Visual Question Answering with Encoder-Decoder Variants

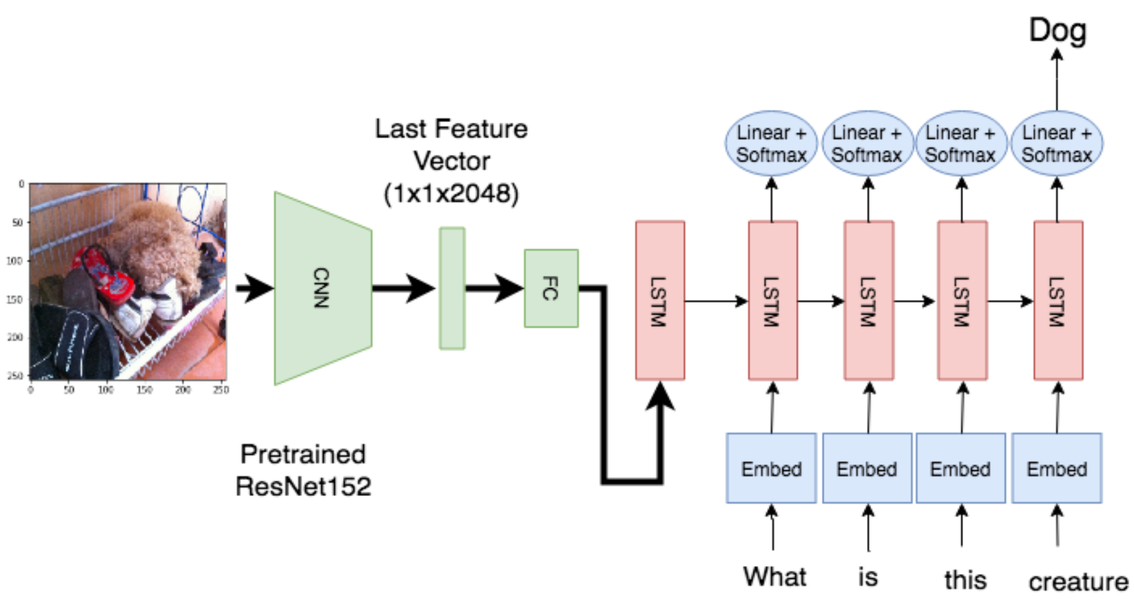
Shiyan Yin, Natcha Simsiri

University of Massachusetts Amherst

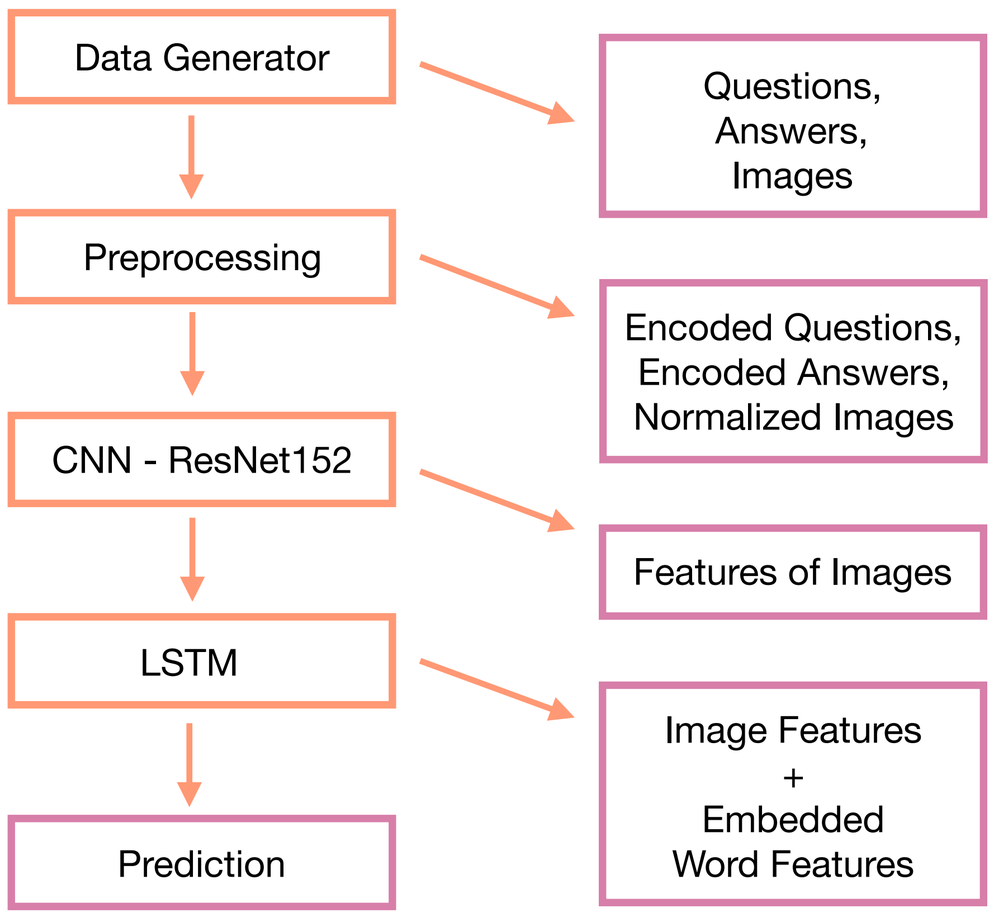
Introduction

Visual Question Answering (VQA) is an interesting problem recently. It uses Deep Learning to learn the information on the image and answer question about the image from the human. This technique can be applied in many areas, one of the most significant use cases is for blind people. VQA is in active research now and has been achieved in several common methods with deep learning. In this project, we used framework CNN + LSTM to achieve the goal, and we will focus on the one-word answer questions which occupy 95% of the entire dataset.

