Transformers in VQA: A Comparative Study of Multimodal and Single-Modality Fusion Approaches

Prince Immanuel Joseph **Shruti Bibra** **Jafar Aliyeh**

Abstract

Visual Question Answering (VQA) focuses on answering questions about images or videos and has grown a lot in recent years. However, current research on VQA lacks a thorough comparison between single-modality and multimodal transformers. VQA requires models to understand and integrate information from both text and images. In this study, we compare a single modality fusion approach—where language and vision features are extracted independently via Roberta (text) and Beit (image) and fused at a later stage—with a multimodal approach using BLIP, a state-of-the-art vision-language model that jointly encodes both modalities. We implement standardized preprocessing, training, and evaluation pipelines for fairness. The findings for this study will help optimize model design for improved accuracy, efficiency, and applicability in real-world scenarios. You can access the project on GitHub at the following link: Transformers-in-VQA: A Comparative Study of Multimodal and Single Modality Fusion Approaches

1 Introduction

VQA addresses the challenge of answering questions about images or videos, combining language understanding and visual perception. VQA is crucial in various applications, including assistive technologies, autonomous systems, and content-based image retrieval, where accurate integration of textual and visual data is essential. Despite its rapid development, previous studies often focus on either modular approaches or specific architectures without comparing single-modality and multimodal transformers comprehensively. This gap limits our understanding of how different architectures contribute to VQA performance and where their strengths and weaknesses lie. In this study, we aim to bridge this gap by exploring and comparing single-modality transformers, which process text and image data independently before fusion, and multimodal transformers, which jointly encode both modalities. We use the DAQUAR dataset, a benchmark for VQA tasks, as it provides structured and diverse examples for testing model capabilities. DAQUAR is particularly suitable for this study due to its balance of visual complexity and linguistic reasoning, making it an excellent dataset for analyzing the performance of different VQA approaches.

Single-modality transformers process language and vision features separately before combining them. For our experiments, we fine-tuned RoBERTa and BEiT as the language and vision components, respectively, because they represent state-of-the-art performance in their domains and are proven to work well for VQA tasks. RoBERTa excels in text understanding with its robust language representation, while BEiT effectively captures image features through its transformer-based architecture. Together, they provide a strong foundation for single-modality fusion approaches. In contrast, multimodal transformers jointly encode text and image inputs, allowing for deeper interactions between the two modalities. For our multimodal approach, we initially considered fine-tuning the LLAVA model on the DAQUAR dataset. However, due to limited computational resources, we opted for BLIP, a vision-language model with fewer trainable parameters. BLIP balances computational efficiency with performance, making it an ideal choice for resource-constrained environments.

To evaluate the models, we used accuracy, F1 score, and WordNet-based semantic similarity (WUP). Accuracy measures the proportion of correct answers, providing a straightforward assessment of model performance. The F1 score, which combines precision and recall, is particularly useful for imbalanced datasets or tasks requiring nuanced understanding. WUP evaluates semantic similarity, helping to gauge how closely the model's predictions align with ground-truth answers in meaning.

These metrics collectively provide a comprehensive evaluation of the models.

We trained all models for 20 epochs with a batch size of 32. While the single-modality approach using RoBERTa and BEiT achieved higher accuracy, the multimodal BLIP model outperformed in terms of F1 score and WUP. This indicates that multimodal transformers better capture the interplay between text and visual inputs, allowing for more semantically meaningful answers, even when exact matches are not achieved. The ability of BLIP to encode relationships between text and image features directly contributes to its superior performance in these metrics. In future, we aim to incorporate spatial reasoning, enabling models to better understand the positional relationships within visual data. Additionally, integrating techniques such as attention-based contextual embeddings and knowledge graphs could further enhance the reasoning capabilities of VQA models. By addressing these areas, we hope to push the boundaries of VQA and its applications in real-world scenarios.

2 Related Work

Many research studies have looked at fine-tuning transformer models for VQA. These studies focus on making different applications better, such as image captioning, object detection, scene understanding, and visual reasoning. The following related works offer helpful ideas for improving and developing our research. One important work is ViLBERT (1), which uses co-attention layers to link image regions and text embeddings, showing how transformers can handle both types of inputs together. Similarly, VisualBERT (2) trains a single transformer model on image-text data and achieves strong performance on VQA tasks. Another key model is UNITER (3), which focuses on aligning image features and words using pre-training tasks, setting benchmarks across multiple vision-language problems. The LXMERT model (4) uses separate transformers for images and text, combining them through cross-attention layers to understand complex visual and text relationships. BLIP (5) improves on this by pretraining a vision-language model with transformers, achieving state-of-the-art results on several VQA datasets. Finally, MDETR (6) uses transformers to link text phrases directly to parts of an image, performing exceptionally well on VQA tasks. These transformer-based models have advanced VQA significantly, making it possible to better connect and reason over visual and text data, and setting new standards in the field.

