NNFL Project: Visual Question Answering from the CLEVR dataset

- Methodology

    1. Load the training images, resize them and create a dictionary of image names as keys and vectors as values
    2. Load the json file, and create 3 lists, one each for questions, image names and answers
    3. Tokenize the questions and pad them to length 200
    4. Create an embedding matrix for the words present in the questions using glove.
    5. One hot encode the labels for each question
    6. Create a generator that returns a concatenated input of image and question to serve as input to our predictive model.
    7. Create and compile a model that uses conv nets for images and lstm/conv1D for questions and concatenates them into a final dense layer.
    8. Train the model on the training set and validate on 20% of the dataset.
    9. Save the model and use it for evaluation.

- How we came to our final model

    1. We started with a basic model.The
       Images were converted using pretrained VGG which was given as input a single dense layer of size 1024 for image
       For the text part we used lstm with 256 units and then a dropout of 0.5 and then concatenated both the models

and then a dense layer of 1024. We got 44.72% validation accuracy with model size of 69 mb.Then we reduced each layer size by half i.e. image single dense layer to 512 , number of lstm units to 128 and final dense layer to 512.We got 46.53% accuracy with model size of 30 mb.

2. Then we removed VGG weights and created our own convolution network for our image(4 conv2d layer each followed by batch normalization).We then trained and got an accuracy of 46.91% validation accuracy with model size of just 13mb.

3. Then we tried out different methods such as using Conv1D to train the text questions instead of using LSTM. We observed a drastic improvement in training time and also helped us getting a validation accuracy of upto 49%.