Architecture



Processing Pipeline



Dataset

We use the dataset V2 from visualqa.org. The questions and answers were made from the visualqa research-team with MS-COCO images. Our data consists of the following:

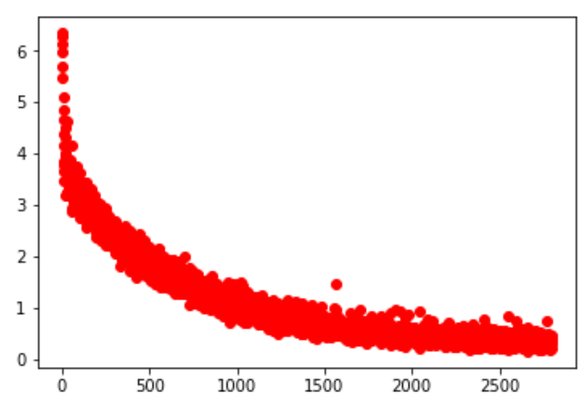
214,354 questions, **2,143,540** answers, **40,504** images.

Under our training mode, we used part of the validation dataset. However, for the feasibility purposes, we will only use **3,000** questions. We will refer to a data-point as a question, image, answer tuple.

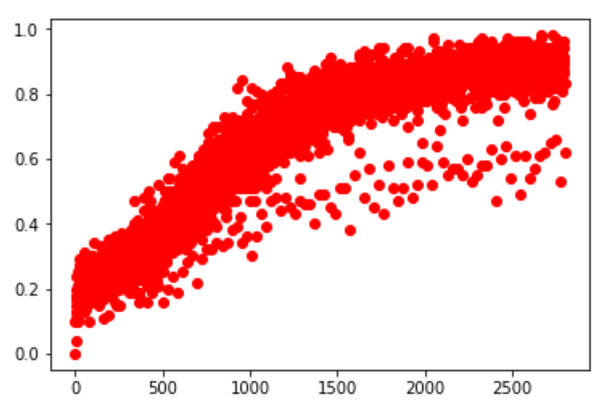
Technical Approach

The encoder-decoder model encodes input of multiple domains and learns a mapping between each one. These models have produced state of the art results such as machine translation and image captions. Our contribution is the application of this model on VQA, where we feed an image and extract the final layer of the ResNet, and perform many to one RNN prediction. Furthermore, we treated QA as a classification problem where the classes in a traditional sense are unique answer vocabulary from our dataset. We trained our model on 3,000 data-points in 100 mini-batches for 100 epochs. Used embedding layer of size 256, hidden size 256 and 1 RNN layer. Furthermore, we optimized with Adam with lr=0.005.

Training Evaluation



Loss, 3000 Questions



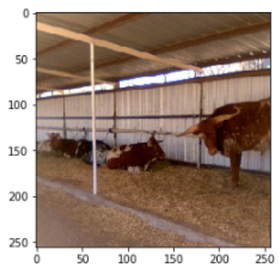
Accuracy, 3000 Questions

Results & Conclusion & Future Work

```
Question: <start> what is to the right of the soup ? <end>
Answer: chopsticks
```

```
Question: <start> is this a creamy soup ? <end>
Answer: no
```

```
Question: <start> why is the cow laying down ? <end>
Answer: tired
```



```
loss: 5.4386467933654785 acc: 0.4 correct: tensor(2, device='cuda:0')
loss: 4.627737522125244 acc: 0.2 correct: tensor(1, device='cuda:0')
loss: 6.835701942443848 acc: 0.4 correct: tensor(1, device='cuda:0')
loss: 4.778101444244385 acc: 0.2 correct: tensor(1, device='cuda:0')
loss: 6.204267978668213 acc: 0.4 correct: tensor(1, device='cuda:0')
loss: 5.908519268035889 acc: 0.4 correct: tensor(1, device='cuda:0')
loss: 4.7726054191589355 acc: 0.3 correct: tensor(1, device='cuda:0')
```

Currently, our ResNet + LSTM encoder-decoder model achieves a result of **17% accuracy on 100 question/images**, With approximately 2000 unique question vocabulary and 500 unique answer vocabulary. This puts random guessing at 0.2% accuracy, and our model performs much better. State of the art non-knowledge based approach achieves approximately 50% accuracy, and we believe our performance (lack of) is attributed to lack of resources on training, fine-tuning and lack of regularization as our model overfits to the training data

VQA is a challenging problem, to train the model well, it also requires more powerful GPU. In the future, we will try another method called Hierarchical Co-Attention Networks (CoAtt). From the experiments, this may be helpful especially for longer questions, which are harder to encode into a single vector representation by LSTMs/ GRUs, so first encoding each word and then using the image to attend to important words helps the model perform better.

Reference

1. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, Yash Goyal Tejas Khot Douglas Summers-Stay Dhruv Batra Devi Parikh
2. Yin and Yang: Balancing and Answering Binary Visual Questions, Peng Zhang Yash Goyal Douglas Summers-Stay Dhruv Batra Devi Parikh
3. VQA: Visual Question Answering, Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh