Deep Learning for Visual Question Answering

Avi Singh

Department of Electrical Engineering, IIT-Kanpur

Introduction

This project proposes Neural Network-based models for Visual Question Answering[1, 2]. In this task, the input is an image and a question based on the image, and the output is the answer (one or more words) to the question. Open-ended question answering requires one to solve several lower-level problems like fine-grained recognition, object detection, activity recognition, common-sense reasoning, and knowledge-based reasoning[1]. Due to the specificity of the task, it can be evaluated automatically, making it easier to track progress. These characteristics make it an ideal "AI-complete" problem, suitable for a Turing Test[3].

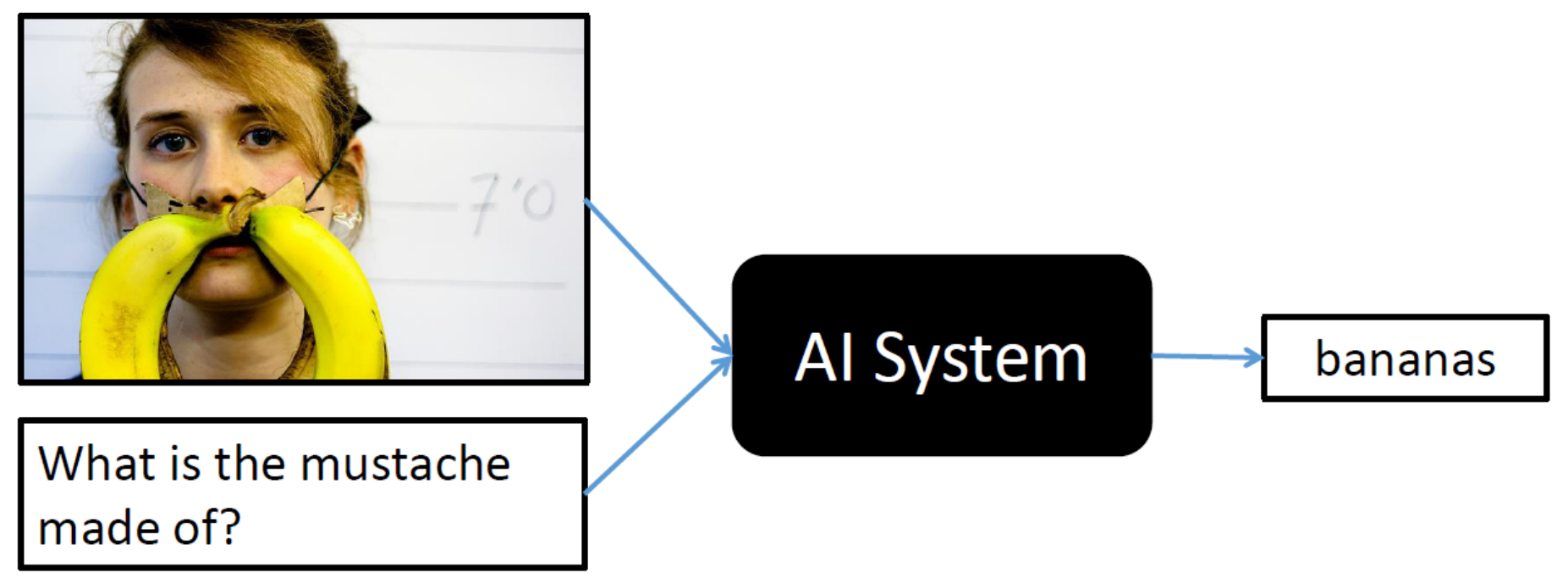


Figure 1: Image from visualqa.org

Preliminaries

- **Convolutional Neural Networks:** ConvNets give state-of-the-art results in supervised computer vision tasks such as recognizing scenes and detecting object. Features learned using CNN are also known to give better performance than hand-designed features when it comes to semantic tasks. We have used the VGG Architecture[4] in our experiments.
- **Recurrent Neural Networks:** RNNs are designed to deal with sequences, and give state-of-the-art performance when it comes to tasks like speech recognition and machine translation. We have used the Long Short Term Memory Network (LSTM), which is a variant of vanilla RNNs, and often give better performance.
- **Word Embeddings:** Methods such as word2vec map words to dense continuous vectors of dimension 300-1000. We have used 300-dimensional Glove Word Embeddings trained on the Common Crawl.

