COMS 4995 Piyush Jena (**pj2400**)

MidTerm Assignment

Q1. Implement an encoder-only **transformer**[1] with the following specifications:

Non-Linearity	tanh
Embedding Size	64
Attention heads	4
Encoder layers	4

Your task is problem. We will use a standa into training set, testing set a hat for simplicity. In total, the ll consider just 4 classes for th

1. Noun: NN
2. Verb: VB
3. Adjective
4. Others: A

To reduce th ned word vectors from **word2** tors are also provided in csv

Initializati

Random Initialization in high dimensional spaces can lead to issues with convergence. This is why we will use **He-initialization**[4] for initialization of weights. Biases are to be initialized to 0.

Training:

While Stochastic Gradient Descent works, it requires 4x epochs as compared to Adam Optimizer. The bonus goal is to implement Adam Optimizer.

Result:

Report the accuracy, precision and recall for each of the classes.

References:

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems (2017).

[2] Erik F. Tjong Kim Sang and Fien De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition." In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 142– 147. (2003)

[3] Mikolov, n. "Efficient estimation of w arXiv:1301.3781 (2013).

[4] He, Kain n. "Delving deep into rectifier classification." In Proceedings vision, pp. 1026-1034. 2015.