What’s in that picture? VQA system

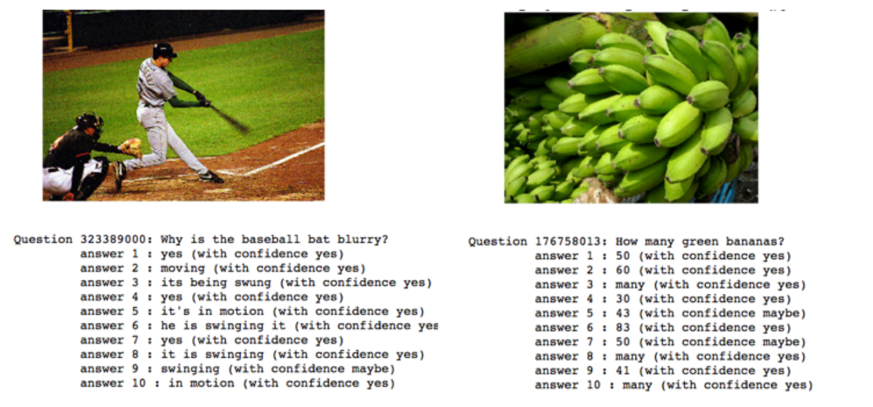
Juanita Ordóñez
Department of Computer Science, Stanford University
ordonez2@stanford.edu

Introduction

Visual Question Answering is a complex task which aims at answering a question about an image. This task requires a model that can analyze the actions within a visual scene and express answers about such a scene in natural language. This project focuses on building a model that answers open-ended questions.

Dataset

- Used Visual Question Answering (VQA)[1] dataset
- 204,721 images
- 3 questions per image
- 10 candidates answers per question
- Wide variety of image dimensions, RGB and grayscale



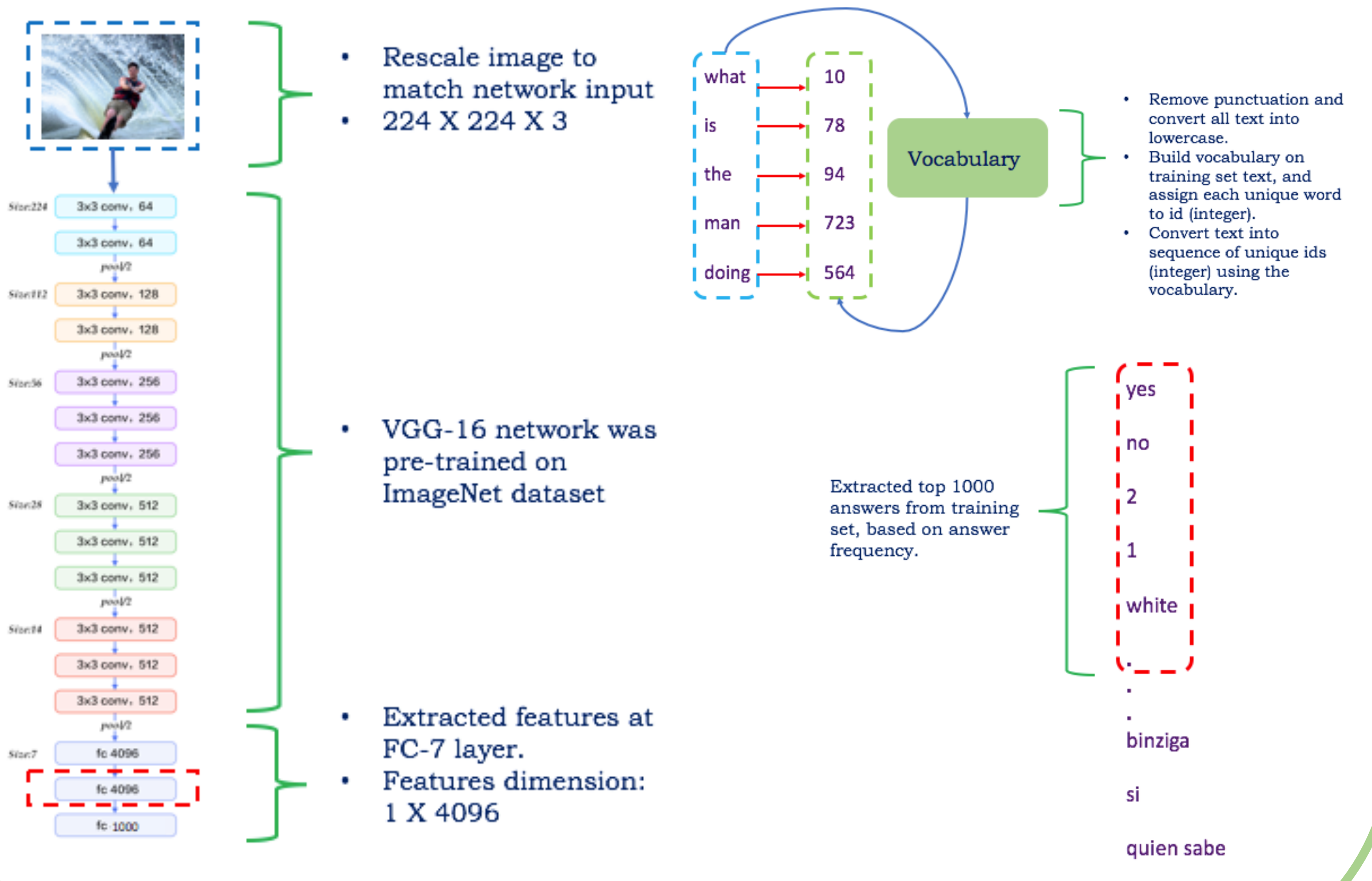
```
Question 323389000: Why is the baseball bat blurry?
        answer 1 : yes (with confidence yes)
        answer 2 : moving (with confidence yes)
        answer 3 : its being swung (with confidence yes)
        answer 4 : yes (with confidence yes)
        answer 5 : it's in motion (with confidence yes)
        answer 6 : he is swinging it (with confidence yes
        answer 7 : yes (with confidence yes)
        answer 8 : it is swinging (with confidence yes)
        answer 9 : swinging (with confidence maybe)
        answer 10 : in motion (with confidence yes)
```

```
Question 176758013: How many green bananas?
        answer 1 : 50 (with confidence yes)
        answer 2 : 60 (with confidence yes)
        answer 3 : many (with confidence yes)
        answer 4 : 30 (with confidence yes)
        answer 5 : 43 (with confidence maybe)
        answer 6 : 83 (with confidence yes)
        answer 7 : 50 (with confidence maybe)
        answer 8 : many (with confidence yes)
        answer 9 : 41 (with confidence yes)
        answer 10 : many (with confidence yes)
```

