Final Project : **Visual Question Answering (VQA) System for Real-World Image Understanding**

CS 6120: Natural Language Processing

Group: 26

Author: *Rishabh Singh ; NUID : 002767904*

Github Repo Link : Multimodal VQA with GenAI utilizing LLM

Instructor: *Prof. Uzair Ahmad*
Department: Computer Science

Contents

1 Abstract

This report presents the implementation and evaluation of Visual Question Answering (VQA) systems using multimodal transformers in PyTorch. The scope of this project includes feature extraction from images and text, multimodal fusion, and answer prediction modeled as a multiclass classification task. In this project, I've used two approaches — classification and generation — are explored using the DAQUAR dataset. Key findings demonstrate that BeRT+ViT (Encoder) with GPT2 (Decoder) achieves superior performance showcasing the effectiveness of robust pre-trained embeddings in multimodal fusion tasks. The report concludes with insights into methodological challenges, evaluation metrics, and future research directions.

2 Introduction

2.1 Background

Visual Question Answering (VQA) is a complex, AI-complete task requiring understanding and reasoning over multiple modalities; natural language and visual data. This task has garnered significant attention due to its applications in assistive technologies, autonomous systems, and human-computer interaction. Recent advancements in transformers for Natural Language Processing (NLP) and Computer Vision (CV) have further bolstered the capabilities of VQA systems.

2.2 Objective

This project aims to explore multimodal fusion models for VQA, focusing on:

- Evaluating feature extraction techniques using text and image transformers.
- Comparing late fusion-based multimodal architectures.
- Using evaluation metrics like accuracy, macro F1 score, and Wu and Palmer Similarity (WUPS).

2.3 Scope

The project is centered on papers and techniques published in the last decade, prioritizing those with high citation counts and relevance to VQA tasks. The DAQUAR dataset was chosen for its suitability in single-word/phrase-answer modeling.

3 Methodology

- Vision Transformers (ViT): For image feature extraction, ViT will be used due to its ability to tokenize images into patches and learn spatial relationships through self-attention mechanisms, providing rich image representations.
- Language Transformers (e.g., BERT, RoBERTa): These models will serve as the text encoder, embedding the questions into semantically meaningful representations.

- Multimodal Fusion Techniques:
    1. Late Fusion: Combines outputs from separate language and image encoders at a high level to capture interactions across modalities.
    2. Attention Mechanism: Implemented to focus on relevant parts of both the image and the text for more contextually accurate answers.
    3. Bilinear Pooling: Applied to jointly represent text and image features through element-wise multiplication or pooling layers, facilitating more complex cross-modal interactions.
- Regularization Techniques:
    1. Dropout: To mitigate overfitting, we will apply dropout layers, especially within the fusion and prediction layers.
    2. Gradient Clipping: Gradient clipping will prevent gradient explosion in backpropagation.
- Implementation Tools:
    - Hugging Face Transformers
    - PyTorch
    - NLTK
    - Scikit-learn
    - Datasets

3.1 Datasets

To ensure the system is trained on diverse and complex visual and textual data, I will be using DAQUAR (DAtaset for QUestion Answering on Real-world images): Contains 12,500 question-answer pairs, focusing on indoor scenes and basic objects, ideal for initial testing and experimentation. The dataset was made available here - DAQUAR Dataset

3.2 Assessment Methodology:

- Wu and Palmer Similarity (WUPS): This metric will be used to evaluate the performance, as it captures semantic similarity between predicted answers and ground truths. Accuracy and F1 scores will also be calculated.
- Ablation Study:
    1. Input Dimensions: Varying the size of image patches in ViT and token embeddings in BERT to analyze the model's sensitivity to input granularity.
    2. Pre-processing Techniques: Testing different image preprocessing (normalization and resizing) and text tokenization methods (simple vs. byte pair encoding).

