Visual Question Answering Using Pretrained Models: A Comparative Study of ResNet+BERT and ResNet+BERT+DETR Architectures with SBERT Similarity Analysis

Muskan Motwani
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
Email: muskanmotwani_24rco01@dtu.ac.in

***Abstract*—This paper presents a comparative study of two Visual Question Answering (VQA) architectures that leverage pretrained models for both visual and textual feature extraction. The first architecture integrates a ResNet-50 backbone for visual representation and a BERT model for textual understanding, while the second architecture extends this combination by incorporating a Detection Transformer (DETR) to capture object-level features. Both architectures are thoroughly evaluated in terms of their accuracy, loss behavior, and semantic quality of predictions, the latter being measured via Sentence-BERT (SBERT)-based similarity metrics. Experimental results demonstrate that incorporating DETR features alongside ResNet and BERT improves not only the prediction accuracy but also the semantic alignment between predicted and ground truth answers. These findings suggest that fine-grained, object-level visual representations can enhance the performance of VQA systems, leading to more contextually relevant responses.**

***Index Terms*—Visual Question Answering, ResNet, DETR, BERT, SBERT, Deep Learning, Pretrained Models, Semantic Similarity**

I. Introduction

Visual Question Answering (VQA) is an interdisciplinary challenge that combines computer vision and natural language processing to enable machines to understand and answer questions about images. Unlike tasks that focus on a single modality, VQA requires effective integration of visual cues and textual semantics. For example, to answer a question such as "What is the color of the car in the image?", a VQA model must first identify the relevant object (the car) and then determine its attributes (its color) from the image. This inherently multi-modal nature makes VQA a complex and rich research area with applications in image search, assistive technologies, and conversational AI systems.

Recent advances in deep learning have strongly benefited from pretrained models. Convolutional Neural Networks (CNNs) like ResNet [1] offer robust visual feature representations, while transformer-based models such as BERT [2] provide powerful language embeddings. More recently, the Detection Transformer (DETR) [3] has shown state-of-the-art performance in object detection tasks, transforming the landscape of how objects are localized and identified in images.

In this work, a comparative study of two architectures for VQA is presented:

1) **ResNet+BERT**: A baseline model that leverages ResNet-50 for visual features and BERT for textual encoding. This model captures a global representation of the image without focusing on individual objects.
2) **ResNet+BERT+DETR**: An enhanced model that introduces DETR-derived object-level features alongside the global ResNet features and BERT's question embeddings. This integration aims to provide a richer, more fine-grained understanding of the scene.

By incorporating Sentence-BERT (SBERT) [4] for semantic similarity analysis, this work goes beyond conventional accuracy metrics to evaluate how closely the predicted answers align with the semantics of the ground-truth answers.

The remainder of the paper is organized as follows: Section II reviews related work in VQA and associated fields. Section III details the methodology for both architectures and explains the SBERT similarity metric. Section IV presents the dataset, experimental setup, and results, including accuracy, loss curves, and similarity scores. Section V discusses the implications, limitations, and potential future work. Finally, Section VI concludes the paper.

II. Related Work

VQA has its roots in early image-captioning and question-answering paradigms. Over time, the field has evolved with increasingly complex models, attention mechanisms, and large-scale datasets such as VQA v2 [5] and CLEVR [6].

A. Early CNN-LSTM Based Models

Earlier VQA models often combined CNN-based image encoders with recurrent neural networks (RNNs) like LSTM to process textual input. They integrated these features through simple concatenation or element-wise operations, achieving modest success. However, such approaches were limited in their ability to focus selectively on image regions relevant to the question at hand.

B. *Attention Mechanisms*

To address this limitation, attention-based models were introduced. Yang et al. [7] proposed the Stacked Attention Network (SAN), which applied multiple layers of attention to progressively refine visual features according to the question context. Similarly, Anderson et al. [8] utilized bottom-up and top-down attention, identifying salient objects in the image to improve alignment with textual queries.

C. *Transformer-Based Approaches*

Transformer architectures have revolutionized both language modeling and, more recently, vision tasks. Models like VisualBERT [10], ViLT [11], and UNITER [12] use transformer blocks to learn joint representations of images and text. DETR [3] introduced a transformer-based approach for object detection, demonstrating that transformers can successfully learn spatial and semantic relationships in images without reliance on handcrafted priors.

D. *IC VQA and Caption-Based Methods*

Kun Zhou [9] introduced an Image Captioning-based VQA framework (IC VQA), leveraging caption generation to guide the VQA model. By first obtaining detailed captions of image content and then using these captions to refine the answer prediction, such methods suggest that rich textual descriptions of images can enhance the reasoning capabilities of VQA models.

E. *Semantic Similarity Measures*

Traditional VQA evaluation uses accuracy based on exact matches of answers. However, semantic similarity metrics, such as those provided by SBERT [4], allow for more nuanced evaluations. Two answers might differ lexically yet be semantically identical (e.g., “car” vs. “automobile”), and similarity metrics can capture these subtle differences.

III. Methodology

Our study focuses on two architectures that share a common BERT encoder for language understanding but differ in how they extract and represent visual information.

A. *ResNet+BERT Architecture*

Visual Feature Extraction: As shown in Fig. 1 a ResNet-50 [1] pretrained on ImageNet [13] is used to extract a global, 2048-dimensional feature vector representing the entire image. The final fully connected layer is removed, and a rich, semantically meaningful vector is obtained from the penultimate layer

Textual Feature Extraction: For question processing,a use BERT-base-uncased [2], a widely used transformer-based language model is employed. The [CLS] token representation (768-dimensional) captures the contextual meaning of the entire question.

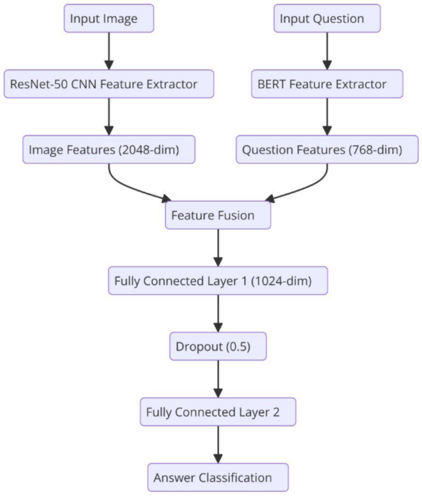

Fig. 1. Architecture: ResNet+Bert

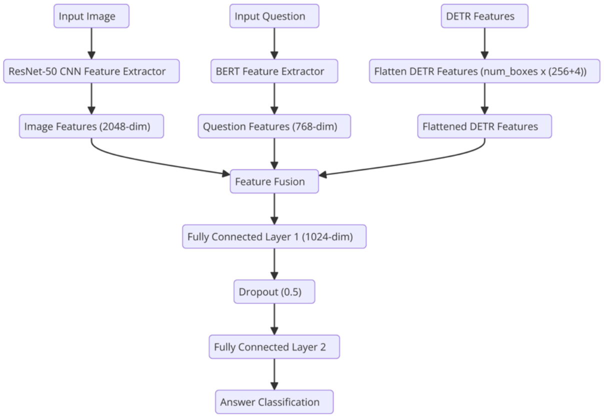
Input Image
Input Question
DETR Features
ResNet-50 CNN Feature Extractor
BERT Feature Extractor
Flatten DETR Features (num_boxes x (256+4))
Image Features (2048-dim)
Question Features (768-dim)
Flattened DETR Features
Feature Fusion
Fully Connected Layer 1 (1024-dim)
Dropout (0.5)
Fully Connected Layer 2
Answer Classification

Fig. 2. ARCHITECTURE: ResNet+BERT+DETR

Fusion and Classification: The image (2048-D) and question (768-D) features are concatenated, resulting in a combined vector of dimensionality 2816. This vector is then passed through fully connected (FC) layers, which project it into the answer space. The final FC layer, followed by a softmax, predicts the probability distribution over possible answers.

