Visual Question Answering

Transforming Visual Data into Knowledge:

Taha Bouhafa

Loubaba Malki L'Hlaibi

2nd Year Big Data and Artificial Intelligence

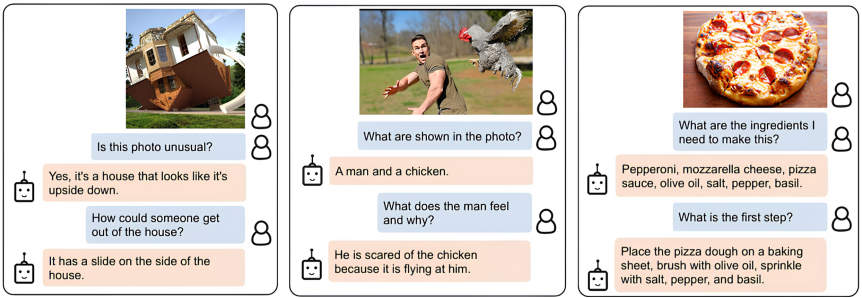
10/01/2025

Prof. Dr.Belcaid Anass

1. Introduction :

Problem Statement

Visual Question Answering (VQA) addresses the challenge of enabling machines to comprehend and process natural language questions in combination with visual content, providing solutions for tasks that require both language understanding and image interpretation to generate accurate answers.



2. Dataset :

- **Description:** The Visual Question Answering (VQA) v2.0 dataset contains open-ended questions about images, requiring an understanding of vision, language, and commonsense knowledge. Due to the large size of the dataset and computational limitations, we used only the training dataset, splitting it into training (20%), evaluation (5%), and testing (5%). This approach was adopted to balance training efficiency and accuracy within the constraints of available GPUs.

Answers Vocab	Question-Image Pairs	Train Split	Val Split	Test Split
1001	443,759	88,751	22,1878	22,1878

Table 2: Statistics of the Dataset Used in the Project

3. General Architecture of the VQA Model :

- **Key components: Vision model, Text model, Fusion mechanism.**

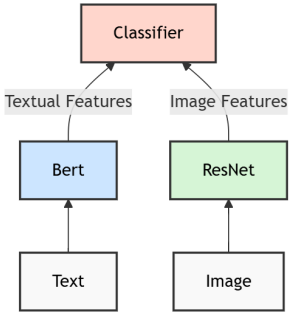


Figure: VQA multi-modal late fusion mechanism

3.1. Vision Models:

- **1. ResNet34 Model:**

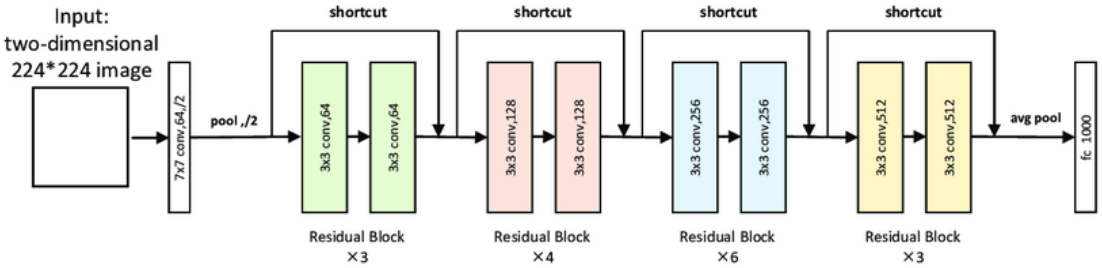


Figure: ResNet34 Architecture

Source: ResearchGate

3.1. Vision Models:

- The ResNet-34 model was trained on our dataset with the following configurations:

Hyperparameters	Values
Input Image Size	224x224
Labels (Number of Vocabulary)	1001
Number of Epochs	8
Learning Rate	1e-3
Batch Size	16
Optimizer	AdamW

Table: Training Configurations for ResNet-34

- **Result:** The model achieved a validation accuracy of **20.41%** after training.

3.1. Vision Models:

- **2. Resnet50:**

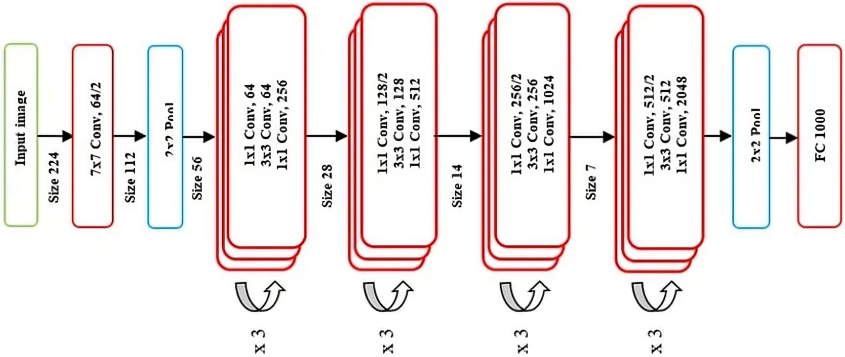


Figure: ResNet50 Architecture

Source: ResearchGate

3.1. Vision Models:

- The ResNet-50 model was trained on our dataset with the following configurations:

Hyperparameters	Values
Input Image Size	224x224
Labels (Number of Vocabulary)	1001
Number of Epochs	5
Learning Rate	1e-3
Batch Size	16
Optimizer	AdamW

Table: Training Configurations for ResNet-50

- **Result:** The best model achieved had a validation accuracy of **19.34%** after training.

