A Simpler Approach to Visual Question Answering

Abstract

Visual Question Answering (VQA) is a multi-modal AI task that requires understanding both images and natural language questions to generate appropriate answers. In this project, we develop a efficient yet high-performing VQA model based on a ResInceptionNet image encoder, a GloVe-initialized BiLSTM question encoder, and a 2-layer Stacked Attention Network (SAN) for multimodal fusion. The final classifier consists of a deep MLP with GELU activations and LayerNorm. Trained on the VQA v2.0 dataset, our model achieves strong performance on binary and common-sense questions while highlighting limitations in fine-grained object recognition. The customized model achieves 56.46% accuracy on the validation set, significantly outperforming a ResNet50+LSTM baseline at 38.76%, demonstrating the effectiveness of our design. Overall, this architecture provides a simple, modular, and interpretable alternative to transformer-heavy models for real-time VQA.

1 Introduction

Visual Question Answering (VQA) is a challenging multi-modal task that lies at the intersection of computer vision and natural language processing. The objective is to answer free-form natural language questions based on the content of a given image, requiring the model to perform both visual understanding and semantic reasoning. As a core task in real-world AI systems, VQA has applications in assistive technologies, search engines, robotics, and general visual intelligence.

Recent advances in VQA have been driven by powerful transformer-based models that combine large-scale visual encoders (e.g., vision transformers or object detectors) with language models (e.g., BERT or GPT). While these approaches achieve strong performance, they come at the cost of high computational complexity, limited interpretability, and reduced accessibility for real-time or resource-constrained applications.

This project explores an alternative approach that emphasizes simplicity, modularity, and effectiveness. We propose a lightweight VQA architecture based on a custom ResInceptionNet image encoder, a GloVe-initialized bidirectional LSTM for question understanding, and a Stacked Attention Network (SAN) for multimodal fusion. The model avoids heavy transformer layers while maintaining competitive performance on the VQA v2.0 dataset.

Our contributions include:

- A custom ResInceptionNet for multi-scale and residual visual feature extraction.
- A GloVe-embedded, BiLSTM-based question encoder for capturing semantic context.
- A 2-layer Stacked Attention Network for aligning question representations with relevant image regions.
- A deep MLP classifier with GELU activations and LayerNorm for final answer prediction.

This architecture achieves 56.46% accuracy on the VQA v2.0 validation set, significantly outperforming a baseline ResNet50+LSTM model that reaches 38.76%. The model performs especially well on yes/no and temporal questions, with room for improvement in fine-grained object recognition and open-ended reasoning.

2 Preprocessing

The model is trained and evaluated on the VQA v2.0 dataset, which contains over 1.1 million question–image–answer triplets derived from MS COCO images. Each image is paired with three questions and ten human-provided answers, making the task inherently ambiguous and noisy. The dataset includes both binary (yes/no) and open-ended questions, as well as multiple categories such as object recognition, counting, color, action, and location.

2.1 Image Preprocessing

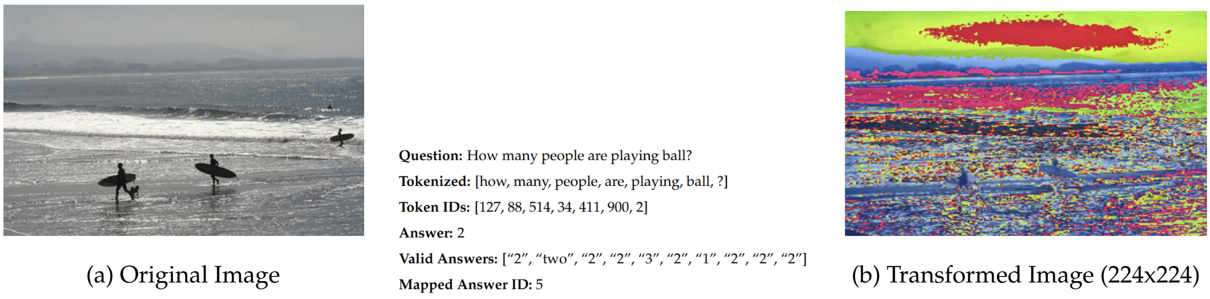
Images were resized to 224 × 224 pixels and normalized using the standard ImageNet mean and standard deviation. Data augmentation techniques such as random horizontal flipping, color jitter, and resized cropping were applied during training to improve model generalization. Image loading and transformation were handled using the `torchvision` library.

2.2 Text Preprocessing

All questions were lowercased and tokenized using a regular expression-based tokenizer. Questions were truncated or padded to a fixed length of 30 tokens. Unknown words not found in the vocabulary were mapped to a special `<unk>` token. The vocabulary was constructed from training data with a minimum frequency threshold, and word embeddings were initialized using 300-dimensional GloVe vectors.