ConvNet and Bag-of-Words

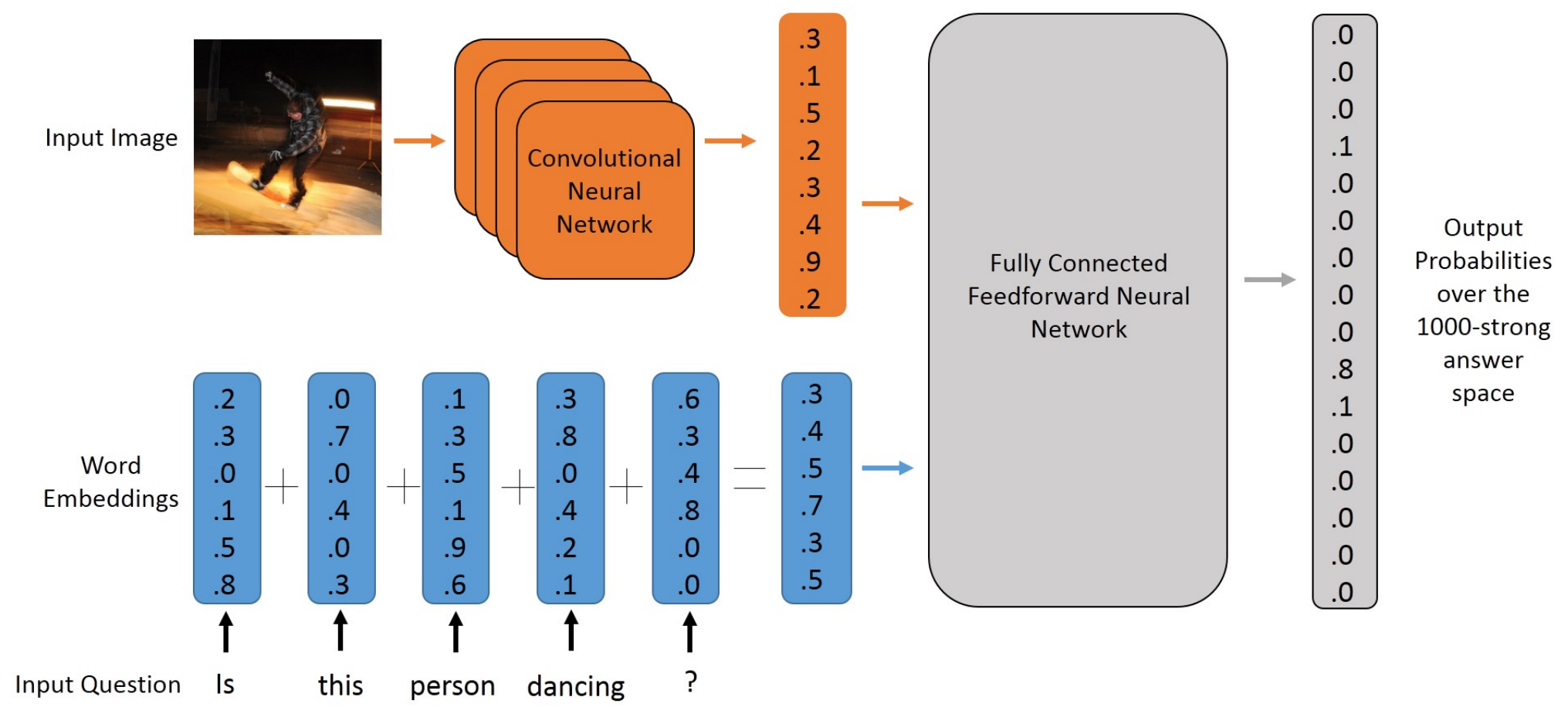


Figure 2: The CNN + BOW model

- We work with a finite answer space by selecting the 1000 most frequent answers from the training set. The problem reduces to multi-class classification.
- The image is passed through a ConvNet, and the activations before the softmax layer are extracted, giving us a 4096-dimensional vector.
- The question is converted to a vector by summing up the word vectors corresponding to all the tokens in the question.
- The two vectors (image and question) are concatenated, and passed through a Multi-Layer Perceptron with three fully connected layers and 50% dropout for regularization.

