VQA with ViLT and Custom Loss

Haowen Xu
Columbia University
hx2346@columbia.edu

Abstract

Visual Question Answering (VQA) is a pivotal task at the intersection of computer vision and natural language processing, requiring precise interpretation of visual data based on textual queries. Our study introduces a resource-efficient approach to VQA by leveraging the pre-trained transformer model ViLT [3] for effective feature extraction and a specialized classifier that incorporates answer confidence into the training mechanism. By embedding answer confidence levels into a custom loss function, we aim to refine the classifier's ability to discern and weigh the veracity of annotated responses. This methodology demonstrates a statistically significant improvement on the VQA v2 test standard comparing to baseline model, showcasing the potential of integrating confidence metrics to enhance the predictive accuracy and reliability of VQA systems in a resource-conscious manner. Code can be found at DragonTail.

1. Introduction

Visual Question Answering (VQA) merges the realms of computer vision and natural language processing to challenge AI systems with the task of interpreting complex visual scenes and answering questions about them. Recent state-of-the-art approaches predominantly involve extensive fine-tuning of large-scale pre-trained transformer models on comprehensive VQA datasets. A prominent example is the Pathways Language and Image model (PaLI) [2], which utilizes immense computational resources to achieve breakthrough performance by training a 4-billion parameter Vision Transformer (ViT) alongside language transformers, demonstrating profound multimodal comprehension.

1.1. Motivation and Project Objectives

Despite their effectiveness, the computational demand and resource intensity of current methods limit their accessibility and scalability. Addressing this, our project proposes a more accessible approach that harnesses the power of pre-trained models for efficient feature extraction, circumventing the need for costly fine-tuning on expansive datasets. By integrating the pre-trained ViLT model [3], which combines the robustness of Vision Transformers and BERT-like architectures for text, we utilize these high-level features to fuel a specialized classifier that is both performance- and resource-optimized.

This project is driven by the goal to demonstrate that using pre-trained models in conjunction with custom classifiers can provide a viable solution to VQA challenges, achieving reasonable accuracy with significantly reduced resource expenditure.

Furthermore, we introduce an innovative element to our methodology by incorporating the confidence levels associated with annotated answers into our training process. This strategy utilizes a custom loss function that adjusts the influence of each training sample based on its annotated confidence, thus enhancing the model's sensitivity to data reliability and annotation quality. It aims to refine how the model handles ambiguities and inconsistencies in the training data, potentially leading to more nuanced and accurate responses.

2. Related Work

2.1. Background in Visual Question Answering

Visual Question Answering (VQA) has grown significantly as an area of research within computer vision and natural language processing. Agrawal et al. [1] introduced the task of VQA, which requires a system to generate a natural language answer in response to a visual input and a related question. This foundational work has paved the way for numerous advancements in the field, focusing on improving the interaction between visual and textual data.

2.2. Advanced Transformer Models in VQA

Recent developments have leaned heavily towards using transformer models to handle the complexities of VQA. The PaLI model described by Chen et al. [2] utilizes large pre-trained encoder-decoder language models alongside Vision Transformers (ViTs) to achieve state-of-the-art performance across multiple vision and language tasks. Similarly, Wang

et al. [5] introduced BEiT-3, which advances image and text representation learning through a unified multimodal model, demonstrating significant improvements on various benchmarks.

2.3. Innovations in Model Architecture

Distinct architectural advancements have also been prominent. Xu et al. [6] presented BridgeTower, which optimizes the integration of visual and textual representations through innovative bridge layers, allowing for effective cross-modal alignment. This architecture addresses the limitations of conventional two-tower designs by enabling deeper and more flexible interaction between modalities.

2.4. Simplification and Efficiency

Efforts to simplify the complex architectures typically used in VQA have also been noted. Wang et al. [4] developed GIT, a Generative Image-to-text Transformer that consolidates the architecture into a single image encoder and text decoder, significantly simplifying the model structure while enhancing performance on diverse vision-language tasks.

2.5. Modular and Co-Attention Approaches

Yu et al. [7] explored deep modular co-attention networks, focusing on the fine-grained interplay between visual and textual elements. Their approach differs by utilizing a cascaded series of attention mechanisms, improving the model's ability to focus on relevant parts of the image in relation to the question posed.

2.6. Relation to Our Work

Our project diverges from these methods that often rely on successive transformer layers to capture complex feature interactions, we directly exploit pre-trained feature extractors followed by a streamlined classification process. This not only simplifies the model architecture but also significantly reduces the computational overhead associated with training.

Our contribution lies in demonstrating the viability of using pre-trained models in a feature extraction role for VQA, supported by a classifier that effectively utilizes confidence information. This approach strikes a balance between performance and computational efficiency, offering a sustainable alternative to the resource-heavy models currently leading the field. This work contributes to the broader discourse on making advanced AI technologies more accessible and feasible within constraints of available resources.

3. Methodology

3.1. Hypothesis

The central hypothesis of our project is two-fold. Firstly, we hypothesize that utilizing pre-trained models for feature extraction, followed by a baseline classifier without incorporating confidence levels, can achieve reasonable accuracy in Visual Question Answering (VQA) tasks.

Secondly, we aim to evaluate whether integrating answer confidence levels into the classifier's loss function offers a significant advantage over the baseline model. By weighting the training samples based on their confidence scores, we anticipate that the model will prioritize learning from more reliable data, leading to improved accuracy and robustness.

3.2. Dataset

The experimental validation was carried out on the well-established VQA 2.0 dataset, a comprehensive benchmark in the field of Visual Question Answering. This dataset encompasses more than 200,000 images paired with over 1.1 million questions, representing a broad spectrum of visual contexts and querying paradigms. For details about the dataset, one can refer to VQAv2 and [1]. For each image-question pair, the dataset includes multiple annotated answers. These annotations come with associated confidence levels, which reflect the degree of consensus among the human annotators regarding each answer's correctness. Furthermore, for our experiments, we filtered the dataset to focus on the top 2000 answers, which collectively account for 91.22% of the training dataset. These answers form our label set, offering a representative subset for training and evaluating our model while ensuring coverage of the most frequent and salient responses within the dataset. This refined dataset allows for a focused and efficient training regimen, enabling us to concentrate computational resources on the most impactful elements of the VQA task.

3.3. Preprocessing

The preprocessing steps involved resizing images to the fixed size expected by the Vision Transformer (ViT) component of the Vilt model [3]. These preprocessed images and texts were then passed to the Vilt model, which has been pre-trained on extensive image-text paired data, to extract high-level features from both the visual and textual inputs. The extracted features were then fed into a Multi-Layer Perceptron (MLP) classifier.

3.4. Baseline Model Architecture

The MLPClassifier consists of two fully connected layers with 1024 units each, interspersed with ReLU activations. A dropout layer with a dropout rate of 0.5 follows the first activation to reduce overfitting. The output layer is

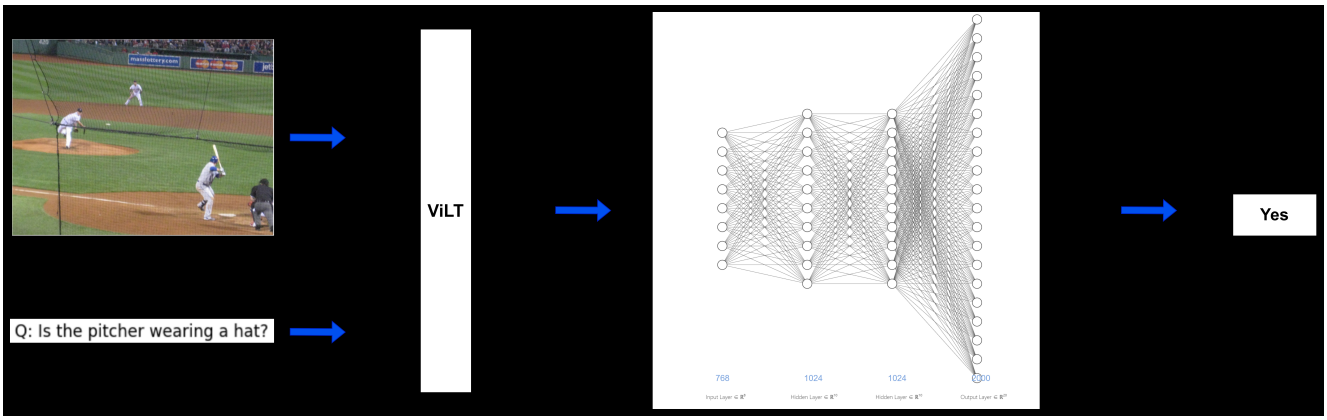

Figure 1. Overall flow of the VQA model, illustrating feature extraction using ViLT followed by the MLP classifier to determine the final answer.

also fully connected, translating the 1024 features from the second hidden layer into 2000 class scores corresponding to the possible VQA answers.

3.5. Custom Loss Function

The novelty in our methodology lies in the custom loss function used during training, which integrates the confidence levels associated with each answer. This custom loss function adjusts the impact of each training sample based on the confidence score, penalizing incorrect answers more severely if they are marked with high confidence.

Given a dataset D with N samples, where each sample i has a feature vector $\mathbf{x}_i$, a label y_i, and a set of answer confidences C_i with corresponding confidence scores $s_{i,j}$ for each answer j. The custom loss function L for the VQA model is defined as:

$$L(D, \theta) = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot \text{CEL}(f(\mathbf{x}_i; \theta), y_i) \tag{1}$$

where:

- θ represents the parameters of the MLP classifier.
- CEL denotes the standard cross-entropy loss function.
- $f(\cdot; \theta)$ is the MLP classifier function.
- w_i is the average weighted confidence for sample i, calculated as:

$$w_i = \frac{1}{|C_i|} \sum_{j \in C_i} \left(s_{i,j} \cdot \mathbb{I}[a_{i,j} = y_i] + \frac{s_{i,j}}{2} \cdot \mathbb{I}[a_{i,j} \neq y_i] \right) \tag{2}$$

Here, $a_{i,j}$ is the answer associated with confidence, y_i is the correct answer, and $s_{i,j}$, $\mathbb{I}[\cdot]$ is the indicator function, which is 1 when the condition is true and 0 otherwise. The loss for each sample is scaled by the average weighted confidence, penalizing incorrect answers with high confidence more severely.

