, and layer freezing configurations were meticulously tuned through a combination of grid search and cross-validation techniques. This iterative optimization process aimed to strike an equilibrium between model convergence speed and final performance metrics, ensuring that the trained ViLT models exhibited both efficiency and effectiveness in addressing the respective tasks.

Data Population and Assumptions

The COCO dataset, renowned for its rich diversity and extensive annotations, provided the foundational corpus for our experiments.

Leveraging its broad spectrum of images and accompanying textual annotations, we assumed that the dataset accurately represented real-world scenarios, enabling us to draw meaningful insights and generalize findings to broader domains.

However, we acknowledge potential biases inherent in the dataset, such as overrepresentation of certain object classes or linguistic patterns, and accounted for these factors in our experimental design and analysis.

Results and Evaluation.

From the Tables, we can observe that as the number of frozen layers increases, the accuracy initially increases until a certain point and then decreases if we compare only the first epoch. This trend can be attributed to the fact that for models with a large number of training parameters, a single epoch is not sufficient for the model to

Model Card - ViLT (Vision-and-Language Transformer)

Model Details

- The ViLT model, a fusion of vision and language transformer architectures, is trained to integrate visual and textual information for diverse multimodal tasks.
- Utilizes the COCO dataset, comprising images paired with questions and annotations, as training data.
- Developed using the Hugging Face Transformers library.

Intended Use

- Intended for a wide array of applications including visual question answering, and multimodal reasoning tasks.
- Not designed for real-time inference in safety-critical systems or fully autonomous decision-making.
- Not intended to make judgments about specific individuals or demographic groups.

Factors

- Multimodal inputs encompassing both image features and textual questions/annotations.
- Effect of frozen layers in training, optimum number of layers to be frozen to find the balance for training time vs performance.
- Number of layers(out-of-12) to be frozen to seek out the best trade-off

Metrics

- Performance evaluated using standard metrics such as accuracy and F1 score for classification, MSELoss for regression.

Training Data

- Primarily trained on the COCO dataset, which contains images with associated textual annotations including questions and answers.
- Two types of data - yes/no ground truths for classification, numerical (0-9) ground truths for regression.

Evaluation Data

- Used the validation data in the COCO dataset provided by the VQADatasets to evaluate the performance.

Caveats and Recommendations

- Since, we need to create a classification problem, we ignored the samples that have answers other than yes/no.
- Filter the output range to be between 0-10, for better performance and the error blows up if there is any answer, other than that in the specified range.

Figure 2 - Model Card for the partially-frozen ViLT transformer model

parameters effectively learn and optimize them.

For both classification and regression tasks, the performance of the ViLT model stops improving noticeably after unfreezing certain layers. For classification, performance plateaus after unfreezing layer 6, with improvements up to this point. Similarly, for regression tasks, performance ceases to improve after unfreezing layer 7. This suggests that further unfreezing does not significantly enhance the model's capability, potentially leading to underfitting or diminishing returns.

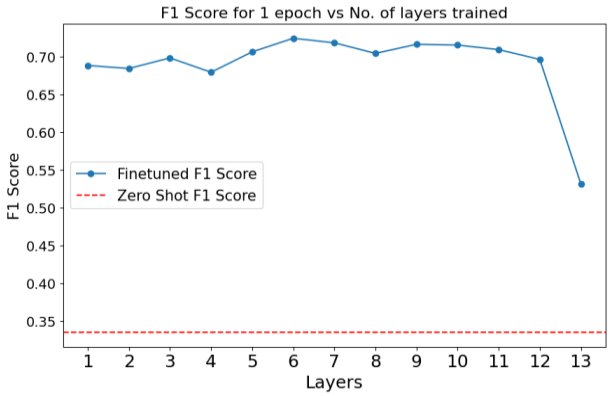

Figure 3 - F1 Scores for 1 epoch

Layers Trained	Accuracy	F1 Score
13	0.531	0.531
12	0.700	0.696
11	0.713	0.709
10	0.715	0.715
9	0.718	0.716
8	0.710	0.704
7	0.719	0.718
6	0.724	0.724
5	0.706	0.706
4	0.680	0.679
3	0.698	0.698
2	0.685	0.684
1	0.688	0.688
Zero Shot	0.495	0.335

Table 1: Classification Metrics for 1 Epoch

Layers	MSE Loss
13	2.586
12	2.347
11	2.323
10	2.352
9	2.297
8	2.347
7	2.440
6	2.236
5	2.275
4	2.501
3	2.560
2	2.490
1	2.272
Zero Shot	9.868

Table 2: Regression Loss for 1 Epoch

Increasing the number of epochs leads to performance improvements, but this comes at the cost of increased training time. Figure 5 shows that while the model's accuracy and regression loss improve with more epochs, the performance-to-training-time ratio declines after unfreezing a certain number of layers.

