A Simpler Approach to VQA: GloVe + LSTM And a Lightweight SAN-ResInception Model

Baldur Hua
Johns Hopkins University

MOTIVATION

Understanding visual scenes through natural language is a critical step toward real-world AI. **Visual Question Answering (VQA)** bridges vision and language by answering free-form questions based on images.

However, most state-of-the-art models are large and compute-heavy. This project explores a lightweight yet effective architecture using a custom **ResIncep image encoder**, **GloVe+LSTM** for question understanding, and **stacked attention** to align modalities efficiently.

Task: Answer the given question based on the image

Q: What is the dog on the right doing?

A: Lying down.

CONTACT

Baldur Hua
Email: chua6@jh.edu

INTRODUCTION

- **Visual Question Answering (VQA)** combines image understanding with natural language reasoning.
- Existing models often use **heavy transformer-based architectures**, which limit deployment and interpretability.
- To address this, we introduce a **lightweight** alternative with:
  - A **ResInception** image encoder,
  - A **GloVe+LSTM** question encoder,
  - And a **Stacked Attention Network (SAN)** for fusion.
- Our experiments use the **VQA v2.0** dataset with over **1.1M** image–question pairs.
- The model classifies among the **top-10 most frequent answers**, balancing **simplicity, accuracy**, and **practicality**.

Preprocessing

Image Preprocessing: Images were resized to 224×224 and normalized using ImageNet stats. Data augmentation included random flips, color jitter, and resized cropping for better generalization.

Text Preprocessing: Questions were lowercased, tokenized, and padded/truncated to 30 tokens. Unknown words were mapped to <unk>. Word embeddings were initialized with GloVe-300.

Answer Preprocessing: Answers were mapped to label IDs using a fixed vocabulary. The most common answer served as the training label; up to 10 valid answers were saved for soft accuracy scoring.

Figure 1. Original Image

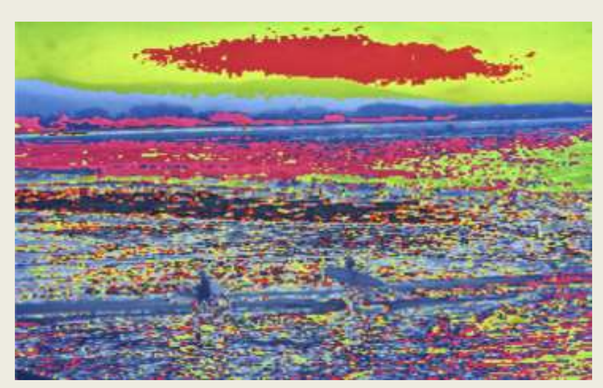


Figure 2. Transformed Image

Q: hazy or sunny ?
Tokens: ['hazy', 'or', 'sunny', '?']
Token IDs: [7433, 10818, 15401, 356]
Valid Answers: ['Sunny']
Answer IDs: [67]

