Annotated Bibliography

Brian Ferrell

October 2020

Topic 1

Comparing BERT against traditional machine learning text classification[1]

In this paper, the authors aim to compare BERT to traditional machine learning algorithms. The authors want to test and see if BERT should be the default when it comes to NLP tasks, and examine BERT's dominance over the older ways of handling NLP. The authors run four comparisons on four different datasets. The first experiment uses the famously known IMDB dataset to conduct a sentiment analysis, classifying movies to be either positive or negative. They compared the default pre-trained BERT to several well-known machine learning models. The second experiment compares BERT to H2OAutoML for a binary text classification task. This dataset(RealorNot tweets) contains tweets about real disasters and tweets that are not about real disasters(ex. use of metaphors). Since BERT continues to outperform their past experiments with the English language, they use a Portuguese news dataset classified into 9 classes for the third experiment and Chinese hotel reviews dataset for the fourth. Both compared BERT to AutoML module. Like this second source, this paper wishes to test how robust BERT really is.

Topic 2

Low-Shot Classification: A Comparison of Classical and Deep Transfer Machine Learning Approaches[2]

In this paper, the authors compare a variety of classical machine learning methods to BERT and ULMFiT using datasets in which the number of samples in each class is between 100-1000. The use of BERT and other transfer learning practices have demonstrated their strengths on much larger datasets, but these authors address the lack of quantitative studies on smaller datasets. They are doing this study because large datasets are too time consuming to obtain or expensive to compute, and real-world datasets, such as surveys or research protocols don't always have the luxury of accessing millions of samples. They focus their efforts on five datasets using sentiment-based classification tasks in two categories: Amazon reviews and Twitter. For the Amazon reviews they use three datasets containing movie reviews, books reviews, and health and personal care product reviews, and the other two datasets from Twitter are under subtask a and ce from SemEval2017. This paper resembles the first source, but more geared towards smaller datasets.

Topic 3

How to Fine-Tune BERT for Text Classification?[3]

This paper attempts to offer a wide-ranging solution for fine-tuning BERT on eight text classification datasets. They examine methods regarding the pre-processing of longer sequences of texts, layer selection, learning-rates for specific layers, low-shot learning, etc... The three strategies proposed by the authors are as follows: fine-tuning, further pre-training, and multitask fine-tuning. For the fine-tuning strategy, BERT contains multiple layers that have different levels of semantic and syntactic meaning which can be optimized for better results. Regarding the further pre-training strategy, BERT contains a general domain knowledge which can be further pre-trained to fit a specific domain. As for the multitask fine-tuning strategy, BERT can be fine-tuned in a multi-task learning framework. Using seven English datasets and one Chinese they investigate the strategies mentioned above using BERT base, uncased BERT, and Chinese BERT base. These datasets vary in number of documents and lengths of documents for sentiment analysis, question classification, and topic classification. This paper examines the parameters and hyperparameters of BERT, whereas in the first two sources they just use BERT to compare on different tasks.

Topic 4

Publicly Available Clinical BERT Embeddings[4]

BERT is pre-trained on BooksCorpus and Wikipedia, which in general can model language well for any NLP task; however, these authors examine ways to improve that general language model in BERT with BERT models that are geared for clinical text and discharge summaries. They demonstrate that performance is improved with domain specific pre-training which is very distinct from general language. The data is used for pre-training from the MIMIC-III database in two ways: Clinical BERT(contains all note types), and Discharge Summary BERT(only contains discharge summaries). The purpose of this is to further downstream the tasks with the clinical data that can be used for even more specific classification problems. They then train two BERT models on the clinical text, where one is initialized from the BERT-base model, and the other is initialized from BioBERT. Their work includes results from the models used on various clinical NLP tasks, and evaluations on the differences between the models. Like the first two sources, it compares on various NLP tasks but for more specific tasks.

Topic 5

ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations[5]

The authors introduce two parameter reduction techniques to lower GPU/TPU memory usage while increasing the training speed of BERT. Their best resulting model was evaluated on the GLUE, RACE, and SQuAD benchmarks against BERT-large. ALBERT is designed to distribute BERT's capacity more efficiently and make finer grained distinctions by reducing its parameters and introducing sentence-order predictions. The parameter reduction is done by splitting the embedding matrix into two smaller matrices and projecting the one-hot vectors to a lower dimensionality followed by projecting the lower dimensional embedding space into the hidden space. They then propose a parameter-sharing strategy within the stacked layers that are all identical weights, unlike BERT. These parameter reduction strategies alone actually hurt accuracy, but they are meant to be scaled up to what is needed to improve(i.e. the larger sized ALBERT models). The model's advantage is the reduction of parameters, which in some cases is not only faster than

BERT but has better benchmark accuracy. This resembles the fourth cited source in that it takes BERT to a whole other level to make it better.