3. Algorithm Complexity: Evaluating the effect of simpler versus more complex fusion mechanisms (e.g., concatenation vs. bilinear pooling) on performance and computational efficiency.
4. Attention Mechanism: Replacing bottom-up attention with alternative attention models to assess their effect on answer accuracy and interpretability.

4 Literature Review

4.1 Thematic Analysis

1. Feature Extraction:
   - Text: Models like BERT, GPT2 were analyzed for their embedding quality.
   - Image: Feature extractors such as ViT demonstrated their ability to encode semantic visual information.
2. Fusion Techniques:
   - Late fusion methods involving concatenation of embeddings followed by linear transformations showed simplicity and effectiveness.
3. Evaluation Metrics:
   - Traditional metrics (accuracy, macro F1) were complemented by WUPS for semantic similarity assessment.

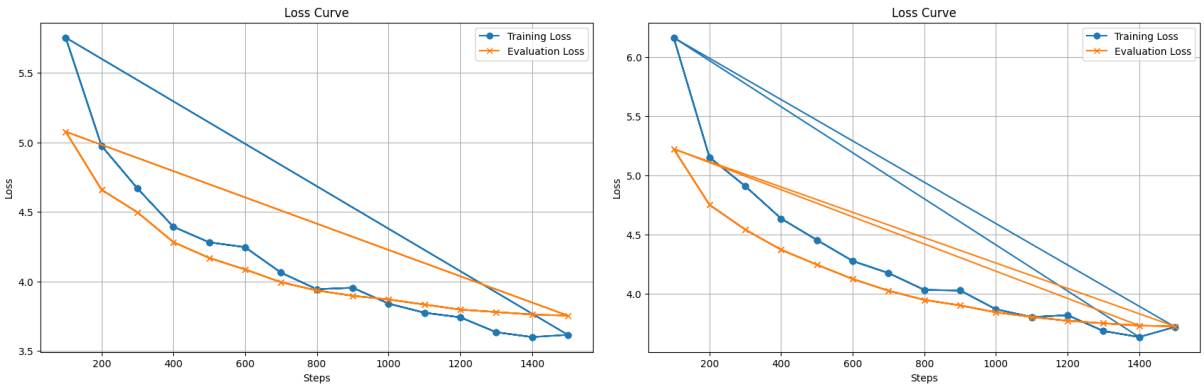
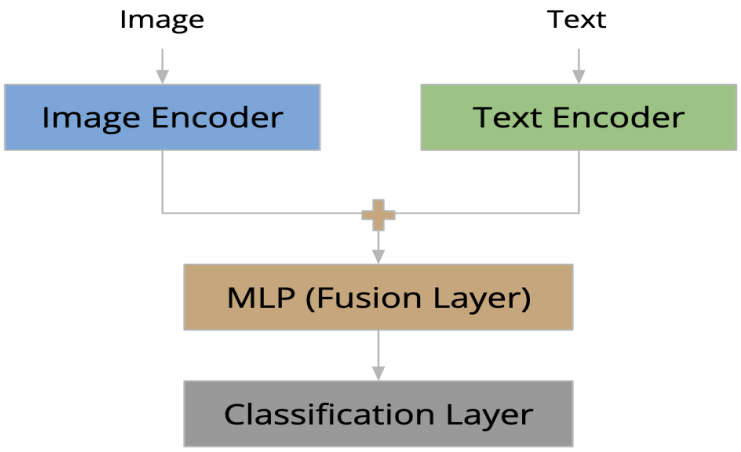


