Visual Question Answering Using Deep Learning: A Multimodal Approach

Aysha Siddika Marua
ID:2104010202229
Department of Computer Science
Premier University
Chittagong, Bangladesh
ayshasiddikamarua@gmail.com

Hasna Hena
ID:2104010202239
Department of Computer Science
Premier University
Chittagong, Bangladesh
hasnahena7665@gmail.com

Antika Dhar
ID:2104010202256
Department of Computer Science
Premier University
Chittagong, Bangladesh
antikadhar7@gmail.com

***Abstract*—Visual Question Answering (VQA) bridges computer vision and natural language processing to interpret images through question-answering. This study evaluates three vision architectures—ResNet-50, EfficientNet-B3, and Vision Transformer (ViT-B/16)—integrated with BERT for multimodal feature fusion. Experiments show ViT-B/16 achieves the highest weighted F1-score (0.169), though class imbalance and small-object detection remain critical challenges. Error analysis reveals frequent mispredictions for dominant classes (“table,” “pillow”), highlighting dataset biases. Hyperparameter tuning identifies optimal performance at a learning rate of 10^{-4} and batch size of 16. This work advocates for advanced fusion techniques, spatial reasoning modules, and balanced datasets to enhance VQA robustness. Future research will explore hierarchical attention mechanisms and synthetic data augmentation to address observed limitations.**

***Index Terms*—Visual Question Answering (VQA), Class Imbalance, Feature Fusion, Fine-Tuning, Deep Learning, ResNet, EfficientNet, Vision Transformer (ViT), Cross-Entropy Loss, Multimodal Learning.**

I. Introduction

Visual Question Answering (VQA) represents a complex yet highly impactful task at the intersection of computer vision and natural language processing (NLP). It involves the generation of accurate textual answers to natural language questions based on the content of a given image. Recent advancements in deep learning have introduced powerful architectures for both image and text processing. Convolutional Neural Networks (CNNs) such as ResNet-50 and EfficientNet-B3 have demonstrated impressive performance in extracting high-level semantic features from images. Additionally, the Vision Transformer (ViT-B/16) introduces a novel self-attention-based mechanism that captures long-range dependencies across visual tokens, proving effective in many visual recognition tasks. On the language side, pre-trained models like BERT have become foundational for understanding contextual word embeddings, making them well-suited for question encoding in VQA systems.

In this study, we investigate the performance of a multimodal VQA framework that integrates visual representations from ResNet-50, EfficientNet-B3, and ViT-B/16 with textual features extracted via BERT. The fusion of these modalities is performed using a joint representation strategy, followed by a classifier trained to predict answers from a space of 582 unique classes. We evaluate model performance on a curated VQA dataset and analyze the impacts of architectural choices, hyperparameter settings, and dataset biases.

Experiments are conducted on a Kaggle-based dataset containing 12,468 image-question-answer triplets across 582 classes. Key contributions include:

- A comparative evaluation of CNN and transformer-based architectures,

- Systematic hyperparameter tuning, and

- Insights from error analysis.

II. Literature Review

Visual Question Answering (VQA) has gained significant attention with advancements in deep learning techniques. Early VQA models relied on convolutional neural networks (CNNs) for visual feature extraction and recurrent neural networks (RNNs) for textual input processing. However, these architectures struggled with long-range dependencies and multimodal alignment, limiting their effectiveness.

The introduction of attention mechanisms significantly improved multimodal representation learning. For instance, the Bottom-Up and Top-Down Attention model enabled object-level attention for VQA tasks. Similarly, transformer-based models such as LXMERT demonstrated strong performance by learning cross-modal interactions through pretraining on large-scale datasets.

For image encoding, ResNet has been widely adopted due to its robust performance across various visual tasks, attributed to its residual connections. EfficientNet further enhanced accuracy-efficiency trade-offs using compound scaling. Recently, Vision Transformers (ViTs) have shown strong performance by modeling global image context through self-attention mechanisms.

In parallel, Bidirectional Encoder Representations from Transformers (BERT) has become a standard for language modeling due to its ability to capture bidirectional context. Combining vision and language models has led to significant improvements in multimodal tasks, motivating our comparative study of CNN- and transformer-based architectures for VQA.

This review highlights the evolution of VQA models, underscoring the need for advanced fusion techniques and robust architectures to address persistent challenges such as class imbalance and dataset biases.

III. Related works

A. Overview of visual question answering

Visual Question Answering (VQA) combines visual perception with natural language understanding to generate accurate answers based on image-question pairs. Early models employed CNNs for image feature extraction and RNNs for question encoding, followed by basic fusion techniques such as concatenation. Although effective for simple tasks, these models lacked the capacity for deeper reasoning and multimodal alignment.

