|  |  |
| --- | --- |
| Paper Title: | BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding |
| Paper DoI: | [10.18653/v1/N19-1423](http://dx.doi.org/10.18653/v1/N19-1423) |
| Research Question | Can BERT overcome the limitation of the standard language model of being unidirectional and produce a new state of the art? |
| Related Work | There are two strategies that are mostly used for applying pre-trained language models to downstream tasks viz feature-based and fine-tuning. Here to pre-train word embedding vector, left-to-right language modeling objectives have been used.  ELMo extract context-sensitive features from left-to-right and right-to-left language model. When the contextual words embedding is integrated with test-specific architecture, ELMo outperforms state-of-the-art for many NLP benchmarks.  Under unsupervised fine-tuning approach, the model has been pre-trained from unlabeled text and fine-tuned for a supervised downstream task which gives the advantage that only few parameters needs to be learned from scratch. Due to this advantage, OpenAI GPT has previously achieved start-of-the-art results. |
| Experiment Design | **Approach:**  The authors have introduced a new pre-trained model BERT. In this framework, there are two main steps: pre-training and fine-tuning.  For pre-training the model is trained on unlabeled data on several pre0training tasks.  For fine-tuning, it is first initialized with pre-trained parameters and all the parameters are fine-tuned using labeled data from the down stream task. BERT fine-tuning is then trained and tested on 11 NLPP tasks. |
| Result | BERT is a multi-layer bidirectional Transformer encoder. The authors report the results on the basis of two model sizes, BERTBASE and BERTLARGE. In terms of GLUE, both of these models outperformed all tasks over the previous state-of-the-art. Additionally, comparing between these two models, BERTLARGE  outperformsBERTBASE especially when the training data is smaller in size. When fine-tuning the SWAG dataset, it was found that BERTLARGE outperformed the baseline ESIM+ELMo system by 27.1% and OpenGPT by 8.6%. This concludes that bidirectional pre-training is important in language representations. It reduces the need for using many heavily-engineered task-specific architectures and use only one. |