2.3 Answer Preprocessing

Answers provided by annotators were normalized and mapped to a fixed answer vocabulary of the top-K most frequent answers (typically $K = 1000$ or 3000). The most common answer among annotators was used as the training label. For evaluation and soft accuracy scoring, up to 10 valid answers per question were stored and matched during inference.



(a) Original Image

Question: How many people are playing ball?
Tokenized: [how, many, people, are, playing, ball, ?]
Token IDs: [127, 88, 514, 34, 411, 900, 2]
Answer: 2
Valid Answers: ["2", "two", "2", "2", "3", "2", "1", "2", "2", "2"]
Mapped Answer ID: 5

(b) Transformed Image (224x224)

Figure 1: Sample image and associated question/answer preprocessing output

3 Methodology

3.1 Overview

Our goal is to design a lightweight yet effective architecture for Visual Question Answering (VQA), balancing performance with interpretability and efficiency. We start with a simple baseline model using ResNet50 and a unidirectional LSTM, and then extend it to a more advanced pipeline that incorporates a custom image encoder (ResInceptionNet), GloVe-initialized BiLSTM for question understanding, and a Stacked Attention Network (SAN) for multimodal fusion.

3.2 Baseline Model: ResNet50 + LSTM

The baseline model uses a fixed, pretrained ResNet50 backbone to extract global visual features from the input image. The image is resized and passed through the ResNet50 network, and the output of the final average pooling layer is used as a dense feature vector.

On the textual side, the question is tokenized, and each token is embedded using randomly initialized vectors. These embeddings are passed into a single-layer unidirectional LSTM. The final hidden state serves as a summary representation of the question.

The image and question vectors are concatenated and passed through a shallow multilayer perceptron (MLP) to predict the answer from a fixed candidate vocabulary. This model performs reasonably on yes/no questions but struggles with object-level or spatial reasoning due to limited feature interaction.

3.3 Proposed Model

Our proposed model introduces several enhancements to address the limitations of the baseline. The full architecture consists of three major components:

3.3.1 Image Encoder: ResInceptionNet

We design a custom image encoder called ResInceptionNet, which integrates residual connections with Inception modules. This architecture captures both local and global features across multiple receptive fields within each block.

Each Inception module applies parallel convolutions of different kernel sizes (e.g., 1×1, 3×3, 5×5), enabling the model to detect both fine and coarse visual patterns. Residual skip connections stabilize deep feature learning and preserve spatial information throughout the network. The final image representation is obtained by flattening the spatial feature map into a sequence of region-level embeddings.

3.3.2 Question Encoder: GloVe + BiLSTM

The question is tokenized, lowercased, and mapped to pretrained 300-dimensional GloVe embeddings. These embeddings are passed through a two-layer bidirectional LSTM to capture both forward and backward context.

The final hidden states are concatenated and linearly projected to produce a question embedding that summarizes the semantic meaning of the entire question. This vector is then used to guide attention over image features in the fusion stage.

3.3.3 Fusion: Stacked Attention Network (SAN)

To align image and question features, we use a two-layer Stacked Attention Network (SAN). In each attention layer, the question embedding is used to compute attention weights over the sequence of image region features:

$$\alpha_i = \mathrm{softmax}(w^T \tanh(W_v v_i + W_q q))$$

These weights indicate the relevance of each image region to the question. The attended image vector is computed as a weighted sum of the region features:

$$v_{\text{att}} = \sum_i \alpha_i v_i$$

This vector is added back to the original question embedding and passed into the next attention layer. After two such refinements, the final joint representation is passed to a deep MLP classifier with GELU activation and LayerNorm, which outputs the predicted answer.

