COL 865: Deep Learning.
Assignment 2

Due: Thursday, November 16, 11:50 pm. Weightage: 16%

Notes:

- You should submit all your code as well as any graphs that you might plot (see below).
- Include a **single write-up (pdf) file** which includes a brief description for each question explaining what you did. Include any observations and/or plots required by the question in this single write-up file.
- You can use PyTorch. If you would like to use any other language, please check with us before you start.
- Your code should have appropriate documentation for readability.
- You will be graded based on what you submit as well as your ability to explain your code.
- This assignment should be done individually.
- Some questions have been borrowed from existing online resources. You may be able to find some online existing implementations, we will include most of these in moss run. Please refrain using those directly and report any references if you refer them.
- You should carry out all the implementation by yourself.
- We plan to run Moss on the submissions. Any cheating will result in a 0 on the assignment. Additional penalties will be incurreed depending on the scale of cheating (going all the way up to a penalty of **-10**). More serious offenses will be referred to the Department's internal committee for disciplinary actions.

1 Visual Question Answering [11 Points]

In this problem, you will be building an AI system for the task of Visual Question-Answering. Specifically, given an image and a question related to the image in natural language, the system needs to answer the question in natural language from the image scene. The underlying system is required to be good at vision, NLP and common-sense reasoning . For more details about the task, please refer to the VQA website. We will be using the dataset given here

- **Base Models [6 points]:** Implement the following base models and optimize the parameters for the same. You need to figure out the hyperparameter details and size of filters/layers for each model. Report your training time, test time, test accuracy and model size for both the cases. Report the test accuracy breakdown for each four types of answers- (Yes/No, Number, Other and All)
  - **CNN+LSTM:** Encode the image by a CNN and encode the question by a LSTM and then combine these for VQA task. You may refer to this paper for details but feel free to optimize hyperparameters keeping the basic model architecture of CNN for image and LSTM for question constant.
  - **Attention based Model:** This model will incorporate attention over the input image. You can refer to this paper for details. Use the LSTM based question model in the paper and keep $k = 1$ (1 attention layer).

- **Challenge [5 points]:** Improve your base models by some of your own innovations or using state-of-the-art methods and submit your code at evalAI server. All the submission instructions and challenge guideline will be same as found here. Your final submissions should be to the *Real test2017 (oe)* phase on the evalAI server. Note that this allows a total of 5 submissions on the test dataset (per account). We will consider your best score for evaluation. Submit a report of atleast 1 page describing the details of your model, your results and analysis and what techniques lead to significant improvement in scores.
  Deadline: The main assignment deadline is Nov 16 but the challenge results for this part and report can be submitted by Nov 25. No buffer days are applicable for challenge.

2 Visual Question Generation[5 Points]

In this problem, we will be using the same VQA dataset for a related task of Visual Question Generation. More details to follow.