Are we asking the machine about that picture? Visual Question Answering

Ashfaq Uddin Ahmed, S M Gazzali Arafat Nishan, Tasfia Tabassum
Supervisor: Dr. Mohammad Ashrafuzzaman Khan
Department of Electrical and Computer Engineering, North South University, Dhaka 1229, Bangladesh.

Abstract

Visual Question Answering (VQA) has piqued the interest of researchers in deep learning, computer vision, and natural language processing, a relatively recent challenge in these fields. VQA requires an algorithm to respond to text-based inquiries about pictures. Because many open-ended answers comprise a few words or a closed set of options in a multiple-choice format, VQA adapts itself to automated review. We used the combination of algorithms CNN (for the Image part) and LSTM/GRU/CNN (for the linguistic part), and a second algorithm which uses Vision Transformer (Vit)(for the Image part) and Bert(for the linguistic part).

Methodology

Current state-of-the-art VQA models generally contain a vision part, a question understanding function, and an answer generation part. A typical fusion model for a VQA system involves the following three steps:

- **Featurization of image and question:** The Bert transformer is used to extract features from the question. Likewise, image features are retrieved using an image transformer (the Vision Transformer). For the traditional models, the question understanding part learns a dense question embedding feature vector to encode question semantics with the recurrent neural network (RNN) model. The vision part extracts visual features through a deep convolutional neural network (CNN).

- **Feature fusion:** As VQA includes comparing the semantic information in the image and the question, the features from both modalities must be represented simultaneously. This is often performed through a fusion layer, which permits cross-modal interaction between image and text information to build a fused multimodal representation.

- **Answer generation:** The answer generation part produces an answer conditioned by visual features and the question embeddings. The correct answers might be generated entirely using natural language using a basic classifier model for one-word answers in a defined response space. CNN-LSTM, CNN-GRU, CNN-CNN models were run for 100 epochs. ViT-Bert model ran for 20 epochs with epochs stating from 0.05.