Topic 6

Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment[6]

The authors in this paper introduce an interesting way to test BERT's performance by creating adversarial sentences. This is done in a simplified way that is applied to two NLP tasks, forcing the model to make wrong predictions. The purpose of doing this is to see how to confuse BERT by replacing the words with the most influence in the final prediction with a semantically similar word, forcing BERT to get it wrong. Their baseline TextFooler algorithm is compared to three deep learning models over five text-classification datasets, and two textual-entailment tasks, and they evaluate the word replacement strategy in two ways(an automated and human form to judge the adversarial sentence). This is noted to be important because it exposes weaknesses in BERT on tasks it should be getting correct, which is shown in the evaluation metrics perturbation and success rates. In two other sources included in this paper, I go over two other ways to create adversarial sentences that are better than the simple TextFooler.

Topic 7

BAE: BERT-based Adversarial Examples for Text Classification[7]

The authors in this paper also introduce a way to test the performance of BERT with adversarial examples, but instead they leverage the power of language models to generate the alternatives. The difference between this one and TextFooler(mentioned previously) is that TextFooler uses rule-based synonym replacement from fixed word embedding space. This simplistic form of adversarial attacks only considers the single word(token) in the sentence rather than the overall semantics and context of how the word is being used. The authors introduce four strategies with BAE(BERT based Adversarial Examples), which uses masked language models. The four strategies are as follows: replacing the token, inserting a token to the left or right of a particular word, choosing to replace or insert but not both, and replacing the token first followed by inserting a token to the left or right. These strategies show that with just a few perturbations of replacing words or inserting them, it can reduce the accuracy significantly compared to TextFooler(baseline model). The adversarial examples are more natural looking examples; considering, they make use of the semantics learned from the language model. Additionally, these experiments were carried out on seven different text classification datasets and are used on three models(word-LSTM, BERT, and word-CNN). The reason for creating such robust ways of forcing models to misclassify inputs, is to expose where powerful models happen to be vulnerable.

Topic 8

BERT-ATTACK: Adversarial Attack Against BERT Using BERT[8]

Creating adversarial examples is becoming more popular, and to compare different styles of testing the robustness of BERT, the authors in this paper present BERT-Attack, which uses the BERT language model against itself to mislead the model in making incorrect predictions, just like in the previous two papers. Both the paper above and this one came out around the same time so I thought it would be interesting to compare them on similar tasks. BERT-Attack uses its language model by finding vulnerable words in each input sentence, then applying BERT to generate substitutes for it. They define “most vulnerable words” by words that have the most impact on the model’s predictions. The approaches in these two papers are very similar and I wish to hopefully summarize them in ways that show their differences. I think to do that, I will plainly describe what each does, then allow the reader to make the connections on how they are different. From what I understand, the true difference between the two is that BERT-attack has what is called sub-word replacement strategy and it does not place words to the left or right of a token, I.e. just substitutes words.

Topic 9

The Utility of General Domain Transfer Learning for Medical Language Tasks[9]

The authors of this paper wish to leverage the use of a medical domain corpus combined with a BERT model on a radiological multi-label text classification problem. They point out the challenges when it comes to electronic health records, and how they are hard to read, hard to classify, hard to do research on, and can be inaccessible sources due to high proportions of the data being unstructured. It is also a very time-consuming task to read and analyze such reports. To help in the automation of classifying and segmenting the textual electronic health report data, the authors randomly sampled 1,977 CT scanner reports(out of 97K) from the hospital’s archiving and communication database system, which were hand-labeled by a team of physicians and neurosurgical/radiology residents. The hand-classified labels are an independent binary multi-label dataset, which means there are multiple labels that can correspond to a single CT report. There can be a 0 or a 1 in front of these 13 labels for every report: normal, hemorrhage, stroke/infarction/ischemia (infarction, venous thrombosis), vascular abnormality (aneurysm, arteriovenous malformation), chronic small vessel disease, periventricular white matter changes, ventricular abnormalities (i.e. hydrocephalus), atrophy (brain), bone abnormality, bony sinus disease, foreign objects, maxillary sinus disease, and carotid siphon calcification. They compare the regular BERT to BioBERT, and a few other models on precision and recall, also included is the ROC score per label.

