Project Step 4: Data Reader and an initial working model

Title Generation

Course: FS22-CSE-842

Course Instructor: Dr. Kordjamshidi

Group Members:
Ye Ma
Kuldeep Singh
Reza Khan Mohammadi

What is the setting of your project? is it supervised, unsupervised or semisupervised?

Supervised. An approach for Text Generation is to break down the dataset sequentially, where a sequence of n words is expected to see the n + 1 word as the to-be-appended token. Hence, an encoder-decoder architecture, for instance, learns how to map a series of input tokens to another series of outputs in the decoder. This mapping demands an X set to be mapped to Y, which makes us see the supervision happening.

How do you represent X?

We used PyTorch's embedding to learn how to represent present words in the article body in a 50-dimensional space. In further steps of this project, we will investigate more means of representation to enhance performance.

How do you represent Y?

Since we have utilized an encoder-decoder architecture to generate titles given the article body, we should use the proper Y representation to make learning possible for the network. Since we aim to generate a series of words, we set the decoder model to make a multi-class prediction among the vocabulary at each time step. This makes Y to be a sequence of multi-class classifications.

Which libraries you will use?

- PyTorch
- re (regex)
- Math
- Pandas
- And more.

Summary of the model's performance on train and test datasets

Generally speaking, even though our architecture utilized the attention mechanism, it lacks fluency and informativeness. This is, as far as we believe, largely due to the scarcity of data which prevents neural networks from properly converging. We trained our network for 50 epochs and the training loss decreased to as low as 3.276. In further steps, we can investigate other representation techniques and approaches to enhance accuracy.