System Diagram

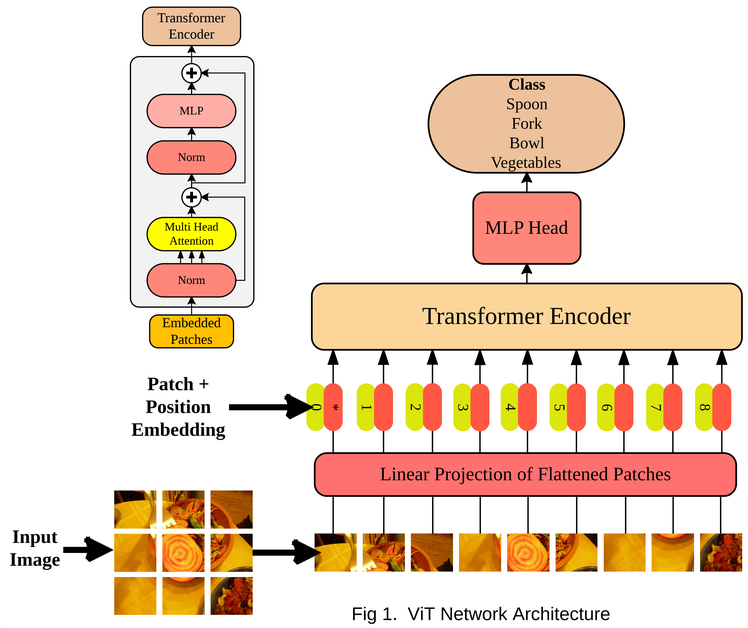

Fig 1. ViT Network Architecture

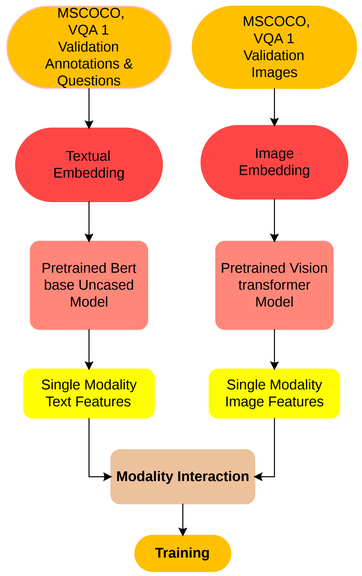

Fig 2. Late Fusion Model Architecture

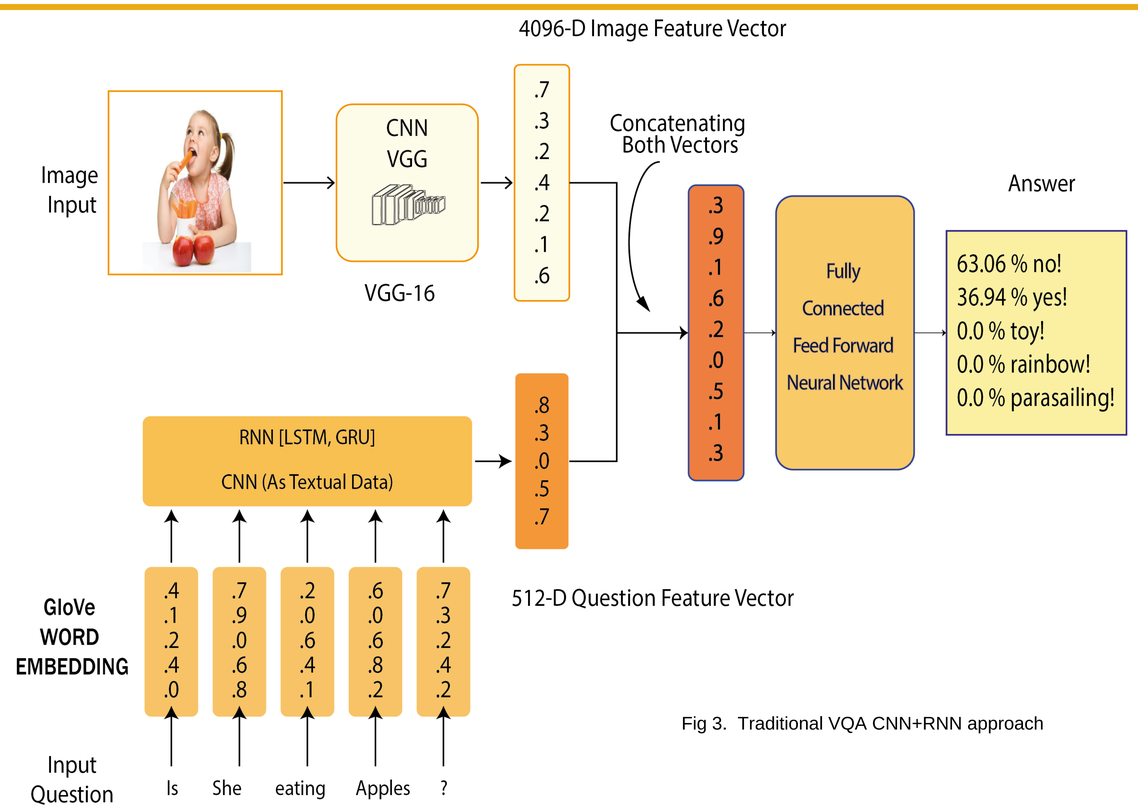

Fig 3. Traditional VQA CNN+RNN approach

Dataset

VQA is a new dataset containing open-ended questions about images. These questions require an understanding of vision, language, and commonsense knowledge.

- The open-source image dataset MS COCO (Microsoft Common Objects in Context) contains 328,000 images of humans and familiar objects.
- VQA dataset is built with images in the MS-COCO dataset.
- Due to computational limitations, we used the VQA V1(MSCOCO) validation dataset from the official VQA website.
- Number of Images in validation dataset 40504
- The dataset contains images with three questions, each with ten potential answers.
- We only evaluate the performance of our approach on the single-word answer QA pairs, which comprise 86.88% of the total QA pairs in the VQA dataset.
- Number of Questions 121,512
- Number of Annotations 1,211,520
- Question and Annotations format (JSON)
- Converted Questions and Annotations from JSON to CSV.

Results

- The output shapes of **GRU and LSTM are the same.** But params are different in the LSTM layer because LSTM models have three instead of two gates.
- CNN seems to be much faster than RNN. CNN appears to be overfitting during training, but the **model's test accuracy was satisfactory**.
- Vision Transformers (ViT) are similar to Transformers in terms of their architecture, analyzing natural language, and CNN's in vision tasks.
- Our result claims that **ViT beats state-of-the-art CNN** with four times fewer CPU resources when trained on enough data.

Dataset	Model	Accuracy	Epoch
V1-VQA Validation	Bert-ViT	0.46	20
V1-VQA Validation	LSTM-VGG	0.43	100
V1-VQA Validation	GRU-VGG	0.43	100
V1-VQA Validation	CNN-VGG	0.44	100