B. *ResNet+BERT+DETR Architecture*

As shown in Fig. 2 To incorporate object-level reasoning, DETR features [3] are introduced. DETR provides a set of object queries, each associated with a 256-D feature vector and 4 box coordinates, resulting in a 260-D feature per object. With 100 object queries, a 100x260 matrix is obtained. Flattening this matrix produces a 26,000-dimensional vector representing object-level embeddings.

Three components are then fused: F_{resnet} (2048-D), F_{bert} (768-D), and F_{detr} (26,000-D). This rich representation is passed through two FC layers to reduce dimensionality and produce the final probability distribution over answers. Dropout layers are used to prevent overfitting.

C. SBERT Similarity Analysis

While accuracy remains a primary metric, it fails to capture semantic equivalences between answers. Sentence-BERT (SBERT) is used for semantic evaluation [4], which produces sentence-level embeddings well-suited for cosine similarity computations. Let F_{pred} be the SBERT embedding of the predicted answer and F_{gt} the embedding of the ground truth answer. Similarity is defined as:

$$\text{Similarity}(F_{pred}, F_{gt}) = \frac{F_{pred} \cdot F_{gt}}{\|F_{pred}\| \|F_{gt}\|} \tag{1}$$

By averaging these similarities over a test set, Insights are gained into how semantically close the predicted answers are to the gold-standard answers.

IV. Experimental Results

A. Dataset Details

The first significant Visual Question Answering (VQA) dataset was the Dataset for Question Answering on Real-world images (DAQUAR), introduced by Malinowski and Fritz. DAQUAR consists of 6794 training and 5674 test question-answer pairs based on images from the NYU-Depth V2 Dataset. On average, there are about 9 question-answer pairs per image. The questions in DAQUAR focus on real-world scenes, testing a model's ability to interpret complex indoor environments.

In this work, a processed version of the DAQUAR dataset is used. The questions have been normalized for easier consumption by tokenizers, and the image IDs, questions, and answers are stored in a structured tabular (CSV) format. This preprocessing allows the dataset to be directly loaded and utilized for training VQA models, streamlining the data preparation process, ensuring that the experiments are grounded in real-world visual complexity and diverse question types.

B. Hyperparameters

The hyperparameters used in the training process are listed in Table I. These values were chosen to ensure effective optimization and generalization of the VQA model.

TABLE I
List of Hyperparameters

Hyperparameter	Value
Number of Epochs	200
Learning Rate (`lr`)	0.0001
Weight Decay	1×10^{-4}
Loss Function	CrossEntropyLoss
Optimizer	Adam
Scheduler	CosineAnnealingLR
Scheduler Step Size	250
Patience for Early Stopping	100
Device	GPU (if available)

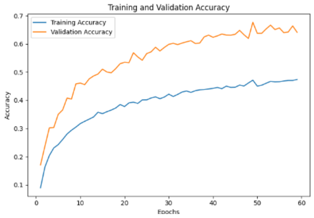

Fig. 3. ResNet+BERT: Training and Validation Accuracy

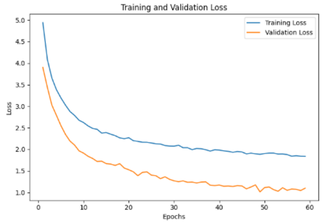

Fig. 4. ResNet+BERT: Training and Validation Loss curves

C. Training Setup and Metrics

Both architectures are trained using the Adam optimizer [14], with cross-entropy as the loss function. Early stopping and a cosine learning rate scheduler helped maintain stable training and prevent overfitting. For semantic evaluation, a pre-trained SBERT model was employed to compute the similarity between predicted answers and ground-truth answers.

D. ResNet+BERT Results

Fig. 3 shows the training and validation accuracy for the ResNet+BERT model. The model gradually improves and converges to a validation accuracy of around 64%. Fig. 4 presents the corresponding loss curves, indicating stable convergence and minimal overfitting.