ConvNet and LSTM

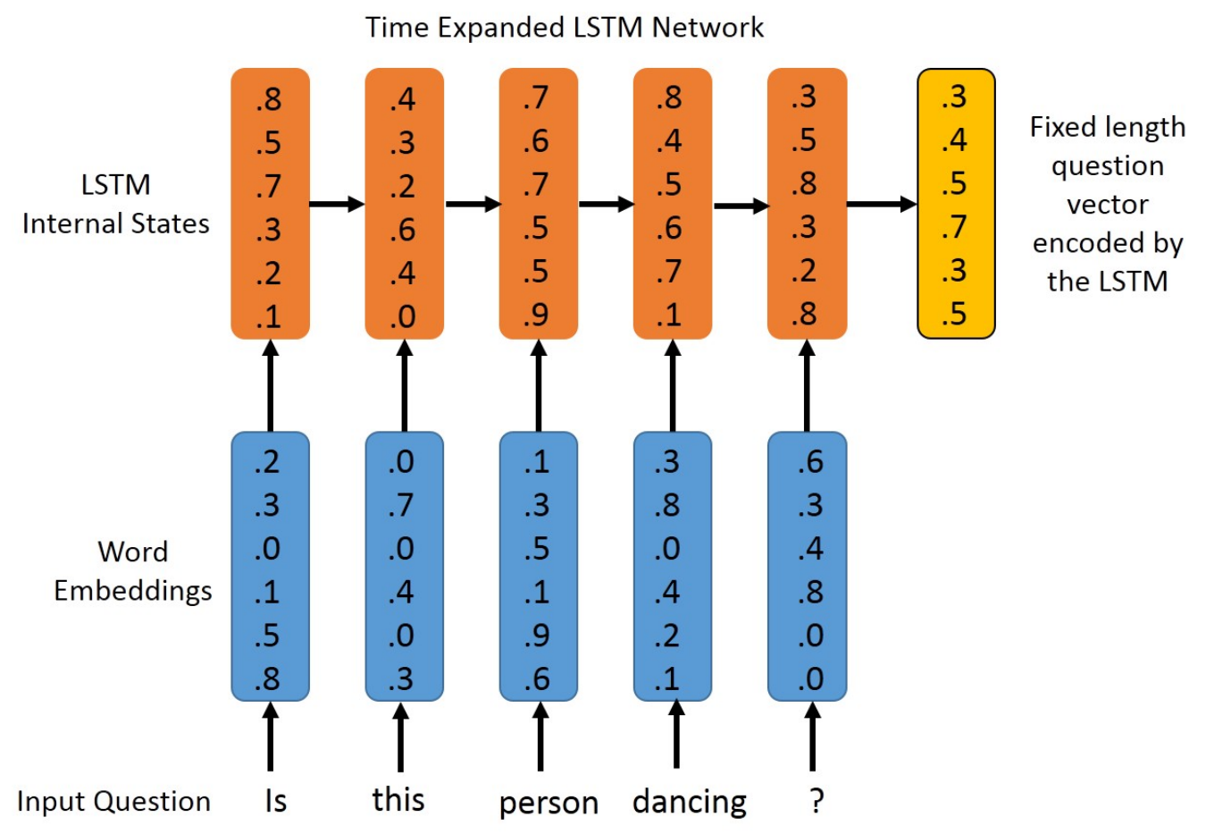
