Fine-Tuning Pre-Trained Compact Bert Models for Fake News Challenge Task

Tom Nguyen

December 2020

Abstract

BERT, a revolutionary pre-trained transformer model, has transformed the field of Natural Language Processing. It has been applied in multiple NLP down-stream tasks with great results. However, the compact models of BERT have been overlooked. In this paper, we experiment fine-tuning these simple compact BERT baseline models on the stance detection task with the data from the Fake News Challenge Stage 1 (FNC-1). The results have proven these compact models can be as effective as the standard Bert Base model.

Keyword: transformer; fine-tuning; fake news; compact model; stance detection.

Introduction

In the last two years, the landscape of Natural Language Processing has been transformed dramatically with the release of BERT, the first deeply bidirectional unsupervised language representation. It has inspired new complex language models and training methods such as TransformerXL[1], GPT-2, XLNet, and RoBERTa[2]. Surprisingly, most research papers on BERT were conducted with either of the two BERT models below:

- BERT Base (12 layers (transformer blocks), 12 attention heads, and 110 million parameters)
- BERT Large (24 layers (transformer blocks), 16 attention heads, and 340 million parameters)

Training these BERT models requires lots of computational resources. In March 2020, Google Research released 24 smaller Bert models (English only, uncased, and trained with WordPiece masking). The goal in this paper is to conduct experiments in fine-tuning these small BERT models with the stance detection task in the Fake News Challenge Stage 1. We hope the results from the experiments will stimulate more applications of these compact models in a wide range of linguistic tasks especially with the restricted computational resource environments.

Background

Four years ago, the Fake News Challenge Stage 1 (FNC-1)'s registration was open for research teams around the world to participate in and submit their work. The competition task is formulated as a stance detection problem. When we try to verify the truthfulness of a claim or statement, we need to gather all news texts about it and determine whether they are related to the claim. We need to find out whether the relevant background information contradicts or supports the claim. A good stance detection solution would allow us to enter a headline then pull out the top articles that agree or

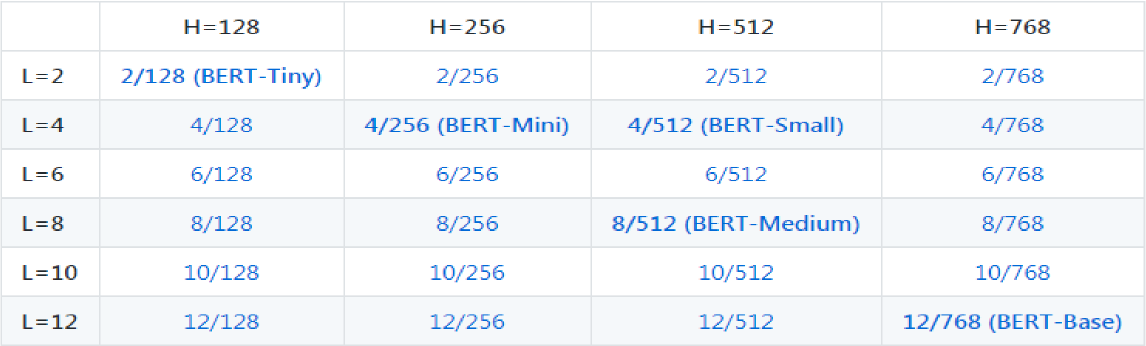
[1] Zihang Dai, Zhilin Yang "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context"
[2] Yinhan Liu, Myle Ott, Naman Goyal "RoBERTa: A Robustly Optimized BERT Pretraining Approach"

disagree. We can look at the arguments for and against the claim, then use our human judgment and reasoning skills to validate the claim.

Out of the 80 participants, 50 made submissions for FNC1 task using a wide array of techniques. The winner team was SOLAT in the SWEN with the highest relative score of 82.02. Since then, new NLP achievements such as BERT, RoBERTa and XLNet transformers helped to improve the result of the FNC-1 classification task. [3]

featMLP	***BERT***	***XLNet***	***RoBERTa***
78.59	*86.16*	*87.90*	*89.17*

However, these new transformers require intensive computational resources for pre-training and fine-tuning. A group of Google researchers have published 24 smaller BERT models [4]with the goal 'to enable research in institutions with fewer computational resources and encourage the community to seek directions in innovation alternative to increase model capacity'.