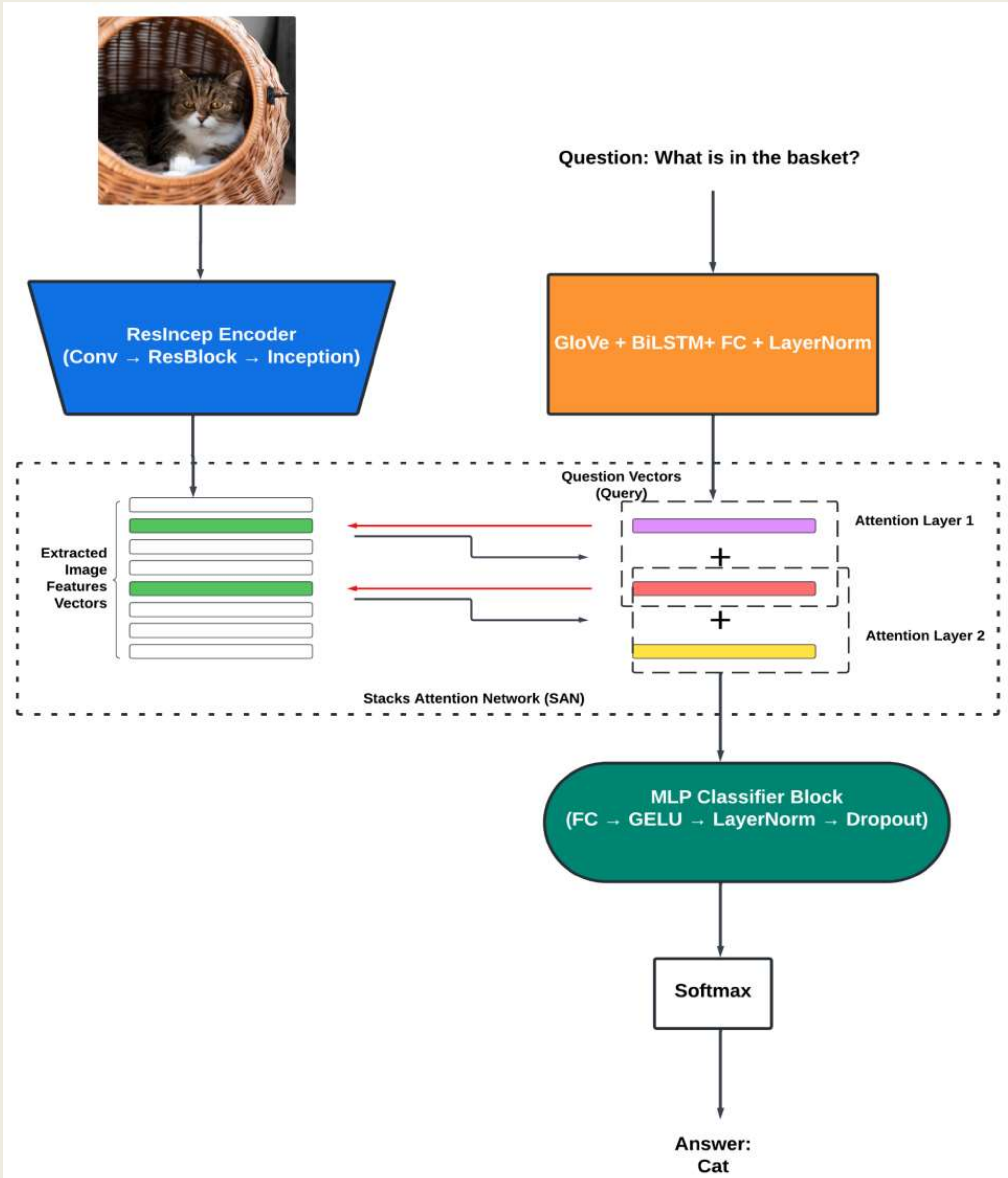
MODEL ARCHITECTURE

We build a **Visual Question Answering (VQA)** model that combines a custom **ResInceptionNet** for image encoding with a **GloVe-initialized BiLSTM** for question understanding. The two modalities are fused using a **Stacked Attention Network (SAN)** to enable reasoning over relevant image regions based on the question. The fused features are passed through **a multi-layer perceptron (MLP)** for answer classification.

Key Components:

- **Image Encoder**: Residual + Inception blocks for rich feature extraction
- **Question Encoder**: GloVe embeddings + 2-layer Bidirectional LSTM
- **Fusion:** 2-layer stacked attention network
- **Classifier:** Deep MLP with GELU and LayerNorm
- **Output:** Top-1 answer from a vocabulary of 1000 candidate answers