Table II provides the SBERT similarity scores, demonstrating a mean similarity of 0.76, which suggests that predicted answers often share core semantic meanings with the correct answers.

TABLE II
ResNet+BERT SBERT Similarity Analysis

Metric	Score
Mean Similarity	0.76

E. ResNet+BERT+DETR Results

Although the current validation accuracy for the ResNet + DETR + BERT model, as shown in Fig. 6, has reached around 15%, there is significant potential for improvement. The integration of DETR for object-level feature extraction has enabled the model to better identify and reason about individual entities within the image. However, the model is still in the process of refinement. Future work will focus on optimizing the architecture, including fine-tuning the hy-

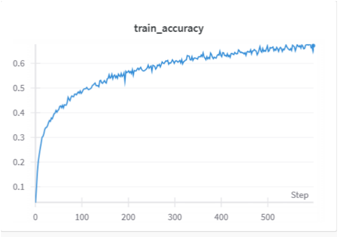

Fig. 5. ResNet+BERT+DETR: Training Accuracy Curve over 200 epochs

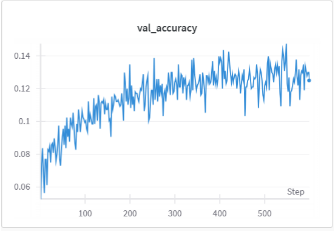

Fig. 6. ResNet+BERT+DETR: Validation Accuracy Curve over 200 epochs

perparameters, incorporating advanced fusion strategies, and exploring additional techniques to enhance its performance. [1]

TABLE III
RESNET+DETR+BERT SBERT SIMILARITY ANALYSIS

Metric	Score
Similarity	0.36

Table III reports SBERT similarity for the enhanced model, with a mean similarity of 0.76. This improvement over the baseline model highlights that not only are the answers more accurate, but they are also semantically closer to the ground-truth responses.

F. Test Results

In this section, examples from the test dataset are presented, with the predictions made by the model highlighted along with their corresponding actual answers and similarity scores.

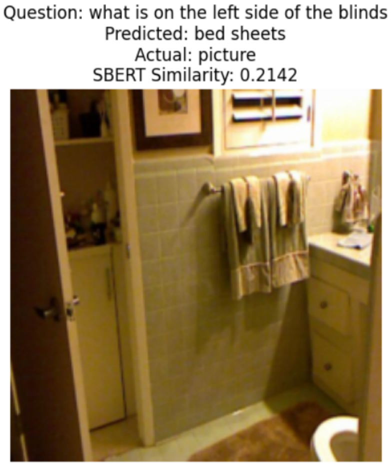

Fig. 7. Example 1: Question: What is on the left side of the blinds?
Predicted: bed sheets
Actual: picture
SBERT Similarity: 0.2142

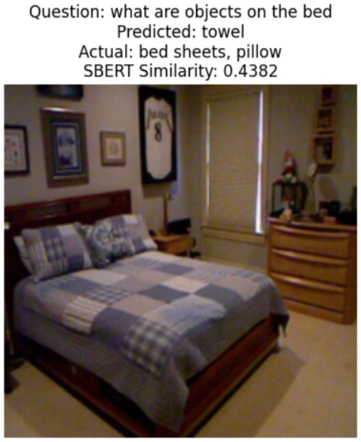
Question: what are objects on the bed
Predicted: towel
Actual: bed sheets, pillow
SBERT Similarity: 0.4382

Fig. 8. Example 2: Question: What are the objects on the bed?
Predicted: towel
Actual: bedsheets,pillow
SBERT Similarity: 0.4382

The above examples demonstrate the model's performance, including cases where the prediction was partially correct but did not fully match the actual answer, as reflected in the similarity scores.