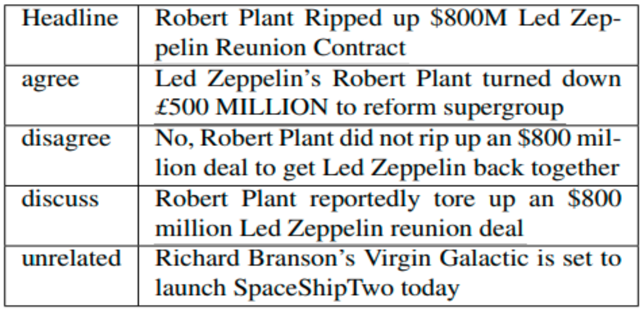
	H=128	H=256	H=512	H=768
L=2	2/128 (BERT-Tiny)	2/256	2/512	2/768
L=4	4/128	4/256 (BERT-Mini)	4/512 (BERT-Small)	4/768
L=6	6/128	6/256	6/512	6/768
L=8	8/128	8/256	8/512 (BERT-Medium)	8/768
L=10	10/128	10/256	10/512	10/768
L=12	12/128	12/256	12/512	12/768 (BERT-Base)

These new compact models were retrained under the same regime as the original BERT model and can be fine-tuned in the same manners.

Approach (Method)

Data

The Fake News Challenge 1 dataset consists of pairs of news article headlines and a text snippet from either the same article or others. Each pair in the training data is labeled with the relation between the headline and text snippet. The labels are "agree", "disagree", "discuss" and "unrelated" as can be seen in the example to the right.



Headline	Robert Plant Ripped up $800M Led Zeppelin Reunion Contract
agree	Led Zeppelin's Robert Plant turned down £500 MILLION to reform supergroup
disagree	No, Robert Plant did not rip up an $800 million deal to get Led Zeppelin back together
discuss	Robert Plant reportedly tore up an $800 million Led Zeppelin reunion deal
unrelated	Richard Branson's Virgin Galactic is set to launch SpaceShipTwo today

[3] Valeriya Slovikovskaya, Giuseppe Attardi "Transfer Learning from Transformers to Fake News Challenge Stance"
[4] Iulia Turc, Ming-Wei Chang, Kenton Lee, Kristina Toutanova "WELL-READ STUDENTS LEARN BETTER: ON THE IMPORTANCE OF PRE-TRAINING COMPACT MODELS

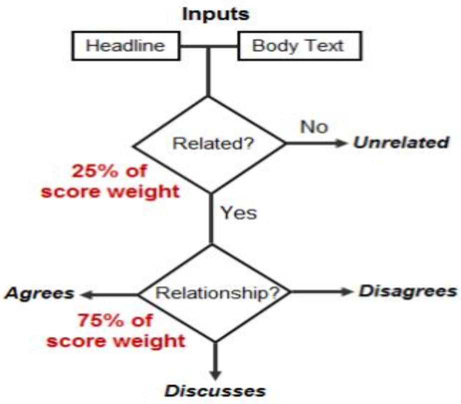
There are 1,683 articles and 49,972 stances in the training dataset in totality. We take notes that the stances are highly unbalanced.

Unrelated	***Agree***	***Disagree***	***Discuss***
36545	*3678*	*840*	*8909*

Baseline

These compact BERT models are used through a simple NLP pipeline. The steps include downloading the model zip files, setting up configuration, processing input data with a BERT tokenizer and an encoder to which is then loaded from model checkpoints and lastly using a classifier either for training or stance inference. We start our base model with the pre-trained Bert-Tiny which has 2 layers (transformer block) and 2 attention heads and 4.3 million trainable parameters.

Base Model Code: https://github.com/tom9ng/W266-Final/blob/main/Base%20Model%20-%20Final%20W266.ipynb



Since the FNC-1 dataset is highly unbalanced with respect to class distribution, we won't use F1 score as our evaluation metrics. Instead a hierarchical evaluation metric, FNC score, is introduced by the contest organizers with the goal to balance out the large number of unrelated instances. The new schema first gives out .25 points if the input article is correctly classified as "related" or "unrelated" to a given headline. If it is classified as "related", .75 extra points are awarded when the model correctly predicts the headline/article pair as "agree", "disagree", or "discuss".