3 Method

In our experiments, each data instance comprises a textual question and a corresponding image, along with a label representing the answer. The dataset is pre-processed to ensure that all images are consistently sized and saved in a uniform format (e.g., PNG), and questions are tokenized into subwords for text-only models and text-encoder components in multimodal models. We compare two classes of approaches: single-modality fusion models and a multimodal model. In the single-modality fusion setting, each modality is processed independently, and a late-fusion step uses either (1) Roberta as a text-only model or (2) Beit as an image-only model to generate separate representations. These are then fused to produce a final prediction. In the multimodal setting, we employ BLIP as a state-of-the-art vision language model capable of jointly encoding textual and visual inputs from the start.

3.1 Dataset

The DAQUAR dataset provides a benchmark for VQA on real-world indoor scenes. It includes natural language questions about indoor RGB-D images captured from the NYU-Depth V2 dataset using Kinect sensors.
Source: Real-world RGB-D indoor images from the NYU-Depth V2 dataset.
Size: Full Dataset: 12,468 questions paired with images. Reduced Dataset: 1,443 questions.
Features: Real-world data: RGB-D indoor scenes, Questions on object identification, counting, spatial reasoning, Ground truth: Single-word or descriptive answers, Depth data for spatial reasoning.

3.2 Data Pre-processing

Single Modality Fusion: Questions are tokenized with the Roberta tokenizer, padded to a fixed length, and accompanied by attention masks and token type IDs. Images are loaded as RGB and processed with the Beit image processor, which resizes and normalizes them into standardized pixel tensors. A custom collator merges these text and image tensors into a single batch dictionary, adding the corresponding labels. This ensures both modalities are consistently pre-processed and aligned for the single-modality fusion model.
BLIP: All samples undergo standardized prepa-

ration before training. For text, we use the BLIP tokenizer to convert questions and answers into token indices, padding them to a fixed length and generating attention masks. For images, we load each file, convert it to RGB, resize it, and normalize it according to the BLIP image processor's specifications, producing a consistent tensor input. A custom dataset class retrieves the raw question, answer, and image path. On access, it applies the tokenizer to the question and answer, and the image processor to the image, returning a dictionary with input ids, attention mask, pixel values, and labels. A custom function then batches these elements into uniform tensors ready for model training. This pipeline ensures every data point is presented to the BLIP model in a consistent, model-ready format.

3.3 Fusion of RoBERTa and BieT

In the single-modality fusion setup, we first produce separate representations from two distinct model streams, one specializing in language input and the other in visual input. The written question is passed through a pretrained language model, such as Roberta, which transforms the sequence of words into a high-dimensional embedding that captures its semantic meaning. The associated image is fed into a pretrained vision model, such as Beit, which extracts a rich visual representation. This representation encodes the visual features, objects, and scene context contained in the image in a vector embedding.

Figure 1: Sample Prediction from RoBERTa and BieT

Once the language and vision models have each generated their embeddings, these two separate outputs are combined. In other words, we take the vector representing the question and the vector representing the image and place them side by side. After we combine the textual and visual representations, the resulting fused vector is passed through a small neural network module. This module, often a series of linear transformations coupled with nonlinear activations, learns how to integrate and interpret the combined information. By doing so, it can produce a prediction. This approach, known as late fusion, treats each modality independently up to the final stage. The model's output is displayed in Figure (1). Although stove burners and plastic boxes differ in both purpose and appearance, the similarity score of 0.4317 indicates that the model's internal representation places these two concepts at a moderately close distance. This positioning likely reflects shared contextual elements—such as appearing together in kitchen environments—rather than any similarity in their intended functions.

