BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

December 19, 2023

1 Summary

0. BERT is a new language representation model that is designed to pre-train deep bidirectional representations from unlabeled texts. The pre-trained model can be topped with one additional layer to create state-of-the-art models for a wide range of NLP tasks.

1. Existing pre-training networks only use one direction architecture while this paper proposes “masked language model” in which the objective is to predict masked words from the context. By doing this BERT achieved state-of-the-art performance in 11 NLP tasks.

2. **Background.** Feature-based and fine-tuning based approaches rooted in training word embedding vectors generalized to coarser granularities like sentences. Existing literature followed a uni-directional approach. The fine-tuning approach has an advantage that fewer parameters need to be learnt from scratch.

3. **BERT.** It has two parts - *pre-training* and *fine-tuning*. The first part is done over unlabeled tasks. For the second part, after initiliazing parameters from pre-training, the network is trained with labeled data from a specific task. Multi-layered transformer blocks are used in the model.

   Since conditioning on both directions allows each word to “see itself”, the paper proposes a simple idea. Mask words arbitrarily and let the model guess based on the context. To mitigate discrepancy between pre-training and fine-tuning masks are let to stay sometimes in fine-tuning.

   Next sentence prediction (NSP) is done simply replacing actual next sentence 50% of the time with a random sentence from a corpus. Pre-training process largely follows existing literature. In fine-tuning task specific inputs and outputs are fed and the network is trained end to end.

4. The paper presents results from the experiments. Then it compares the model with existing models using similar network size and shows that the results are state-of-the-art.

2 Analysis

This is a small but effective improvement over GPT model. Since it is more general, a better is expected. But surprisingly they do this with similar network size on all tasks considered. Perhaps the “bidirectional” should be replaced with “contextual”.