3.6. Performance Metrics

To objectively assess the performance of our proposed VQA model, we adhere to the official VQA evaluation metrics [1], which account for the variability in human responses. This metric reflects the inherently subjective nature of the VQA task, as there can be multiple valid answers to a given question. Specifically, the accuracy of a generated answer is determined as follows:

$$\text{accuracy} = \min \left(\frac{\text{\# humans that provided that answer}}{3}, 1 \right) \tag{3}$$

In other words, an answer is considered completely accurate if at least three annotators have provided the same response.

4. Experimental Results

To assess the impact of integrating confidence levels into our VQA model's loss function, we conducted a series of experiments using the VQA 2.0 test-standard dataset. Our baseline model, trained for 50 epochs with a learning rate of 10^{-4} and using the Adam optimizer, served as the reference point for our comparative analysis.

4.1. Training Loss

The training process involved two models: one with a custom loss function that accounts for confidence levels and another that uses a standard loss function. As illustrated in Figure 2, the model with the confidence-aware loss function consistently exhibited lower loss values across training

epochs. This decrease in loss suggests improved learning efficiency and a better fitting model, potentially translating to more accurate predictions when dealing with complex, real-world VQA tasks.

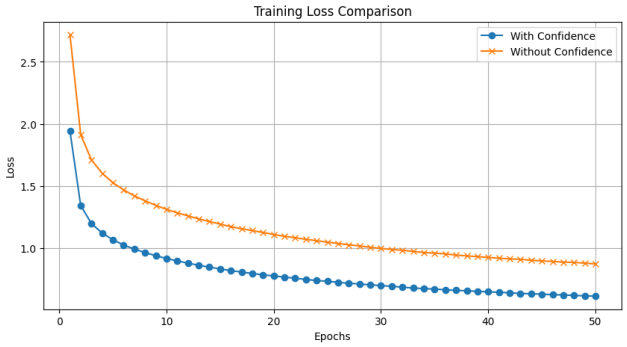

Figure 2. Comparison of training loss between the models.

4.2. Model Performance Comparison

To contextualize our model's performance, we juxtaposed it with state-of-the-art systems like PaLI-X [2] and the established deeper LSTM Q + norm I model [1]. Figure 3 provides a comprehensive view of this comparison. Although our model does not surpass the leading-edge accuracy of PaLI-X or the official baseline, it achieves reasonable performance, especially considering the significantly lower computational resources and simplicity of our methodology.

Performance Comparion on Test Standard

Method	All	Yes/No	Number	Other
PaLI-X	86.06	96.78	74.14	79.46
Deeper LSTM Q + norm I	57.75	80.5	36.77	43.08
Model with confidence	51.52	77.39	34.07	32.99
Model without confidence	50.94	76.98	34.09	32.13

Figure 3. Model performance comparison on the VQA test-standard dataset.

The results of our study suggest that incorporating confidence levels into the training of VQA models can indeed refine their predictive accuracy. Our model exhibits marginal yet meaningful improvements over the baseline, particularly in the 'Other' category of questions. While not outperforming state-of-the-art models like PaLI-X, our approach achieves respectable results, showcasing its potential as a resource-efficient alternative for the VQA task.

5. Conclusion

In this study, we proposed a Visual Question Answering (VQA) model that utilizes pre-trained feature extractors and a custom loss function integrating confidence levels to compensate for the limited resources available for training.

Figure 4. Sample model output with confidence.

Our results, though not surpassing state-of-the-art methods, demonstrate the model's potential to achieve reasonable accuracy within its constraints.

The open challenge of balancing resource efficiency with high performance in VQA remains unresolved. Due to the intrinsic complexity of VQA tasks and the limited ability to capture deeper feature interactions without extensive fine-tuning, our model's performance lags behind more resource-intensive approaches. The dependency on data quality, particularly the accuracy and consistency of annotator-provided confidence levels, also poses a significant challenge. The model's success hinges on the reliability of these confidence scores, making the integrity of the training data a critical factor.

Looking ahead, future work should explore methods to enhance feature representation without prohibitive resource expenditure. Investigating strategies such as transfer learning, few-shot learning, or meta-learning could yield improvements in model performance without substantial computational costs. Additionally, measures to ensure and possibly improve the quality of annotator confidence need to be developed, which may include robust data preprocessing or post-hoc correction algorithms that can refine or reweight unreliable annotations.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2015. 1, 2, 3, 4

[2] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2022. 1, 4

[3] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021. 1, 2

[4] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022. 2

[5] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022. 2

[6] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. Bridgetower: Building bridges between encoders in vision-language representation learning, 2022. 2

[7] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering, 2019. 2