Mini-Project Proposal
Information Retrieval
Instructor: Venkatesh Vinayakarao
Chennai Mathematical Institute

Prateek Chandra Jha (MDS201922)
Saikrishna Ranganathan (MDS201928)

.

Submission Date: 06/11/2020

Introduction: Exploring the Cutting Edge Transformers in Google Search and Information Retrieval

As recently as 15th October 2020, Google published a blog on how Artificial Intelligence is powering a more helpful Google[1]. Mr. Raghavan writes:

> *At the heart of Google Search is our ability to understand your query and rank relevant results for that query. We've invested deeply in language understanding research, and last year we introduced how BERT language understanding systems are helping to deliver more relevant results in Google Search. Today we're excited to share that BERT is now used in almost every query in English, helping you get higher quality results for your questions.*

In 2018, Google introduced and open-sourced a neural network-based technique for natural language processing (NLP) pre-training called Bidirectional Encoder Representations from Transformers[2], or BERT, in short. This technology enables anyone to train their own state-of-the-art question answering system. This breakthrough was the result of Google's research on transformers: models that process words in relation to all the other words in a sentence, rather than one-by-one in order. BERT models can therefore consider the full context of a word by looking at the words that come before and after it—particularly useful for understanding the intent behind search queries. In our IR course[3] we have seen a flavour of Probabilistic Information retrieval but the Attention mechanism[4] employed by the transformer models takes it up a notch. ***I hope to explore the mathematics behind Attention mechanism, causal attention, bidirectional attention and multi-headed attention employed by Google Search's BERT models to explain the procedure of conducting a translation retrieval mechanism and text summarization mechanism via mini-code snippets***.

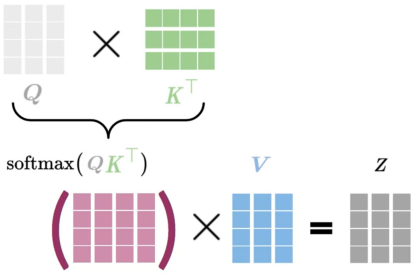

Figure 1: Dot-product Attention is the heart and soul of transformers. In general terms, the attention takes as inputs, queries, keys and values, which are matrices of embeddings. Dot-product Attention is composed by just two matrix multiplications and the softmax function. My project submission shall involve the usage of transformer models for tasks Google Search employs it for on a daily basis whether it is Autocompletion, Named Entity Recognition, Question-Answering, Neural Machine Translation, Text Classification and everyday Spell Checking.

[1] Prabhakar Raghavan (Senior Vice President, Search & Assistant, Geo, Ads, Commerce, Payments & NBU) writes on how AI is powering a more helpful Google. Available at: `https://blog.google/products/search/search-on/`

[2] Pandu Nayak (Google Fellow and Vice President, Google Search) writes on Understanding searches better than ever before using BERT. Available at: `https://blog.google/products/search/search-language-understanding-bert/`

[3] Available at: `http://vvtesh.co.in/teaching/IR-2020.html`

[4] Attention Is All You Need (Vaswani et al, 2017). Available at: `https://arxiv.org/abs/1706.03762`