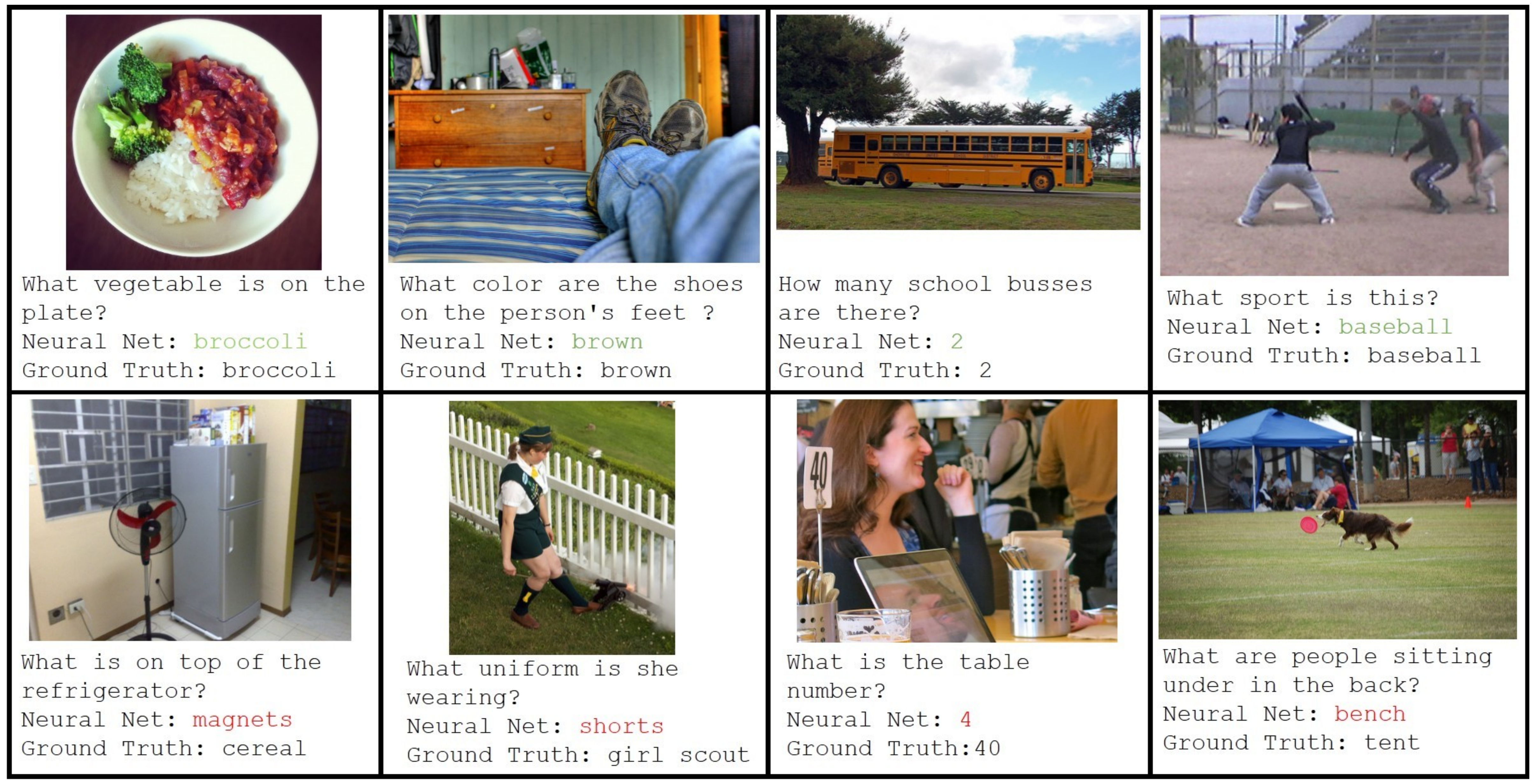