Figure 3. Model Structure



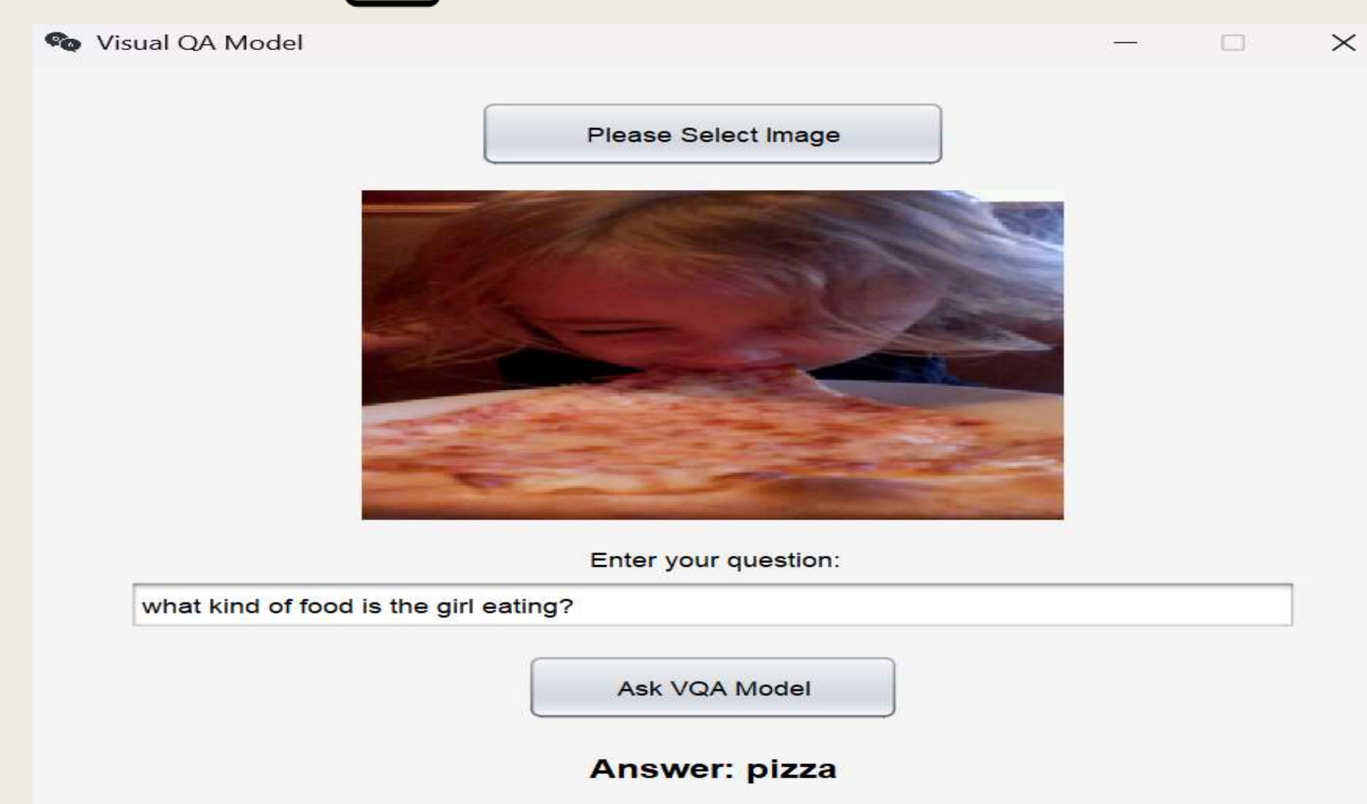
Result

Category	Description	Accuracy
Yes/No	Binary questions (is/are/do)	66.26%
Counting	"How many", "What number"	38.43%
Color	"What color is..."	43.48%
Location	"Where...", rooms, scenes	32.60%
Object	"What is", "What are"	34.83%
Action	"What is x doing", "why..."	43.17%
Temporal	"What time..."	70.02%
Other	Miscellaneous/unclassified	41.88%

Question: how many are playing ball?

Top Predicted Answers:
1 : 0.2734
2 : 0.2590
3 : 0.1247
4 : 0.1051
5 : 0.0558
0 : 0.0537

GUI



Conclusion

- The model successfully answers a wide range of visual questions with reasonable accuracy, especially in yes/no and temporal categories.
- However, performance drops in open-ended and fine-grained questions due to limited contextual understanding and answer coverage.
- Integrating a richer answer vocabulary and refining image-text alignment remain key steps for improving prediction quality.