A Comparative Study On Visual Question Answering Using CNN, LSTM, Vision Transformer, and Bert

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 photos. Because many open-ended replies 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).

Method with System Diagram

Traditional LSTM-CNN uses vgg_feast.mat, the file where we have pre-trained model weights for the given dataset. This is fed into vgg16 architecture, with 16 layers of convolutional neural network and a Dense layer(with 4096 units), except the last dense layer(with 1000 units). Concatenating each feature vector of every image provides us with the matrix of body 4096 X 123287.

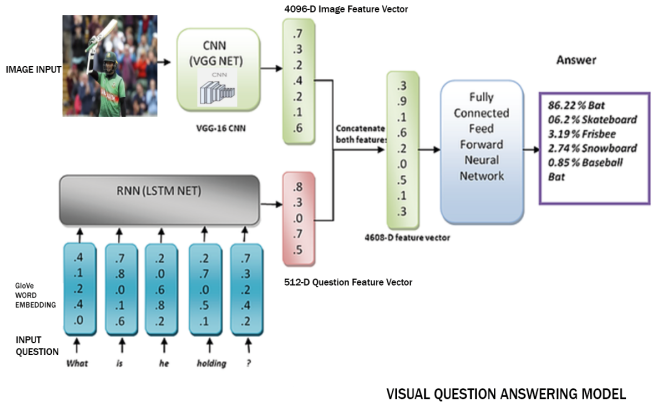

Figure 1: System Diagram

Firstly, we transformed the dataset from JSON formatted questions and annotations to csv format for better comprehension. Then, we have used BERT (Bidirectional Encoder Representations from Transformer) text transformer for encoding questions and ViT (Vision - Transformer): 'google/vit-base-patch16-224-in21k' image transformer for encoding images. Then, we have trained our model for 20 epochs.

Results

In our circumstance, we trained our Bert-ViT model for 20 epochs (epochs start from 0.05) on Google colab pro on a Tesla T4 GPU, which took around 19.5 hours. After completing the training process, it created the three best models, with the greatest accuracy of the model being 46%, which is relatively good for 20 epochs. The rest of the algorithms, CNN(VGG16)-LSTM, VGG-GRU, and VGG-CNN, ran for 100 epochs on the same hardware with an accuracy of 43%, with a run time of 2 hours each. We believe if we train our model on a more extensive dataset, we should see an increase in character recognition rate. The LSTM-CNN model has trainable params: 14,273,315 and The Bert-ViT model has trainable params: 86,567,656

Model	Accuracy	Dataset
Bert-ViT	evaluation accuracy: 0.46 | Epoch: 20	V1-VQA Validation
LSTM-VGG	evaluation accuracy: 0.42 | Epoch: 100	V1-VQA Validation
GRU-VGG	evaluation accuracy: 0.42 | Epoch: 100	V1-VQA Validation
CNN-VGG	evaluation accuracy: 0.43 | Epoch: 100	V1-VQA Validation

Figure 2: Validation accuracy of the models

Novelty of Project

The authors provide a cutting-edge, productive response in the paper VQA: Visual Question Answering. Only the traditional CNN and LSTM were used to implement this strategy. In contrast to their model, ViT and Bert produce very different kinds of prediction output. We suggest two additional algorithms in our study and train and test the Bert and ViT model using the MSCOCO VQA dataset.

Impact on society/environment

Compared to humans, a VQA system can concentrate on different aspects of an image, which can result in many options like, medical VQA, AI-based medical image understanding, associated medical question-answering, or even a VQA for persons who are blind. A VQA that automatically responds to common inquiries could enable blind individuals to live without visual limitations.

Business Model/Feasibility/Financial Scalability plan

In the hot topics of Deep Learning, Computer Vision, and Transformers-based architecture, this research intends to shed light on 4th generation alternatives. This study aims to clarify some of these problems and provides a starting point for describing the environment for the next generation. The VQA community should pay close attention to the Vit & Bert paradigm to achieve performance capable of enhancing practical domain-specific applications. In the long term, we want to create an API that users can use in their software. To generate revenue, we intend to implement VQA for corporations and government organizations based on their needs, as this is far from over.

Conclusion

On the official MSCOCO VQA dataset, we have successfully developed four VQA models, possibly the first to implement ViT and Bert. We have not yet created a demo for that model, though. We have not yet developed an API as suggested above, but we anticipate doing so after we make demos of all of our algorithms.