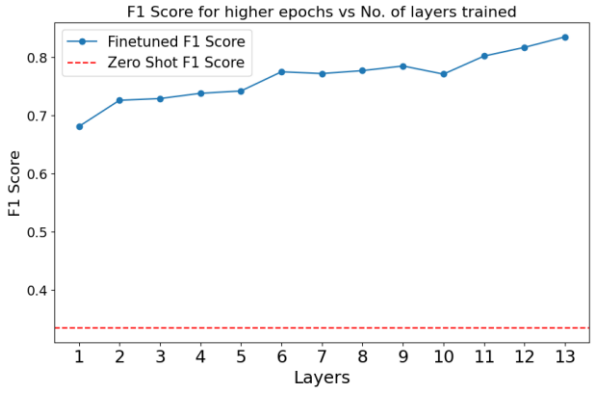

Figure 4 - F1 Scores for higher epochs

Specifically, after unfreezing layer 7, the time required for training becomes disproportionately high compared to the marginal gains in performance. This highlights the trade-off between achieving higher accuracy and the computational resources required, emphasizing the importance of balancing model performance with training efficiency. This trade-off is crucial for practical applications where computational resources and time are limited, necessitating an optimal balance between the number of unfrozen layers and training duration.

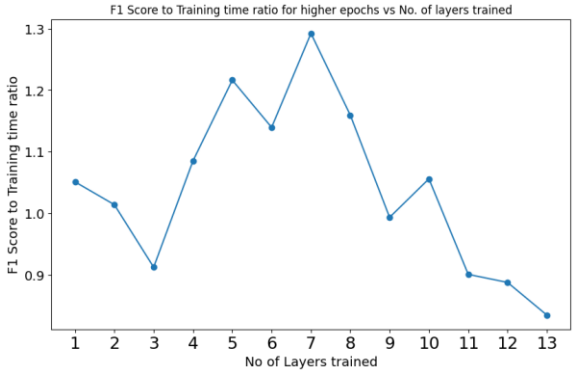

Figure 5 - F1 Score to Training time ratio

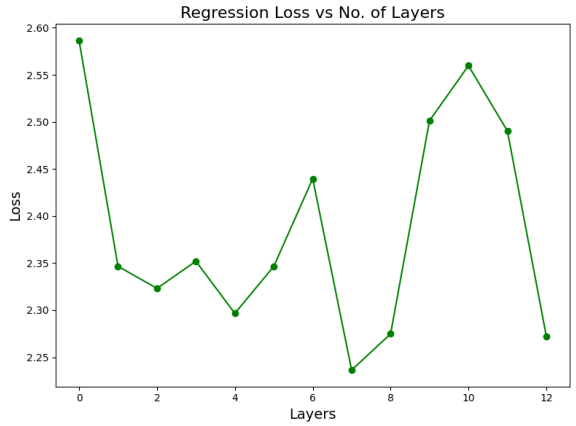

Figure 6 - Regression Loss for 1 Epoch

Conclusion.

Fine-tuning all the layers of a model is beneficial for optimal performance but is sub-optimal considering the extensive training time required. This finding might be useful in scenarios with limited computational resources, where sacrificing some performance for reduced computational burden is acceptable. By strategically freezing certain layers, one can balance performance and resource usage, making model fine-tuning feasible even with constrained resources.

References.

[1] arXiv:2201.11316 [cs.CV] https://doi.org/10.48550/arXiv.2201.11316
[2] Yamada, Moyuru, et al. "Transformer module networks for systematic generalization in visual question answering." arXiv preprint arXiv:2201.11316 (2022).
[3] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).
[4] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (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 (pp. 6904-6913).
[5] Surya Prakash, Miranda & Devananda, S. (2024). A Review of Recent Advances in Visual Question Answering: Capsule Networks and Vision Transformers in Focus. Indian Journal Of Science And Technology. 16. 4525-4546. 10.17485/IJST/v16i47.2643.
[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[7] Ashish, V. (2017). Attention is all you need. Advances in neural information processing systems, 30, I.
[8] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.