Modeling

We use a "grid-search" approach to modeling Bert-Tiny and Bert- Mini by exploring different combinations of epochs (1,2,3,5,5 and batch sizes (16,32,64)) in order to fine-tune the stance detection task. We notice that good FNC-1 score (86.1) can be obtained with one epoch training on these tiny and mini models. We also experience Google Colab time-out while trying to fine-tune some larger BERT models with multiple epochs. As a result, we decided to apply the same "grid-search" method for all 24 Bert models with one epoch fine-tuning varying different batch sizes (16, 32, 64) to study whether we can achieve good FNC-1 score with only one epoch training.

Code: https://github.com/tom9ng/W266-Final/blob/main/Experimental%20Models%20-%20Final%20W266.ipynb

Features

Every BERT input includes the three following embeddings: Position, Segment and Token Embeddings.

- **Position Embeddings**: Express the position of words in a sentence.
- **Segment Embeddings**: BERT can also take sentence pairs as inputs for tasks.
- **Token Embeddings**: These are the embeddings learned for the specific token from the WordPiece token vocabulary.

Data Split

The training data is divided into two sets (80/20) for training and validation. These two sets will be used throughout the 24 Bert models training.

Results

The table below shows the results for the batch sizes that achieve the highest score FNC-1 (86.1) while fine-tuning 24 pre-trained BERT models.

Bert Model	***H=128 (A=2)***	***H=256 (A=4)***	***H=512 (A=8)***	***H=768 (A=12)***
L=2	*B=16,32*	*B=32*	*X*	*B=16,32*
L=4	*B=32*	*B=16,64*	*X*	*X*
L=6	*B=64*	*X*	*X*	*B=16*
L=8	*B=16,64*	*B=16*	*B=64*	*B=16*
L=10	*B=16,32,64*	*B=16*	*B=16,32,64*	*B=16,32*
L=12	*B=64*	*B=64*	*X*	*B=64*

Notes: L: layers; A: attention heads

Data sheet: https://github.com/tom9ng/W266-Final/blob/main/FNC-1-COMPACT-BERTS-EXPERIMENT-DATA.xlsx

Based on the results, the top FNC-1 scores can be achieved under one epoch training with 18/24 models. This can prove that BERT modes are well-known for being more sample-efficient than any other language models. These compact models can make the most out of every single example in the training data.

We noticed the Small Bert model (L=4, A=8) couldn't score high enough with even more epoch training given. Most of the failed experiments also occurred with compact models with 8 attention heads. This can be another research topic worth looking into.

Another advantage with one epoch training for these compact models is that we don't have to worry about over-fitting. Over-fitting happens when the model goes through multiple epoch training and it learns the details and noises in training data to the extent that it impacts the performance negatively.

We are also still interested in learning what the right size of the training dataset is for these compact models. However, due to time constraints on this project, we cannot perform experimentation on different training data set splits.

Conclusion

We conducted extensive experiments to learn the effectiveness of pre-trained compact Bert models on the stance detection task of Fake New Challenge 1. We found the small compact models can be as highly effective as the large ones. Moreover, these models could be trained with one epoch in order to achieve that high performance. The result encourages us to try fine-tuning and applying these compact models for other NLP tasks especially in the environment with the restricted computational resources.

REFERENCES

Iulia Turc, Ming-Wei Chang, Kenton Lee, Kristina Toutanova **“WELL-READ STUDENTS LEARN BETTER: ON THE IMPORTANCE OF PRE-TRAINING COMPACT MODELS”** < https://arxiv.org/pdf/1908.08962.pdf>

Valeriya Slovikovskaya, Giuseppe Attardi “**Transfer Learning from Transformers to Fake News Challenge Stance**” <https://www.aclweb.org/anthology/2020.lrec-1.152.pdf>

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, Kalina Bontcheva **“Stance Detection with Bidirectional Conditional Encoding”** <https://arxiv.org/pdf/1606.05464.pdf>

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova “**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”** < https://arxiv.org/pdf/1810.04805.pdf>

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, Ruslan Salakhutdinov **“Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context”** <https://arxiv.org/pdf/1901.02860.pdf >

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov **“RoBERTa: A Robustly Optimized BERT Pretraining Approach”** https://arxiv.org/pdf/1907.11692.pdf

Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf **“DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”** https://arxiv.org/abs/1910.01108