B. Vision and Language Backbones

For visual feature extraction, CNNs like ResNet-50 and EfficientNet-B3 have been widely adopted due to their strong performance on classification tasks. Recently, Vision Transformers (ViT) have gained attention for their ability to capture global context using self-attention, showing competitive results in VQA and related vision tasks. On the language side, BERT has emerged as a dominant model for question encoding, providing deep contextual embeddings that enhance understanding of complex queries.

C. Fusion Techniques and Architectural Comparisons

Multimodal fusion remains a critical component in VQA systems. Beyond simple concatenation, research has explored bilinear pooling, gated fusion, and co-attention to effectively combine visual and textual information. However, most studies evaluate these techniques within a single architectural pipeline.

In contrast, our work conducts a comparative analysis of three visual encoders—ResNet-50, EfficientNet-B3, and ViT-B/16—each integrated with BERT for text understanding. This unified framework allows for direct performance comparison and highlights the strengths and limitations of different vision backbones under the same VQA setup.

IV. Methodology

This study implements a multimodal deep learning framework for Visual Question Answering (VQA). The process includes dataset preparation, feature extraction, model architecture design, and supervised training.

A. VQA Datset Description

The Visual Question Answering (VQA) dataset used in this project consists of 12,468 entries, each containing an image, a natural language question, and a corresponding answer. The images cover a wide range of real-world scenes, and the questions vary in type, including object recognition, color identification, counting, and spatial reasoning. Answers are selected from a predefined set of 582 unique labels. Each image is referenced by a unique ID and stored in a designated folder, while the questions and answers are stored in a CSV file.

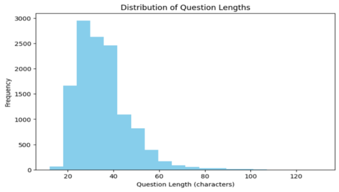


Fig. 1. Histogram of question lengths,showing distribution.

Fig. 2. Sample images from the dataset with id shown

B. Text and Image Processing

To prepare the dataset for training, both the image and text modalities underwent extensive preprocessing. The text data was first cleaned by removing null or empty records, followed by the elimination of irrelevant tokens, punctuation, and special characters. To ensure uniformity and reduce computational overhead, questions were truncated to a maximum sequence length of 30 tokens. A word cloud was generated before and after the cleaning process to visually validate the effectiveness of the text preprocessing steps. On the image side, all images were resized to 224×224 pixels, converted to tensors, and normalized using standard ImageNet statistics. These steps ensured consistency across inputs and compatibility with pretrained vision models used for feature extraction.

C. Exploratory Data Analysis

The Visual Question Answering (VQA) dataset contains 12,468 entries with 582 unique answer classes.

D. Data Splitting

The dataset was randomly split into training, validation, and testing subsets using an 80-10-10 ratio.A total of 8,976 samples were allocated for training, 998 samples for validation, and 2,494 samples for testing.

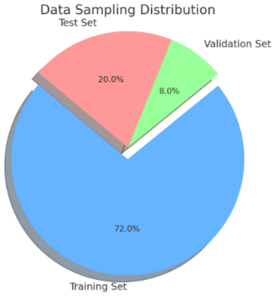
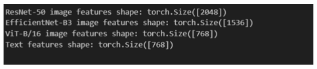


Fig. 3. Data Sampling



```
ResNet-50 image features shape: torch.Size([2048])
EfficientNet-B3 image features shape: torch.Size([1536])
ViT-B/16 image features shape: torch.Size([768])
Text features shape: torch.Size([768])
```

Fig. 4. Caption

V. FEATURE EXTRACTION

A. *Texual Features*

Texual Features were extracted using BERT.The mean pooles output from the final hidden layer(768-dimentional) was used to represent each questions.

1) Vision Features: - ResNet-50:2048-dimentional output. -EfficientNet-B3:1536-dimentional output. -ViT-B/16: 768-dimentional output.

VI. Model Architecture

The VQA model projects BERT-based text features into a lower-dimensional space using a linear layer. These are concatenated with visual embeddings and passed through a fully connected layer with ReLU activation and dropout ($p = 0.5$), followed by a classification layer mapping to 582 answer classes.

A. *Fusion Model Design*

1.Text Projection Layer: A linear layer that projects 768-dimensional BERT embeddings to a hidden dimension (512).

2.Feature Fusion Layer: The projected text features and vision features are concatenated and passed through a fully connected layer followed by ReLU activation and dropout.