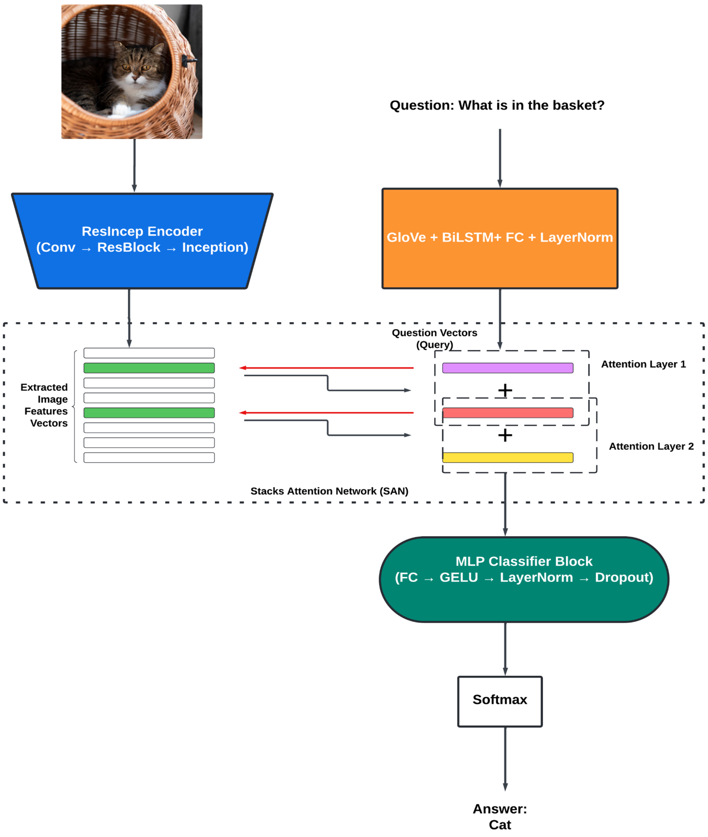

Question: What is in the basket?
ResIncep Encoder
(Conv → ResBlock → Inception)
GloVe + BiLSTM+ FC + LayerNorm
Question Vectors
(Query)
Attention Layer 1
Extracted
Image
Features
Vectors
+
+
Attention Layer 2
Stacks Attention Network (SAN)
MLP Classifier Block
(FC → GELU → LayerNorm → Dropout)
Softmax
Answer:
Cat

Figure 2: model architecture

4 Results

4.1 Quantitative Evaluation

We evaluate both the baseline and the proposed model on the validation split of the VQA v2.0 dataset. Accuracy is computed as the percentage of correctly predicted answers based on soft voting against multiple annotator labels.

Table 1 shows the accuracy of the baseline ResNet50 + LSTM model and our full pipeline.

Model	**Accuracy (%)**
ResNet50 + LSTM (Baseline)	38.76
ResInceptionNet + GloVe + BiLSTM + SAN (Ours)	56.46

Table 1: Validation accuracy for baseline and customized model

The proposed model improves accuracy by 17.7 percentage points over the baseline, demonstrating the benefit of multimodal attention and better feature encoding on both the visual and language side.

4.2 Detailed Performance

To better understand the model's strengths and limitations, we break down accuracy by question type as categorized in the VQA v2.0 dataset. Table 2 summarizes results across common categories.

Category	**Accuracy (%)**
Yes/No	66.26
Counting	38.43
Color	43.48
Location	32.60
Object	34.83
Action	43.17
Temporal	70.02
Other	41.88

Table 2: Accuracy by question type

The model performs best on binary and temporal questions, where visual grounding and commonsense reasoning play a larger role. It performs less consistently on open-ended object recognition or spatial reasoning, where precise localization and object-level features are critical.

5 Discussion

5.1 Findings

Qualitative evaluation of the model reveals several interesting behaviors in how it learns to align visual and linguistic information. For example, when presented with the question *"What is on the plate?"*, and an image of a banana on a plate, the model answers *"food"* instead of the more specific *"banana"*. While this may seem generic, it suggests that the model is aligning key question tokens like "plate" with frequently associated answers such as "food." In this sense, the model is leveraging dataset priors and question–answer co-occurrence patterns in a way that reflects shallow reasoning rather than deep visual understanding.

Stacked attention appears to help in filtering out irrelevant regions and focusing on the overall context. The model performs well on yes/no and commonsense questions, where global visual cues and language priors are often sufficient. However, this behavior may also explain the model's tendency to fall back on "safe" or "statistically likely" answers when faced with object-specific questions that require precise recognition or localization.

5.2 Limitations

Despite improvements over the baseline, the model has several notable limitations. The most prominent is its inability to reject or abstain from answering irrelevant or unanswerable questions. Since the model is trained as a classification system with a fixed answer vocabulary, it is forced to choose the most probable answer for any input, regardless of whether the question can be answered from the image. For example, when shown a picture of a cat in a basket and asked *"What is under the bus?"*, the model still outputs a plausible answer like *"dog"*, even though no bus is present.

This limitation stems from the lack of an "I don't know" or "not answerable" option in the training setup. Without supervision on unanswerable examples, the model cannot learn when to withhold an answer. Additionally, the current architecture lacks explicit object detection or bounding-box level grounding, which limits its ability to distinguish between visually similar or overlapping objects in open-ended scenarios.

6 Future Work

While the current model strikes a balance between simplicity and effectiveness, several avenues remain for future improvement and extension.

6.1 Enhancing Visual Sensitivity

The current ResInceptionNet encoder is effective for capturing global and mid-level visual features, but it lacks the ability to explicitly identify and differentiate between fine-grained object categories. To address this, future models could integrate object-aware mechanisms such as region-based CNNs (e.g., Faster R-CNN) or attention-guided region proposal modules. This would allow the model to reason more precisely about object attributes, spatial relationships, and counts — especially for questions that require distinguishing between visually similar items.

Another direction is to pretrain the image encoder on object-centric datasets such as COCO or Open-Images, or apply self-supervised methods like contrastive learning (e.g., SimCLR or MoCo) to improve feature robustness across varied scenes.

6.2 Improving Question Understanding

On the language side, while the GloVe + BiLSTM encoder is lightweight and effective, more sophisticated NLP encoders such as transformers (e.g., BERT or RoBERTa) can capture deeper semantic relationships and handle more complex question structures. Integrating transformer-based encoders would likely enhance the model's ability to reason through multi-step or compositional questions that require parsing nested clauses or understanding temporal sequences.

6.3 Future Direction

The current model formulates VQA as a classification task over a fixed answer vocabulary, which limits expressivity and prevents the model from expressing uncertainty. A long-term goal is to extend the model to support full-sentence answer generation. This would allow the model to provide more human-like and informative responses, such as *"There is a yellow banana on the white plate."* rather than just *"banana"*.

Achieving this requires a shift toward generative architectures, such as encoder–decoder models or multimodal transformers (e.g., BLIP, OFA, or Flamingo), trained with autoregressive or sequence-to-sequence objectives. These models would better align with the natural variability of human answers and enable the system to handle both answerable and unanswerable questions more flexibly.

Overall, these improvements aim to make the model more detailed, context-aware, and linguistically expressive — advancing it closer to human-level visual understanding and communication.

7 User Interface (GUI)

To demonstrate the functionality of the model in a practical setting, we developed a simple graphical user interface (GUI) that allows users to interact with the system in real time. The interface provides a user-friendly way to test the model by uploading images and entering free-form questions.

The GUI supports the following features:

- Upload an image file from the local device.
- Type a natural language question related to the image.
- View the model's predicted answer instantly after submission.
- Display confidence or top-k predictions (optional).

The interface is built using Python `FastAPI` and `Java Swing`, which supports the same preprocessing pipeline used during training. Once the user submits an image and question, the input is processed and passed through the model backend, and the answer is displayed on the screen.

This interface serves as both a proof-of-concept and a lightweight deployment tool, showcasing the potential for integration into larger real-world applications such as assistive vision, education, or interactive AI assistants.

8 Conclusion

In this project, we designed and implemented a lightweight yet effective Visual Question Answering (VQA) model that integrates a custom ResInceptionNet image encoder, a GloVe-initialized BiLSTM question encoder, and a Stacked Attention Network (SAN) for multimodal fusion. Our goal was to strike a balance between architectural simplicity, real-time responsiveness, and performance — without relying on transformer-heavy or resource-intensive components.

We trained and evaluated the model on the VQA v2.0 dataset, achieving 56.46% accuracy on the validation set and significantly outperforming a baseline ResNet50+LSTM model. The proposed model demonstrates strong performance on binary and commonsense questions, and its stacked attention mechanism helps in aligning relevant image regions with question semantics.

In addition to model development, we built a user-facing graphical interface that enables interactive VQA in real time, allowing users to upload images, ask questions, and view predictions. This helps bridge the gap between research and application, showing how such models can be deployed in practical scenarios.

While the current model has limitations — particularly in fine-grained object recognition and its inability to reject irrelevant questions — it offers a strong foundation for future extensions. With improvements to object-level visual grounding, question understanding, and natural language generation, the architecture can evolve toward more human-like VQA capabilities that handle complex reasoning and produce full-sentence answers.

Overall, this project demonstrates the feasibility of building a modular and interpretable VQA pipeline with competitive performance, opening pathways for both lightweight deployment and further research in multi-modal AI.

References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the ieee international conference on computer vision (iccv).*

Chandrasekaran, A., Prabhu, V., Yadav, D., Chattopadhyay, P., & Parikh, D. (2018). *Do explanations make vqa models more predictable to a human?* Retrieved from `https://arxiv.org/abs/1810.12366`

Desta, M. T., Chen, L., & Kornuta, T. (2018). *Object-based reasoning in vqa.* Retrieved from `https://arxiv.org/abs/1801.09718`

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 (cvpr).*

Huynh, N. D., Bouadjenek, M. R., Aryal, S., Razzak, I., & Hacid, H. (2025). *Visual question answering: from early developments to recent advances – a survey.* Retrieved from `https://arxiv.org/abs/2501.03939`

Kim, J.-H., Jun, J., & Zhang, B.-T. (2018). Bilinear Attention Networks. In *Advances in neural information processing systems 31* (pp. 1571–1581).

Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr).*

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp).*

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr).*

Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., & Xie, W. (2023). Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415.*