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

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.

Approach

Softmax layer for 1000 class.

\[
\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_j e^{z_j}}
\]

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

\[
\Lambda^e = \text{tanh}(W_e x_{e}^{(0)})
\]

\[
B^e = \text{tanh}(W_b^{(0)} x_{b}^{(0)}, \Lambda^e)
\]

\[
\Lambda^l = \text{tanh}(W_l^{(0)} x_{l}^{(0)}, B^e)
\]

Encode question information using LSTM network

\[
i_t = \sigma (W_i x_t + W_h i_{t-1} + W_c c_{t-1} + b_i)
\]

\[
c_t = i_t \odot c_{t-1} + f_t \odot \text{tanh}(W_c x_t + W_h c_{t-1} + b_c)
\]

\[
s_t = \sigma (W_s x_t + W_h s_{t-1} + W_c c_{t-1} + b_s)
\]

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

Results

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

\[
\text{VQA Score}(a) = \min \left( \frac{a}{3} \right)
\]

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