3.Classification Layer: A final linear layer outputs logits corresponding to 582 answer classes.

B. *Hyperparameter Tuning*

The model is optimized using the Adam optimizer and trained for 10 epochs. The loss function employed is CrossEntropyLoss, and a hidden layer of size 512 is used for feature fusion.

VII. Results

We evaluated three VQA models—ResNet-50 + BERT, EfficientNet-B3 + BERT, and ViT-B/16 + BERT—on a test set of 2,494 samples using accuracy and weighted F1-score as evaluation metrics.

A. *Model Performance*

ViT-B/16 + BERT achieved the highest performance, with 32.27% accuracy and a weighted F1-score of 0.169. EfficientNet-B3 followed with 28.57% accuracy and an F1-score of 0.15, while ResNet-50 scored 20.88% accuracy and 0.13 F1-score. Table I summarizes the results.

TABLE I
Performance Comparison of VQA Models

Model	Test Accuracy (%)	Weighted F1-Score
ResNet-50 + BERT	20.88	0.13
EfficientNet-B3 + BERT	28.57	0.15
ViT-B/16 + BERT	32.27	0.169

B. *ViT-B/16+BERT*

The ViT-B/16 + BERT model achieved a training accuracy improvement from 5.46. Validation accuracy increased from 10.62. On the test set, the model achieved an accuracy of approximately 23. These results suggest that Vision Transformers can effectively capture complex visual features for the VQA task .

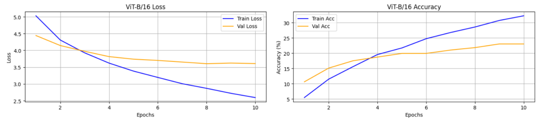


Fig. 5. Training and Validation Loss/Accuracy Curves for ViT-B/16+BERT

C. *ResNet-50+BERT*

The ResNet-50 + BERT model showed a steady improvement, with training accuracy rising from 4.71. Validation accuracy improved from 9.72. On the test set, the model achieved an overall accuracy of approximately 20. These results reflect the dataset's complexity due to high class imbalance and answer diversity .

D. *EfficientNet-B3+BERT*

The EfficientNet-B3 + BERT model demonstrated superior learning behavior compared to ResNet-50 . Training accuracy increased steadily from 5.27. Validation accuracy rose from 10.32. The model achieved a test set accuracy of approximately 24. EfficientNet-B3's compound scaling strategy enabled better representation of visual patterns . This result highlights its suitability for complex multi-modal tasks like Visual Question Answering .

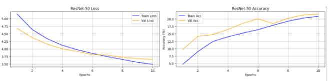


Fig. 6. Training and Validation Loss/Accuracy Curves for ResNet-50+BERT

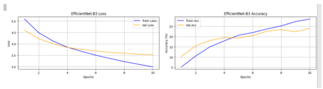


Fig. 7. raining and Validation Loss/Accuracy Curves for EfficientNet-B3+BERT

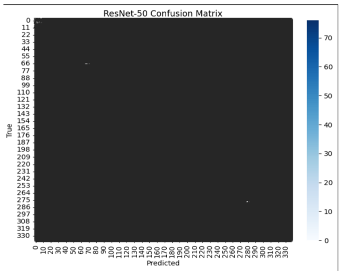


Fig. 8. Confusion matrix for ResNet-50 on test data

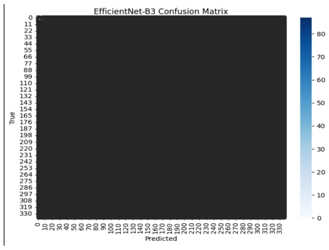


Fig. 9. Confusion matrix for EfficientNet-B3 on test data

VIII. Model Evaluation

Confusion matrices show that ResNet-50 and EfficientNet-B3 correctly predict frequent classes but struggle with rare categories, while ViT-B/16 achieves better overall generalization. ViT-B/16 records the highest weighted F1-score (0.169) compared to the others. All models face challenges with class imbalance and small-object detection.

Model evaluation is crucial to understand how well a model generalizes to unseen data. It highlights strengths, weaknesses, and biases in predictions, such as favoring dominant classes or missing rare ones. Using metrics like confusion matrices, F1-scores, and accuracy ensures a comprehensive assessment beyond just training performance, guiding improvements in model design, data balancing, and future work.

