CS 401R Section 1 Homework 3

Part 1:

After reading the paper, viewing the Power Point, and discussing the topic in class, I would best describe the NPI as a small neural network that contains replacement activations for some of the hidden layers of the large language model (GPT2, etc.) which are then added back into the network in order to guide or restrict output without changing the weights of the original model. The diagrams explain the process very well; GPT2 receives input and generates text, and the hidden layer activations are recorded. From there, the NPI is trained using those values, and used once again with the original GPT2 model to produce more controlled output. The process is assisted by the classifiers and discriminator which respectively help determine how to produce the desired input while keeping the text understandable and representative of how GPT2 generates text. I understood bits and pieces of the more technical aspects like the Loss function used and the data set generation, but I will need to re-read to understand those more fully. Overall, I think I have a better grasp on the overall flow of how GPT2 and the NPI work together and what the bigger picture is.

Part 2:

**Generator:**

Contruct\_data.py Lines 41-448 (the whole file essentially).

This one also I was not 100% sure if this is the right exact spot, but based on the description of the generator, this seemed like the class that would make the examples for training

**Classifier:**

train\_classifier.py Line 141-421

**Discriminator:**

Either

train\_npi Line 707 – 756

or train\_npi Line 644 - 704

his one I am not totally sure if it is the right part of the code, but it has parameter such as the discrim\_coeff and style\_coeff and seems to be determining the loss as the difference between expected style and the actual output. However, from reading comments and the diagrams, it looks like the GenerationClassifier works with the ContentClassifier class to train the NPI adversarially, which would make the GenerationClassifier the discriminator.

**NPI:**

train\_npi.py Line 488 - 472