Topic 10

Comparing deep learning architectures for sentiment analysis on drug reviews[10]

The authors of this paper present comparisons on different advanced NLP models on sentiment analysis tasks for drug reviews. The models used for this task are the following: LSTM, CNN, SVM, Hybrid models, and BERT+LSTM. They use a dataset taken from Drugs.com, which comprises drug reviews with their corresponding score(reflects patient’s degree of satisfaction with the drug). They do this task in two ways: Classify ratings based on a 10-point scale, and then reduce the ten-point scale to a three-point scale by combining the ratings into three levels of polarity(negative, positive, and neutral). The 3-point scale task improves performance for obvious reasons. For models other than BERT they use different approaches for their word-embeddings: one was trained on PubMed, PMC, and Wikipedia, and the other was trained on tweets about drugs. Results for both styles are shown in the paper as well. Results show the difference between the 3-point and 10-point scale using the F1 evaluation score

Topic 11

To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection[11]

The authors of this paper compare the results of hand-crafted feature-based models versus a fine-tuned BERT model for detecting Alzheimer's Disease. There are two tasks within this paper; however, I will just focus on the binary text classification one. The author's motivation behind this task is to find a less time-consuming way to classify these texts, due to the process that comes with feature engineering can be quite long and it requires expert domain knowledge. There is also the risk of missing features in the data, as humans tend to be limited in finding these characteristics that best inform the predictions. They use the ADReSS Challenge dataset which has 156 speech samples that are split 50/50 for either non-Alzheimer's or Alzheimer's participants. They manually extract 509 key features from the transcripts for the following machine learning models: SVM, Neural Network, Random Forest, and Naive Bayes. All the models(including BERT) are trained using 10-fold cross validation as the amount of data they have is quite small, and another cross-validation strategy(not for BERT, because of memory constraints) called leave-one-subject-out CV. Results shown are of the training and testing data, showing BERT outperforming the ML models on all metrics. In the literature review, I wish to go into more in depth on the data they are classifying and what exactly is inside Alzheimer's Disease transcript data.

Topic 12

Antisocial online behavior detection using deep learning[12]

This paper shows how Deep learning models and BERT are being used overseas in places like Germany for malicious online behavior detection and how people outside of America view speech and protecting social welfare. They focus their work on harassment or threatening situations in an online communication setting over social media platform datasets such as Twitter, Facebook, Wikipedia, and Formspring. They have a wide variety of models that they compare using stratified 5-fold cross-validation such as traditional machine learning models, fully connected neural networks, RNN's, LSTM, GRU's, Bi-LSTM/GRU, transformers, attention models, and more. DistilBERT and BERT do better on the number of binary classification tasks they had for each dataset predicting whether something was malicious or not. I think this task is very interesting, because how do we draw the line between what is malicious intent versus what is sarcastic, and even if we can differentiate between what is considered malicious(which I am sure we can), what do we do with that kind of power when it comes to people's opinion varying. It also seems that having a perfect dataset for their task doesn't seem to exist now, as they point out the unintended biases in their paper. These biases occur from different ways the datasets are retrieved, and it leads to discriminating against people or groups differently in unfair ways, like the use of gender, religion, or sexual orientation apparently can have higher "malicious scores" even if the comment is not intended to be hateful.

References

1. Santiago González-Carvajal and Eduardo C Garrido-Merchán. Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*, 2020.

2. Peter Usherwood and Steven Smit. Low-shot classification: A comparison of classical and deep transfer machine learning approaches. *arXiv preprint arXiv:1907.07543*, 2019.

3. Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019.

4. Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

5. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

6. Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *arXiv*, pages arXiv–1907, 2019.

7. Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970*, 2020.

8. Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*, 2020.

9. Daniel Ranti, Katie Hanss, Shan Zhao, Varun Arvind, Joseph Titano, Anthony Costa, and Eric Oermann. The utility of general domain transfer learning for medical language tasks. *arXiv preprint arXiv:2002.06670*, 2020.

10. Cristóbal Colón-Ruiz and Isabel Segura-Bedmar. Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics*, 110:103539, 2020.

11. Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2008.01551*, 2020.

12. Elizaveta Zinovyeva, Wolfgang Karl Härdle, and Stefan Lessmann. Antisocial online behavior detection using deep learning. *Decision Support Systems*, page 113362, 2020.