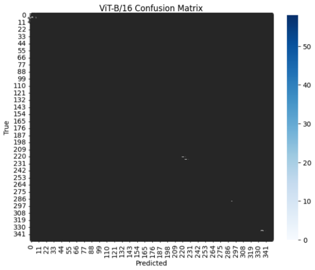
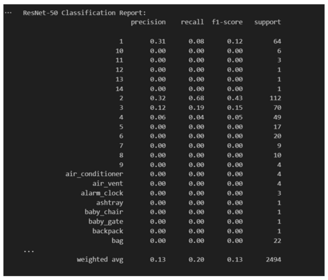


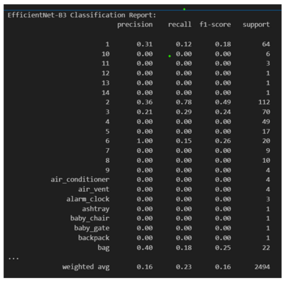
Fig. 10. Confusion matrix for ViT-B/16 on test data



```
...   ResNet-50 Classification Report:
                            precision    recall  f1-score   support

                      1        0.31      0.08      0.12        64
                     10        0.00      0.00      0.00         6
                     11        0.00      0.00      0.00         3
                     12        0.00      0.00      0.00         1
                     13        0.00      0.00      0.00         1
                     14        0.00      0.00      0.00         1
                      2        0.32      0.68      0.43       112
                      3        0.12      0.19      0.15        70
                      4        0.06      0.04      0.05        49
                      5        0.00      0.00      0.00        17
                      6        0.00      0.00      0.00        20
                      7        0.00      0.00      0.00         9
                      8        0.00      0.00      0.00        10
                      9        0.00      0.00      0.00         4
        air_conditioner        0.00      0.00      0.00         4
               air_vent        0.00      0.00      0.00         4
            alarm_clock        0.00      0.00      0.00         3
                ashtray        0.00      0.00      0.00         1
             baby_chair        0.00      0.00      0.00         1
              baby_gate        0.00      0.00      0.00         1
               backpack        0.00      0.00      0.00         1
                    bag        0.00      0.00      0.00        22
...
           weighted avg        0.13      0.20      0.13      2494
```

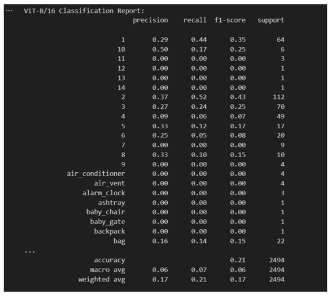
Fig. 11. report summary for ResNet-50



```
EfficientNet-B3 Classification Report:
                        precision    recall  f1-score   support

                      1      0.31      0.12      0.18        64
                     10      0.00      0.00      0.00         6
                     11      0.00      0.00      0.00         3
                     12      0.00      0.00      0.00         1
                     13      0.00      0.00      0.00         1
                     14      0.00      0.00      0.00         1
                      2      0.36      0.78      0.49       112
                      3      0.21      0.29      0.24        70
                      4      0.00      0.00      0.00        49
                      5      0.00      0.00      0.00        17
                      6      1.00      0.15      0.26        20
                      7      0.00      0.00      0.00         9
                      8      0.00      0.00      0.00        10
                      9      0.00      0.00      0.00         4
        air_conditioner      0.00      0.00      0.00         4
               air_vent      0.00      0.00      0.00         4
            alarm_clock      0.00      0.00      0.00         3
                ashtray      0.00      0.00      0.00         1
             baby_chair      0.00      0.00      0.00         1
              baby_gate      0.00      0.00      0.00         1
               backpack      0.00      0.00      0.00         1
                    bag      0.40      0.18      0.25        22
...
           weighted avg      0.16      0.23      0.16      2494
```

Fig. 12. report summary for EfficientNet-B3



```
...   ViT-B/16 Classification Report:
                            precision    recall  f1-score   support

                      1        0.29      0.44      0.35        64
                     10        0.50      0.17      0.25         6
                     11        0.00      0.00      0.00         3
                     12        0.00      0.00      0.00         1
                     13        0.00      0.00      0.00         1
                     14        0.00      0.00      0.00         1
                      2        0.37      0.52      0.43       112
                      3        0.27      0.24      0.25        70
                      4        0.09      0.06      0.07        49
                      5        0.33      0.12      0.17        17
                      6        0.25      0.05      0.08        20
                      7        0.00      0.00      0.00         9
                      8        0.33      0.10      0.15        10
                      9        0.00      0.00      0.00         4
        air_conditioner        0.00      0.00      0.00         4
               air_vent        0.00      0.00      0.00         4
            alarm_clock        0.00      0.00      0.00         3
                ashtray        0.00      0.00      0.00         1
             baby_chair        0.00      0.00      0.00         1
              baby_gate        0.00      0.00      0.00         1
               backpack        0.00      0.00      0.00         1
                    bag        0.16      0.14      0.15        22
...
               accuracy                            0.21      2494
              macro avg        0.06      0.07      0.06      2494
           weighted avg        0.17      0.21      0.17      2494
```