Figure 1: VQA dataset image, question, and candidates answers examples.

Feature Extraction

- **Images** – Used VGG[2] CNN pre-trained on ImageNet, scaled images to 224x224x3 prior to feeding in network and extracted features from FC-7
- **Text** – Removed all the punctuation, converted to lowercase and built vocabulary on training set.
- **Answers** – Extracted top-1000 most frequent answers from training set. Model predicts a score for each.



- Rescale image to match network input
- 224 X 224 X 3

- VGG-16 network was pre-trained on ImageNet dataset

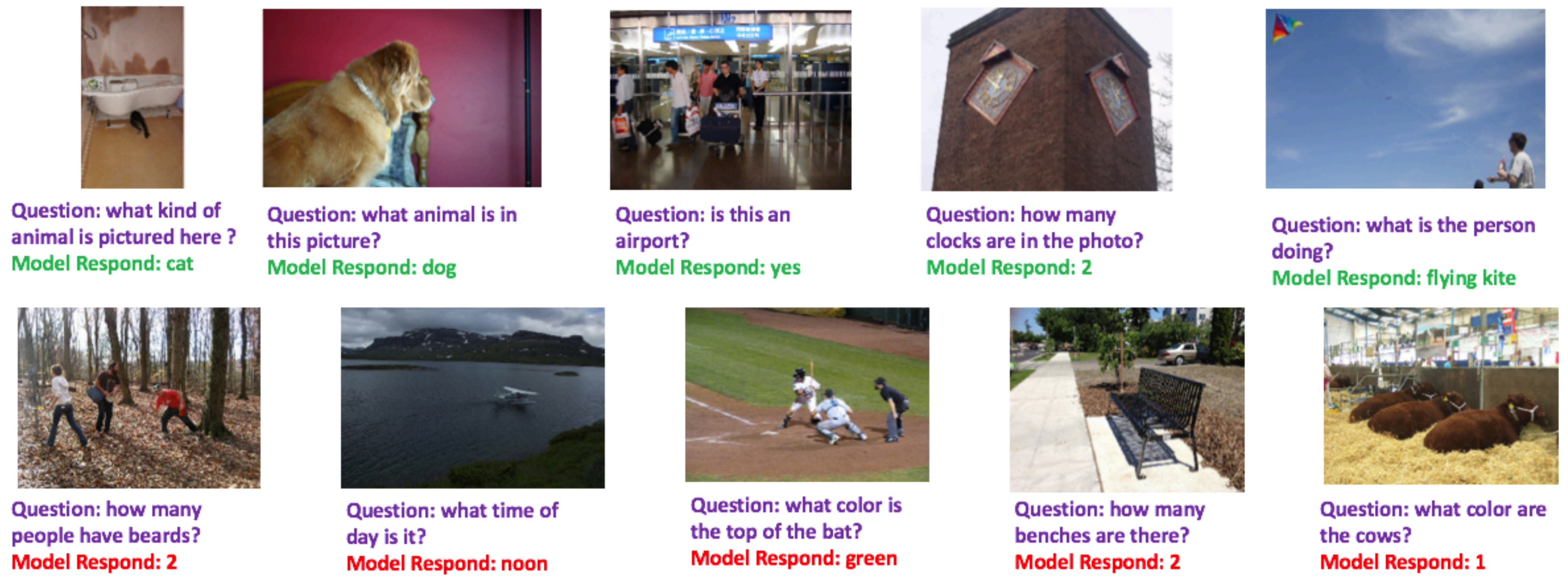
- Extracted features at FC-7 layer.
- Features dimension: 1 X 4096