Figure 1: 1.VQA Classification loss Curve ; 2.VQA Generation loss Curve

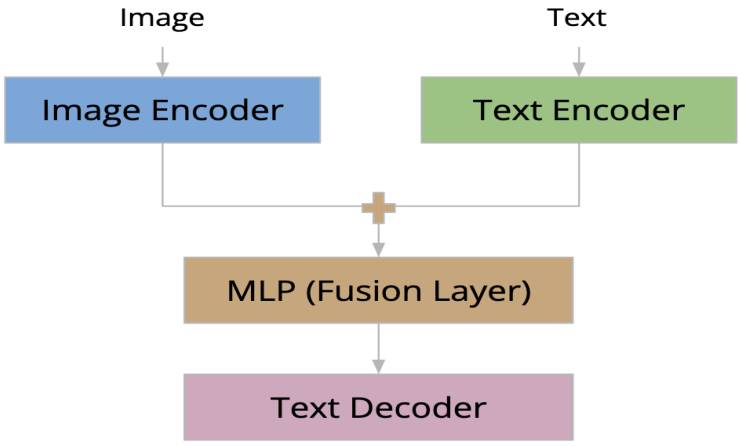
4.2 Comparative Analysis

- The late fusion model for Visual Question Answering (VQA) treats the task as a classification problem. It uses separate encoders for text and image inputs, which are fused together before making a classification prediction.

- BERT + ViT: Good performance with moderate complexity, with a WUPS score of 0.26 and eval loss of 3.76, eval accuracy of 0.21, eval f1 of 0.018.



- The generation model treats VQA as a sequence generation problem. It integrates separate encoders for text and image inputs and uses a decoder to generate textual answers.
- BERT + ViT + GPT2: Highest WUPS and accuracy due to robust embeddings, with a WUPS score of 0.27 and eval loss of 3.72, eval accuracy of 0.22, eval f1 of 0.018.



5 Critical Analysis

5.1 Research Quality

The methods demonstrate high validity through thorough experiments on DAQUAR. Pre-trained transformer models enhance reproducibility and generalization.

5.2 Gaps Identified

- Limited dataset size in DAQUAR restricts model scalability.
- Semantic metrics like WUPS lack coverage for sentence-level answers.

5.3 Implications

Advances in VQA can drive real-world applications in education, healthcare, and autonomous systems. Integrating external knowledge sources could enhance reasoning capabilities.

5.4 Limitations

The study's reliance on pre-trained models may limit adaptability to unseen domains. High computational requirements pose accessibility challenges.