Fig. 13. report summary for ViT-B/16

IX. Hyper parameter

Hyperparameter tuning was performed to optimize model performance. Different learning rates and batch sizes were tested and a batch size of 16 provided the best results across all models. Lower learning rates led to slower convergence, while higher rates caused instability. Batch size variations affected model generalization, with larger batches reducing overfitting but requiring more training time. Optimal tuning helped achieve better validation accuracy and improved F1-scores.

X. Error Analysis

A. Model Performance:

All models—ResNet-50, EfficientNet-B3, and ViT-B/16—demonstrated low performance, achieving only 20 to 24 percentage test accuracy. The classification reports revealed that a majority of classes had near-zero precision, recall, and F1-scores, suggesting significant issues in both prediction accuracy and generalization capability.

B. Class Imbalance:

The dataset exhibits severe class imbalance, with a few classes having relatively high support (e.g., ”sink”, ”books”) and many classes having less than five samples. Classes with minimal samples recorded zero precision, recall, and F1-scores. The models were biased towards frequent classes and unable to predict rare classes effectively.

C. Overfitting and Underfitting:

Training accuracy steadily increased, while validation accuracy improved marginally and plateaued early, indicating potential overfitting. For example, EfficientNet-B3 achieved 28.57 percentage training accuracy and 24.05 percentage validation accuracy at epoch 10, showing a gap typical of moderate overfitting combined with under-capacity to generalize over the validation data.

D. Feature Representation:

Image features were extracted using pre-trained ResNet-50, EfficientNet-B3, and ViT-B/16 models without fine-tuning, leading to a domain mismatch since these models were trained on ImageNet. Text features were extracted from a pre-trained BERT model without any VQA-specific fine-tuning. The feature fusion utilized simple concatenation followed by shallow multi-layer perceptrons, which is insufficient for capturing complex multimodal relationships.

E. Loss Function and Optimization:

Although CrossEntropyLoss combined with the Adam optimizer is standard, the complexity induced by 582 output classes demands more sophisticated handling, such as class-balanced loss, focal loss, or label smoothing. The current setup did not address the extreme class imbalance.

F. Dataset and Preprocessing Issues:

During preprocessing, only the first available answer per sample was selected from possibly multiple valid answers. This strategy likely introduced label noise, thereby affecting model training and evaluation.

G. Specific Observations:

Isolated perfect F1-scores (e.g., ”projector” with F1-score = 1.0) were observed only for classes with a single sample, which do not reflect the model’s generalization ability. Frequent classes such as ”chair” and ”bed” also showed low recall values, indicating difficulty in consistent classification even among majority classes.

XI. Discussion

The overall poor performance highlights the need for more sophisticated methods in VQA tasks. Handling class imbalance and improving feature representations are critical to boosting performance. Furthermore, effective fusion strategies that can model complex interactions between image and text modalities are necessary. Fine-tuning pre-trained models on the specific VQA domain can help close the domain gap, and incorporating adaptive loss functions can better manage the challenging multi-class setting. Future work should explore transformer-based multimodal architectures and apply targeted data augmentation to better address these challenges.

XII. Conlusion

Conclusion:

In conclusion, the study reveals that while pretrained models provide a strong foundation, significant challenges such as class imbalance, inadequate multimodal integration, and lack of fine-tuning severely limit performance in VQA tasks. Addressing these shortcomings through balanced data sampling, deeper model fusion, fine-tuning strategies, and improved loss designs will be essential to advance the system’s ability to answer visual questions accurately. Future enhancements must target both model architecture and dataset preparation to achieve meaningful improvements.

XIII. Reference

[1] K. He, X. Zhang, S. Ren, and J. Sun, ”Deep Residual Learning for Image Recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[2] M. Tan and Q. Le, ”EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in Proceedings of the International Conference on Machine Learning (ICML), 2019.

[3] A. Dosovitskiy et al., ”An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019.

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

[6] A. Vaswani et al., ”Attention Is All You Need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[7] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, ”What Does BERT Look at? An Analysis of BERT’s Attention,” in Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2019.

[8] P. Anderson et al., ”Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[9] A. Zellers, Y. Bisk, R. Farhadi, and Y. Choi, ”From Recognition to Cognition: Visual Commonsense Reasoning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[10] D. Hudson and C. Manning, ”GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.