- Remove punctuation and convert all text into lowercase.
- Build vocabulary on training set text, and assign each unique word to id (integer).
- Convert text into sequence of unique ids (integer) using the vocabulary.

Extracted top 1000 answers from training set, based on answer frequency.

Figure 2: Visual representation of the preprocessing step.

Qualitative Results



Question: what kind of animal is pictured here ?
Model Respond: cat

Question: what animal is in this picture?
Model Respond: dog

Question: is this an airport?
Model Respond: yes

Question: how many clocks are in the photo?
Model Respond: 2

Question: what is the person doing?
Model Respond: flying kite

Question: how many people have beards?
Model Respond: 2

Question: what time of day is it?
Model Respond: noon

Question: what color is the top of the bat?
Model Respond: green

Question: how many benches are there?
Model Respond: 2

Question: what color are the cows?
Model Respond: 1

Figure 4: Qualitative results of model prediction, red indicates the model got the incorrect answer and green represents the model got the correct answer.

Metric

The model performance was evaluated using the VQA score metric. Which a is the model’s answer matches to question candidates responses.

$$VQA_{score}(a) = \min(\frac{a}{3}, 1)$$

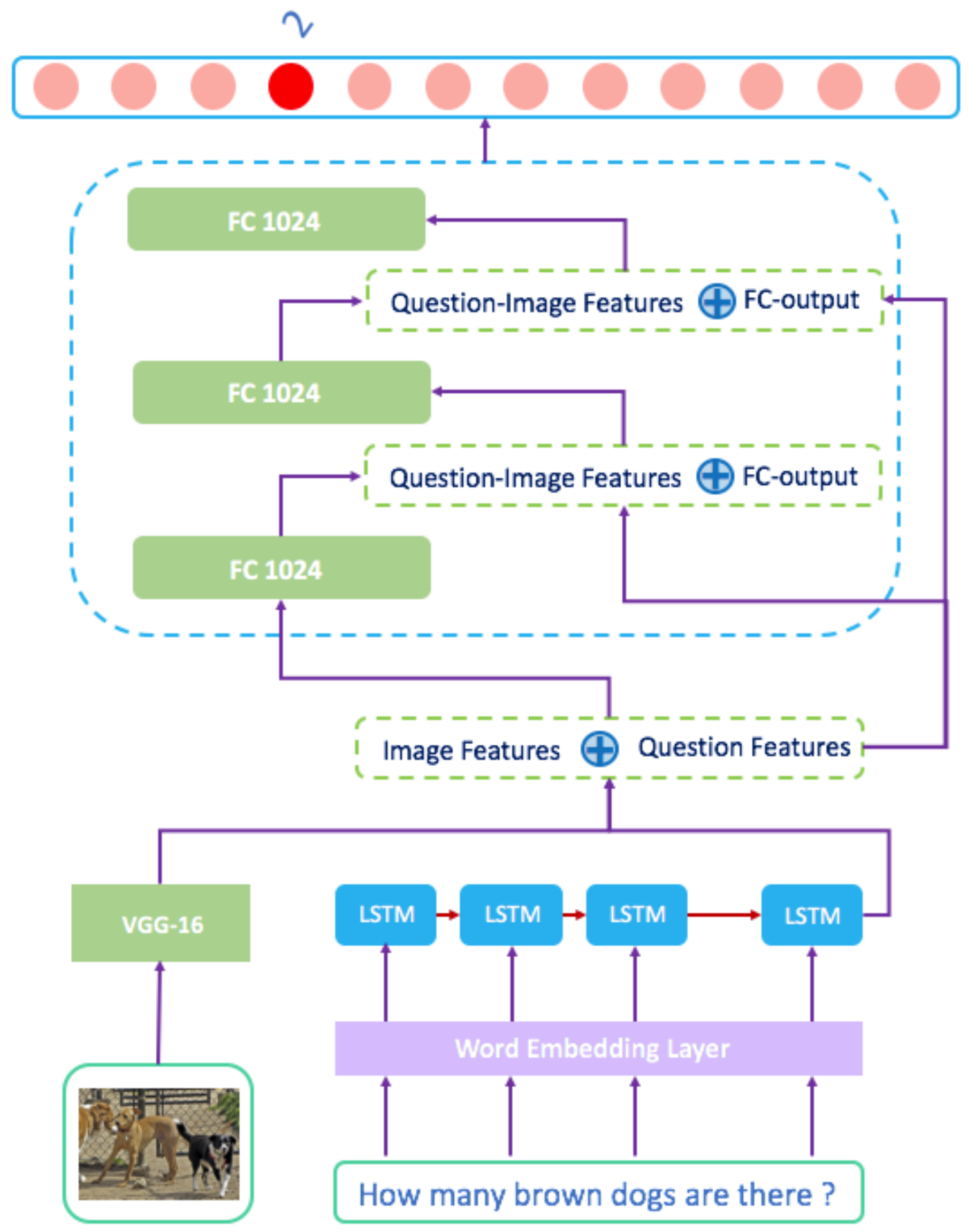
Results

Table 1: Results on my val-test dataset

	all	other	count	yes/no	train all	val-dev all
MLP	48.02	36.67	32.68	63.16	48.55	48.05
RMLP	49.32	36.67	**32.68**	63.14	71.64	49.09
LSTM-RMLP	**51.89**	**41.05**	32.52	**67.76**	78.68	51.8
Language Only	47.67	31.36	32.72	67.22	47.83	47.67

Table 1. results evaluated on the val-test dataset. Each model was trained for a total of 50 epochs with the same hyper-parameters. We show evaluations for the following models: MLP baseline, Recursive MLP with bag of words and LSTM-RMLP. We also show the results of a language-only LSTM-RMLP model wherein no image information is used