6 Conclusion

6.1 Summary of Findings

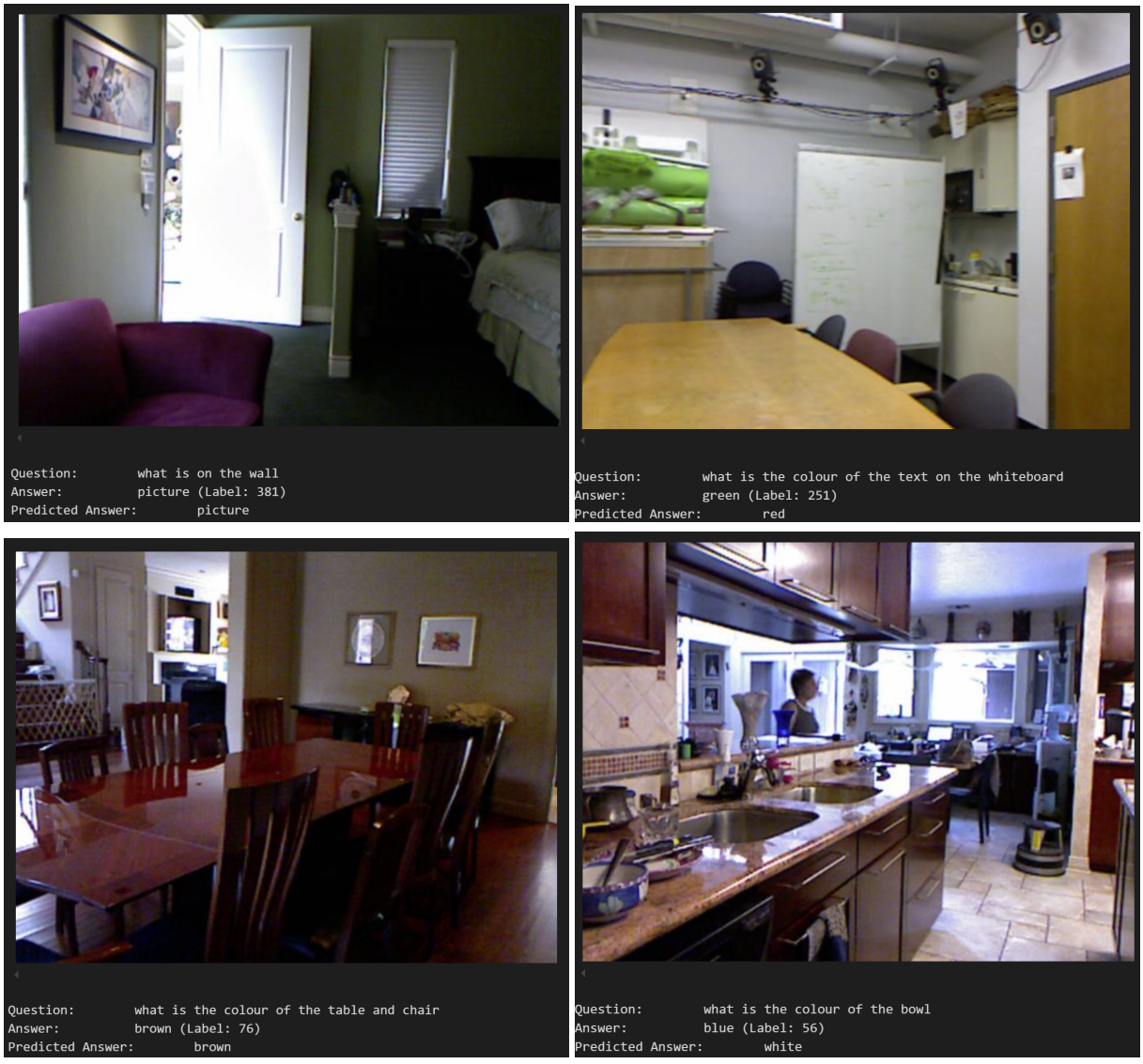


Figure 2: Sample Results of VQA Classification

The BERT+ViT+GPT2 model excelled in VQA tasks on DAQUAR, highlighting the synergy between robust text and image embeddings. Late fusion emerged as a practical approach for multimodal representation.

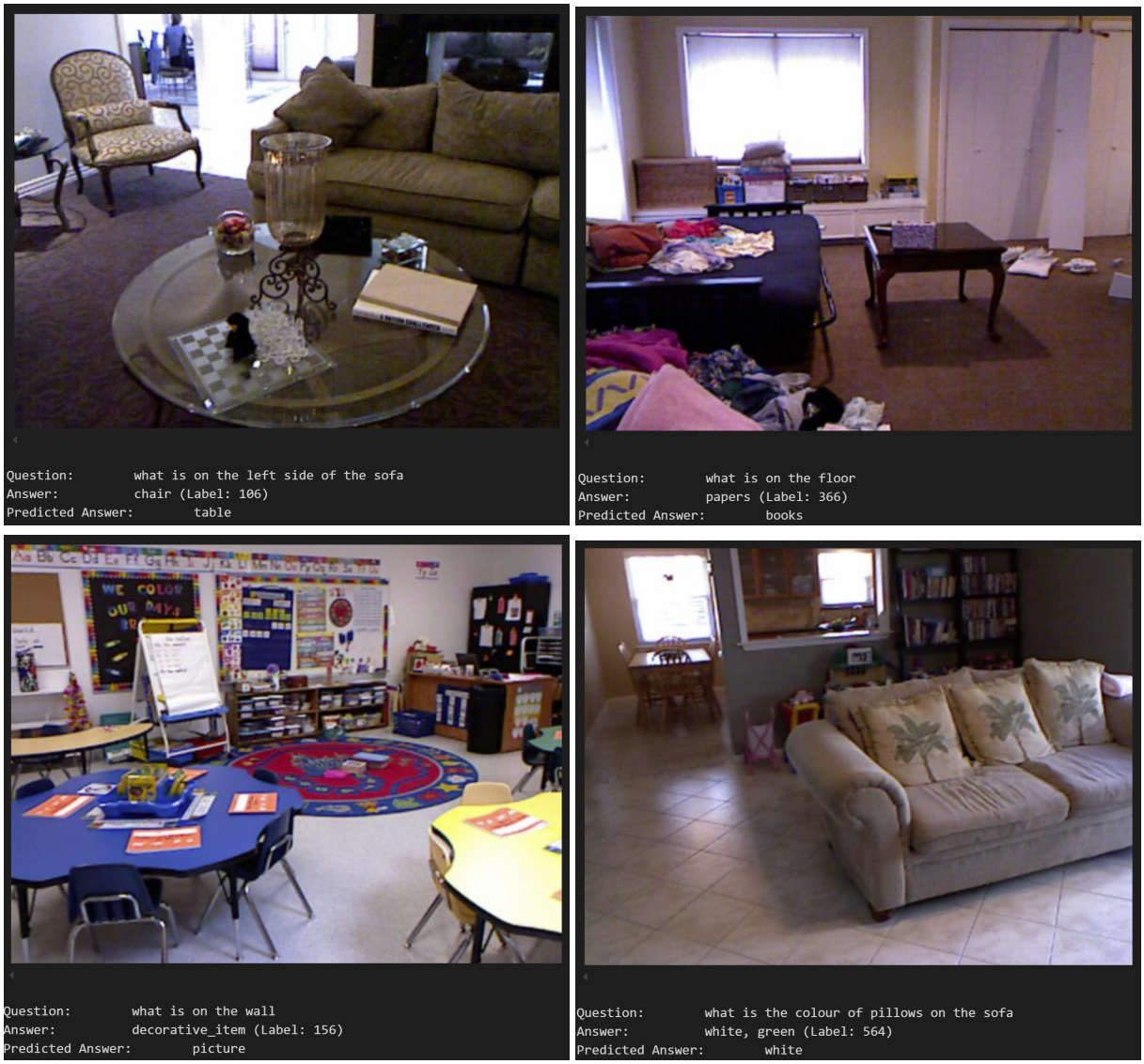


Figure 3: Sample Results of VQA Generation

6.2 Future Directions

- Investigate transfer learning for larger, more diverse datasets.
- Incorporate external knowledge graphs for reasoning-based VQA.
- Optimize computational requirements for broader adoption.

References

[1] Paperspace Blog. Vision transformers explained, 2021. Available at: https://blog.paperspace.com/vision-transformers/.

[2] Tryolabs Blog. Introduction to visual question answering, 2018. Available at: https://tryolabs.com/blog/2018/03/01/introduction-to-visual-question-answering.

[3] Hugging Face Documentation. Transformers v4.15.0, 2022. Available at: https://huggingface.co/docs/transformers/v4.15.0/en/index.

[4] MPI-Inf. Visual turing challenge, 2022. Available at: https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/visual-turing-challenge/.

[5] Nithin Rao. Visual question answering: Attention and fusion-based approaches, 2019. Available at: https://medium.com/@nithinraok_/visual-question-answering-attention-and-fusion-based-approaches-ebef6

[6] Avi Singh. Visual question answering, 2018. Available at: https://avisingh599.github.io/deeplearning/visual-qa/.

[7] Haileleol Tibebu. Data fusion, 2020. Available at: https://medium.com/haileleol-tibebu/data-fusion-78e68e65b2d1.

[8] YouTube. Visual question answering tutorial, 2022. Available at: https://www.youtube.com/watch?v=2sQp7jJJmeg.