3.2. Text Model :

- **3. BERT Model for Question Understanding in VQA:**
- We used 'BertForSequenceClassification' from Hugging Face, adding a classification layer to BERT's architecture for answer prediction.

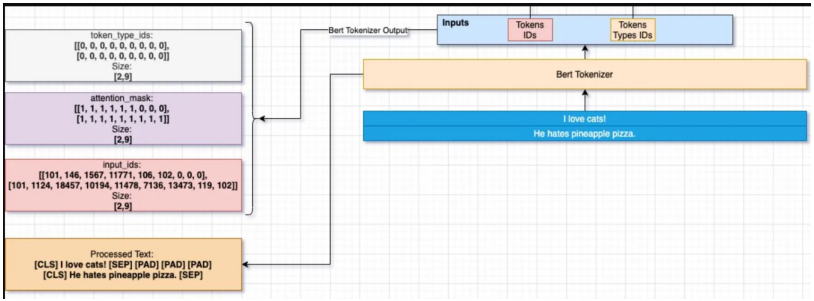


Figure: Input and outputs of the BERT Tokenizer

Source: Medium

3.2. Text Model :

- **BertForSequenceClassification Architecture:**

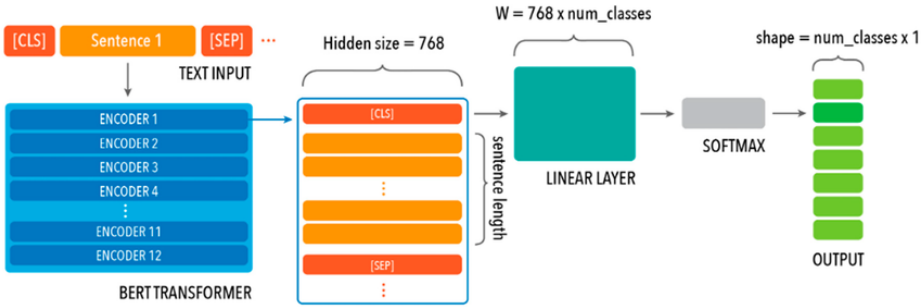


Figure: BERT Model Architecture (12 Transformer Blocks)

Source: ResearchGate

3.2. Text Model:

- **We trained the BERT model on our dataset with the following configurations:**

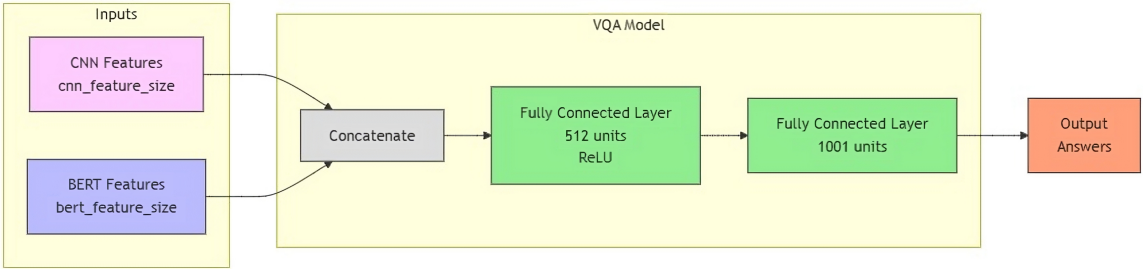
Hyperparameters	**Values**
Input Type	Tokenized Questions
Labels (Number of Vocabulary)	1001
Number of Epochs	10
Learning Rate	5e-5
Batch Size	16
Optimizer	AdamW

Table: Training Configurations for BERT Model

- **Result:** The model achieved a validation accuracy of **40.48%** after training.

3.3. VQA Model

- **VQA Model Combined Architecture:**



3.3. VQA Model

- **We trained our VQA Model on our dataset was with the following configurations:**

Hyperparameters	**Values**
Input Image Size	224x224 image + question
Output Vocabulary Size	1001
Number of Epochs	10
Learning Rate	5e-5
Batch Size	16
Optimizer	AdamW

Table: Training Configurations for VQA Model

- **Result:** The model achieved a validation accuracy of **43.73%** and a **43.46%** test accuracy.

4. Comparison with State-of-the-Art Algorithms

Table 5: Test accuracy analysis with State-of-the-Art algorithms that use BERT.

Model Name	**Test Acc**
ViLBERT	71.79%
VisualBERT	70.80%
Our Model	43.46%

Source: *From Image to Language: A Critical Analysis of Visual Question Answering (VQA) Approaches, Challenges, and Opportunities. (2024)*

5. Challenges and Future Directions :

- **Challenges:**
  - **Hardware Limitations::** Training was hindered by low-performance machines and frequent crashes on Kaggle due to insufficient RAM and GPU memory.
  - **Out of Memory Issues:** Large models like ResNet and BERT caused out-of-memory errors, slowing down training and limiting experimentation.
  - **Training Time:** Long training times and frequent interruptions due to hardware limitations delayed model iteration and optimization.

5. Challenges and Future Directions :

- **Future Directions:**
  - **Improved Training Techniques:** Use mixed-precision or distributed training to speed up the process and manage memory more efficiently.
  - **Larger Datasets:** Utilize larger, more diverse datasets to improve model robustness and generalization.
  - **Domain-Specific Fine-Tuning::** Fine-tune models on specialized datasets (e.g., medical or technical) for more accurate domain-specific VQA.