Approach



FC 1024
Question-Image Features ⊕ FC-output
FC 1024
Question-Image Features ⊕ FC-output
FC 1024
Image Features ⊕ Question Features
VGG-16
LSTM
LSTM
LSTM
LSTM
Word Embedding Layer
How many brown dogs are there ?

Figure 3: High level visual representation of the model.

Softmax layer for 1000 class.

$$\mathrm{softmax}(y)_i = \frac{\exp(y_i)}{\sum_j \exp(y_j)}$$

Kept Image and question information throughout MLP, this is done by concatenating FC output with question image context vector.

$$h^w = tanh(W_w(v_{qi}^w)))$$
$$h^p = tanh(W_p[(v_{qi}^p), h^w])$$
$$h^s = tanh(W_s[(v_{qi}^s), h^p])$$

Encode question information using LSTM network

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)$$
$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)$$
$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)$$
$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)$$
$$h_t = o_t \tanh(c_t)$$

Where each word vector fed into LSTM cell, and the last hidden state is concatenated with VGG features

Discussion & Future Work

Our results show that encoding the question using an LSTM, as we do in the LSTM-RMLP module, our VQA scores went up by 3.87%. The language-only model only did around 4% worse in comparison to the full information LSTM-RMLP. This result is extremely surprising as it means that the model does quite well in answering questions about an image without ever seeing it. For my next steps I will remove the softmax and generate a respond in a way similar to Sequence to Sequence models. I would also like to explore reinforcement learning training techniques. Finally, I want to experiment with training the VGG-16 model end-to-end.

References

[1]Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[2]Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[3]https://www.tensorflow.org/get_started/summaries_and_tensorboard