Figure 3: The LSTM question encoder

- The previous model ignores the order in which the words appear in the question, and there is a loss of information when summing up the word vectors.
- To capture the sequential nature of language data, we model the questions using LSTMs. Every word in the question is first converted to its embedding, and these embeddings are passed into the LSTM in a sequential fashion. The final output of the LSTM is used as the *question embedding*.
- This question embedding is concatenated with the 4096-dimensional image vector, and we then apply the same Multi-Layer Perceptron architecture that we used in the previous BOW model. The entire network is trained end-to-end, except the ConvNet.



What vegetable is on the plate?
Neural Net: broccoli
Ground Truth: broccoli

What color are the shoes on the person's feet ?
Neural Net: brown
Ground Truth: brown

How many school busses are there?
Neural Net: 2
Ground Truth: 2

What sport is this?
Neural Net: baseball
Ground Truth: baseball

What is on top of the refrigerator?
Neural Net: magnets
Ground Truth: cereal

What uniform is she wearing?
Neural Net: shorts
Ground Truth: girl scout

What is the table number?
Neural Net: 4
Ground Truth:40

What are people sitting under in the back?
Neural Net: bench
Ground Truth: tent

Figure 4: Some sample outputs generated by the Neural Networks

Results

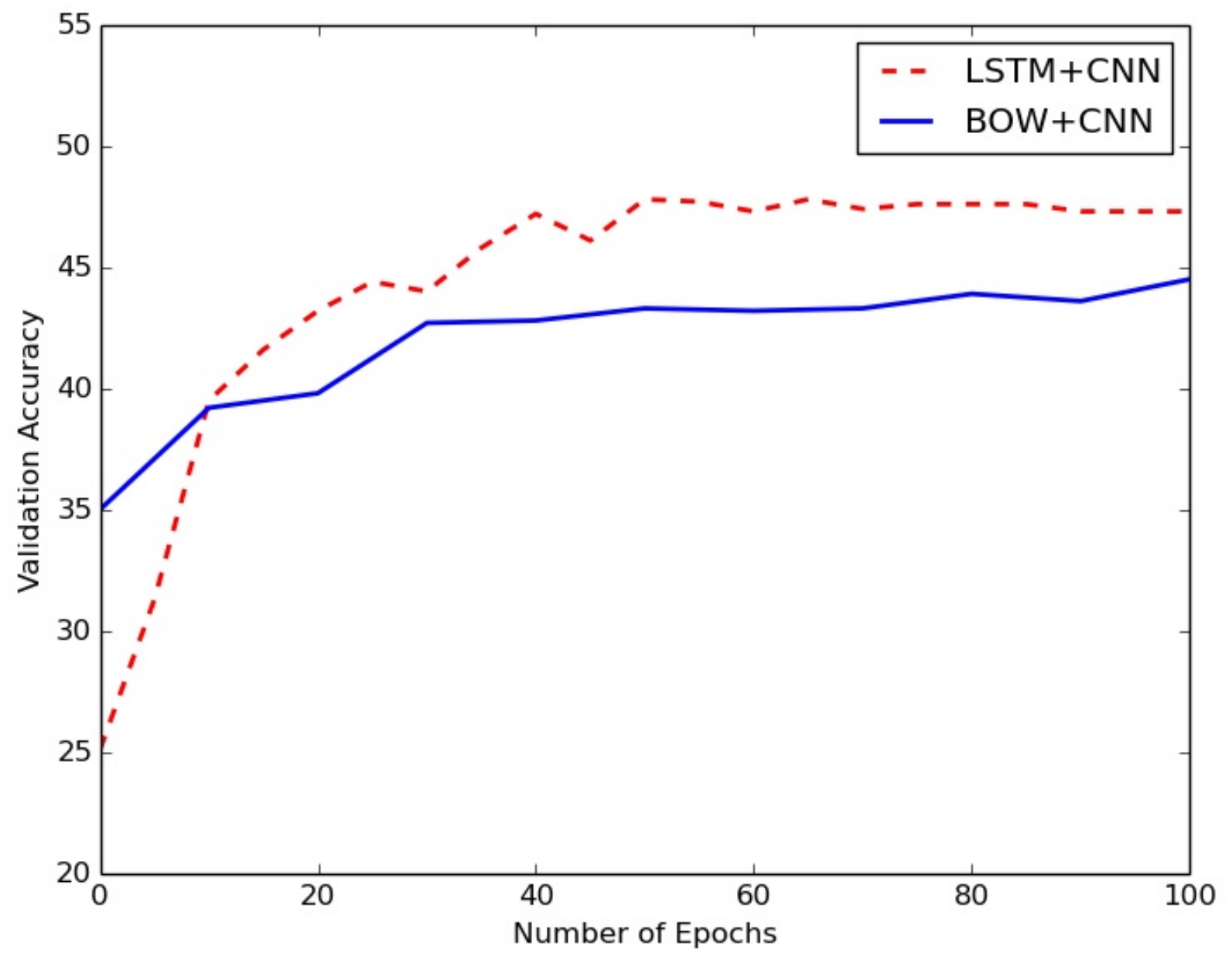


Figure 5: The LSTM needs more epochs, but gives better accuracy.

All our experiments were conducted on the VQA dataset[1] with 240K training set questions for learning and 120K validation set questions for evaluating performance.

Model	Accuracy
BOW + CNN	44.47%
LSTM - Language Only	42.51%
LSMT + CNN	47.80%

Table 1: All results on the validation set of the VQA dataset

References

[1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D.
VQA: Visual Question Answering.
International Conference on Computer Vision (ICCV) (2015).

[2] Malinowski, M., and Fritz, M.
A multi-world approach to question answering about real-world scenes based on uncertain input.
Neural Information Processing Systems (NIPS) (2014).

[3] Malinowski, M., and Fritz, M.
Towards a visual turing challenge.
NIPS Workshop (2014).

[4] Simonyan, K., and Zisserman, A.
Very deep convolutional networks for large-scale image recognition.
In *International Conference on Learning Representations* (2015).