3.4 BLIP

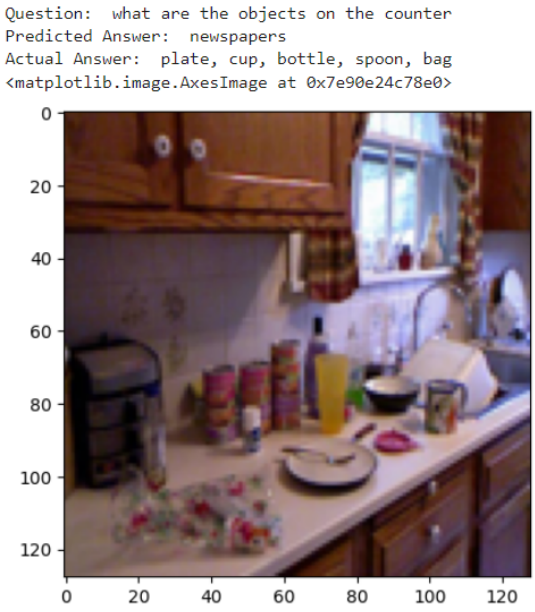

Figure 2: Sample Prediction from BLIP

For the multimodal baseline, we employ BLIP, a pre-trained vision-language model designed to handle both modalities jointly from the start. BLIP's architecture integrates textual and visual embeddings through transformer-based cross-attention layers. By using BLIP, we eliminate the need for separate encoders and a late fusion step. The question and image are directly fed to BLIP, which outputs a shared representation space optimized for multimodal understanding. BLIP is fine-tuned on the given VQA task, allowing the model to leverage its

pre-trained multimodal representations and adapt them to the target domain. The sample prediction can be seen in figure (2).

4 Results

4.1 Models comparison

We assessed our models using three metrics: Accuracy, F1 score, and WUPs(Wu-Palmer similarity) score. These metrics are effective for evaluating VQA transformers because they cover correctness, balance, and how closely answers match the true meaning.

Table 1: COMPARISON BETWEEN RoBERTa + BieT Vs BLIP

Model	Accuracy	F1 Score	WUPs Score
RoBERTa + BieT	25.33%	3.48%	30.49%
BLIP	21.66%	18.62%	43.92%

RoBERTa + BieT achieved higher accuracy as compared to BLIP. However, it has a very low F1 Score indicating difficulties in accurately understanding the images and questions.The WUPs Score of 30.49% suggests that the model was somewhat able to grasp the meaning of the answers. BLIP performed better with an F1 Score of 18.62%, indicating better semantic understanding but still challenges with precision. Its WUPs Score of 43.92% suggests that while BLIP's all the answers were not always accurate, they were closer in meaning to the correct answers.

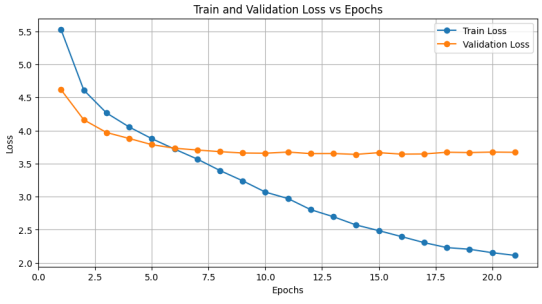

Figure 3: Traning and Validation Loss vs Epochs for RoBERTa + BieT

For RoBERTa + BieT, as the epochs increase, both the training and validation losses decrease significantly, indicating that the model is learning and improving its prediction accuracy over time. However, the validation loss plateaus earlier than the training loss, suggesting potential overfitting—where the model may have learned too much from the training data and struggles to generalize to new, unseen data.

For BLIP, the validation loss, which starts high, quickly decreases and stabilizes at a low value. This shows that the model has effectively learned from the training data and is not overfitting, as the validation loss remains low and does not diverge significantly. This graph reflects good model performance and efficient learning throughout training. While RoBERTa + BieT took longer to stabilize and had a higher validation loss, BLIP achieved faster convergence with better generalization to unseen data.

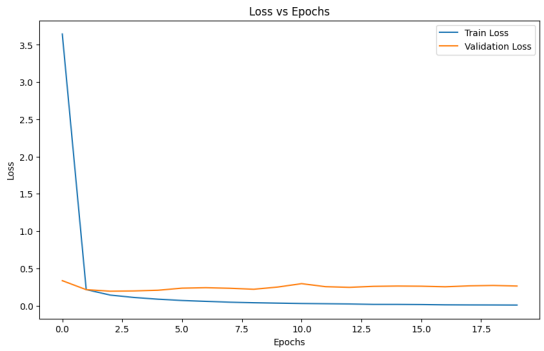

Figure 4: Traning and Validation Loss vs Epochs for BLIP

4.2 Comparison with other studies

Although our research focuses on the comparison between single modality fusion using RoBERTa and BERT with the multimodal BLIP. We also compared our models with other studies that used the DAQUAR dataset for VQA.

Table 2: COMPARISON BETWEEN DIFFERENT STUDIES

Model	WUPs Score
RoBERTa + BieT	0.3049
BLIP	0.4392
2-VIS+BLSTM (7)	0.8215
MULTI-WORLD (8)	0.5147

The (8) uses a Bayesian multi-world framework, which handles uncertainty in image analysis by considering multiple possible interpretations of the scene. It combines this with symbolic reasoning to answer questions and achieves a WUPS score of

0.5147. (7) uses two sets of image features—one from the start of the sentence and one from the end—each transformed using separate learned functions. It also uses two LSTMs one moving forward through the sentence and one backward. Both LSTMs contribute to the final prediction, which is made using a softmax layer at the end. This model performs even better for DAQUAR dataset, achieving a WUPS score of 0.8215.

5 Discussion and conclusion

Our results highlight a nuanced interplay between model complexity, training resources, and performance in VQA tasks. Contrary to initial expectations, the single-modality fusion approach achieved higher accuracy as compared to the multimodal BLIP mode under limited computational resources. This suggests that well-established unimodal feature extractors like Roberta and Beit can provide strong initial representations that. BLIP is designed to excel through deeper cross-modal interactions and joint representation learning, these advantages may become evident with longer training durations or more comprehensive hyperparameter tuning. The multimodal model demonstrated a superior F1 score. This indicates that BLIP's richer, joint representation learning may lead to more balanced predictions across classes, despite not achieving higher accuracy in the constrained training scenario. Hence, these models depends not only on raw accuracy but also on more nuanced metrics like F1 scores. Our study thus highlights the importance of evaluating multiple performance measures and considering both data efficiency and model complexity in VQA tasks.

5.1 Limitations

One limitation of this study is the reliance on a single dataset, which may not fully represent the variety of real-world image-question pairs, potentially limiting the generalization of the results. Fine-tuning the hyperparameters could lead to improved accuracy and generalization, but this study did not investigate this aspect thoroughly.

5.2 Future Scope

Building on our current work, we plan to enhance the model by incoperating spatial reasoning capabilities by using 3D data representations. Converting the existing dataset into a 3D LiDAR-based format would provide richer geometric cues, enabling the model to better interpret object positions, depth, and spatial relationships. By integrating LiDAR-derived point cloud features, we aim to extend the model's understanding from flat, 2D imagery to more realistic, three-dimensional scenes.

Additionally, we intend to apply Low-Rank Adaptation techniques to the BLIP model. LoRA can streamline the fine-tuning process by injecting a minimal set of learnable parameters into pre-trained layers, thus reducing computational overhead and improving model adaptability. This approach would allow BLIP to more efficiently incorporate spatial reasoning skills gained from the 3D data, ultimately enhancing its performance on vision-language tasks that demand richer spatial awareness. Also, increasing computational resources and training for more epochs will allow us to fully leverage the potential of these enhancements.

6 Statement of contributions

Each team member contributed to all tasks and played an active role in preparing the report. Prince focused on BLIP, Shruti worked on RoBERTa and BieT, while Jafar assisted with both the report and the code.

References

[1] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*.

[2] Li, L. H., Su, W., et al. (2019). VisualBERT: A Simple and Performant Baseline for Vision and Language. In *arXiv preprint arXiv:1908.03557*.

[3] Chen, Y. C., Li, L., et al. (2020). UNITER: Universal Image-Text Representation Learning. In *European Conference on Computer Vision (ECCV)*.

[4] Tan, H., & Bansal, M. (2019). LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

[5] Li, J., Li, D., et al. (2022). BLIP: Bootstrapped Learning from Image-Text Pairs for Vision-Language Pretraining. In *Proceedings of the 2022 Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Kamath, A., et al. (2021). MDETR: Modulated Detection for End-to-End Multi-Modal Understanding. In *Proceedings of the 2021 International Conference on Computer Vision (ICCV)*.

[7] Ren, M., Kiros, R., & Zemel, R. S. (2015). Exploring Models and Data for Image Question Answering. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

[8] Malinowski, M., & Fritz, M. (2014). A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Salesforce. (2022). BLIP: Bootstrapping Language-Image Pretraining. GitHub. Retrieved from https://github.com/salesforce/BLIP.

[10] Victor, B. (2022). BLIP Medical Visual Question Answering. In *Kaggle*. Retrieved from https://www.kaggle.com/code/basu369victor/blip-medical-visual-question-answering#Inference.

[11] Chiio, D. (2022). BLIP VQA Fine-Tuning. GitHub. Retrieved from https://github.com/dino-chiio/blip-vqa-finetune.

[12] Sahu, T. (2022). VQA With Multimodal Transformers. GitHub. Retrieved from https://github.com/tezansahu/VQA-With-Multimodal-Transformers.

[13] Tang, H. (2022). Vision Question Answering System Based on RoBERTa and ViT Model. In *2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, Xi'an, China, 258-261. doi: 10.1109/ICICML57342.2022.10009711.

[14] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, Haruo Takemura, "A comparative study of language transformers for video question answering,"2024.