V. DISCUSSION

The integration of DETR features into the ResNet+BERT VQA pipeline shows some improvements, but the results are not yet as expected. While DETR has the potential to enhance object-specific information by enabling the model to better distinguish between similar objects and more accurately link textual queries to the correct visual content, these benefits have not fully translated into the desired performance outcomes. For example, while DETR allows the model to focus on object attributes, its impact on providing more precise and contextually aligned answers remains limited.

[1] Code is available at https://github.com/InfiSmile/Visual_Question_Answering.git.

There are several factors contributing to the suboptimal results. One major challenge is the high computational cost associated with DETR's feature dimensionality, which requires significant memory and processing power. This, combined with the complexities of fine-tuning the ResNet+BERT pipeline, has hindered the expected improvements. The trade-off between achieving better performance and managing efficiency is more pronounced than anticipated, especially when considering real-world applications where resources are often constrained.

Additionally, while SBERT similarity provides a more nuanced evaluation compared to raw accuracy, it still depends on pre-defined embeddings that may not capture all the intricacies of model performance. Some errors, particularly those involving subtle differences in answers, may not be fully accounted for by SBERT alone. Moving forward, exploring more advanced semantic evaluation metrics or incorporating human-in-the-loop assessments could offer a more comprehensive understanding of model performance and guide further improvements.

VI. Conclusion

This paper compared two VQA architectures, one using ResNet+BERT and the other integrating ResNet+BERT with DETR. Our experiments demonstrated that object-level features derived from DETR require further fine-tuning to achieve the desired accuracy and semantic closeness in the answers. While initial results show promise, they also indicate that the model's performance is not yet optimal and requires additional work to better align with expected outcomes.

The use of SBERT-based similarity analysis highlights that the generated responses still need improvement in both correctness and semantic appropriateness. Future directions for this research involve refining the integration of object-level information, exploring computationally efficient methods such as sparse attention mechanisms, model compression, or knowledge distillation. Additionally, experimenting with other pretrained vision models (e.g., ViT, CLIP) and language models (e.g., RoBERTa, T5) may help address current limitations.

As VQA research progresses, integrating richer and more effective multi-modal representations remains a critical step toward building models capable of reasoning more like humans.

Acknowledgment

I would like to express my sincere gratitude to Dr. Anurag Goel for assigning this project and providing invaluable guidance throughout its course. This project introduced me to the fascinating domain of multimodal systems, enabling me to explore the integration of textual and visual modalities. The journey of working on this project was highly insightful and engaging, deepening my understanding of cutting-edge research in this area.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, ”Deep Residual Learning for Image Recognition,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL-HLT*, 2019.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, ”End-to-End Object Detection with Transformers,” *European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[4] N. Reimers and I. Gurevych, ”Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 3982–3992.

[5] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, ”Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6904–6913.

[6] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C.L. Zitnick, and R. Girshick, ”CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning,” *CVPR*, 2017, pp. 1988–1997.

[7] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, ”Stacked Attention Networks for Image Question Answering,” *CVPR*, 2016, pp. 21–29.

[8] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, ”Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering,” *CVPR*, 2018, pp. 6077–6086.

[9] K. Zhou, ”IC-VQA: Image Captioning-based Visual Question Answering Framework,” *arXiv preprint arXiv:2106.01613*, 2021.

[10] L. Li, M. Yatskar, D. Yin, C. Hsieh, and K.-W. Chang, ”VisualBERT: A Simple and Performant Baseline for Vision and Language,” *arXiv preprint arXiv:1908.03557*, 2019.

[11] W. Kim, B. Son, and I. Kim, ”ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision,” *ICML*, 2021, pp. 5583–5594.

[12] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, ”UNITER: Learning Universal Image-Text Representations,” *ECCV*, 2020, pp. 104–120.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ”ImageNet: A Large-Scale Hierarchical Image Database,” *CVPR*, 2009, pp. 248–255.

[14] D. P. Kingma and J. Ba, ”Adam: A Method for Stochastic Optimization,” *International Conference on Learning Representations (ICLR)*, 2015.