Introduction to Deep Learning and Reinforcement Learning

Project Proposal

November 1st, 2019

# Group Members

Li, Yansong

Qu, Shuzheng

Su, Xuanyu

Yang, Siyuan

Linkletter, Maurice

# Statement of Problem

Our team has decided to compete in a superGLUE[1] task. The task we selected is **Choice Of Plausible Alternatives (COPA)** [2]. This task evaluates a model’s commonsense causal reasoning. It is tested with a series of questions where each question gives a premise and two plausible causes or effects. The correct choice is the alternative that is more plausible than the other.

# Approach

# To design our own pre-trained word representation model, we follow the unsupervised pre-training architecture [3] as below:

# ../Desktop/Screen%20Shot%202019-10-30%20at%202.05.05%20PM.png

# Where U is the context vector of tokens and n is the number of layers. We is the token embedding matrix, and Wp is the position embedding matrix. A transformer block is a set of transformers connected together. Transformer uses Multi-Head Attention, which means it computes attention h different times with different weight matrices and then concatenates the results together [3]. Then the output from the transformer block will be input into a linear layer and finally we use softmax to calculate the probability. This model will be used as the pre-trained word representation model.

# The key is the design of the transformer block.

# 

The transformer model has an encoder-decoder structure. The encoder maps an input sequence of symbol representations (x1; :::; xn) to a sequence of continuous representations z = (z1; :::; zn). Given z, the decoder then generates an output sequence (y1; :::; ym) of symbols one element at a time.

The Transformer follows this overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, shown in the left and right halves of the Figure above, respectively.[4]

**Encoder**: Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a position-wise fully connected feed-forward network. Each of the two sub-layers is followed by normalization layer. [4]

**Decoder**: In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack. The self-attention sub-layer is modified in the decoder stack to prevent positions from attending to subsequent positions. This masking, combined with fact that the output embeddings are offset by one position, ensures that the predictions for position i can depend only on the known outputs at positions less than i.[4]

# Goals and Objectives

[What are our goals for this project? Maybe list 3 or 4 goals. For example:]

* Obtain the highest possible superGLUE score using state-or-the-art models and pre-trained transfer learning.
* Gain a deep understanding of Natural Language Processing problem space by compete in superglue tasks.

# Action Plan

**Sept 28th:** Project group formed.

**Oct 5th:**  First group meeting to review project choices.

**Oct 12th:** Second group meeting. Project Selection. Initial self-directed student topics identified.

**Oct 26th:** The team meet to discuss 3 possible approaches for the COPA superGLUE task.

**Oct 26th - Oct 30th:** Project proposal created with additional research on the three approaches.

**Nov 1st:** Review project proposal with professor.

**Nov 1st – Nov 15th:** Implemented our model.

**Nov 16th – Nov 17th:** Our first submitting to the superGLUE online benchmark.

**Nov 18th – Dec 1st:** Refine model and continue to submit to superGLUE.

**Nov 25th – Dec 2nd:** Create project poster.

**Dec 2nd – Dec 7th:** Create project report.

**Dec 5th:** Poster presentation

**Dec 8th:** Report due date

# References

[1] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems, 2019. arXiv:1905.00537.

[2] Roemmele, M., Bejan, C., and Gordon, A. (2011) Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford University, March 21-23, 2011.

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018. URL https://blog.openai.com/language-unsupervised/

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010, 2017.