John Dukewich
7 February 2020

Executive Summary: Using BERT for Text Representation in Machine Learning

Machine learning is the process of collecting data and using a computer to mathematically generalize the data to make predictions on new, unseen data examples. A subcategory of machine learning is natural language processing (NLP), which aims to utilize machine learning methods to analyze, predict, and classify text. NLP is utilized by large technology companies for various purposes, such as categorizing reviews of products. However, one of the most difficult parts of natural language problems is how to represent text mathematically so a computer can perform analyses (Yu, Su, & Luo, 2019). Various current methods for creating machine representations of text utilize variations of neural network models. These models are computationally expensive to initially train, costing companies excessive amounts of money and time (Sun, Qiu, Xu, & Huang, 2019). Recent advances have shown a method not using neural networks, but instead using a transformer model to more effectively capture essential context between words within a text.

This report evaluates the effectiveness of a Bidirectional Encoder Representations from Transformers model (BERT) to determine how NLP problems can improve accuracy while lowering temporal and monetary costs. Much of the research on BERT has only been done in 2018 and 2019, but its proven success deems it necessary for technology companies to start using BERT on any attempts to solve NLP problems.

BERT has proven effective in various types of NLP problems, such as sentiment analysis, text classification, and text generation. All NLP problems will initially require a mathematical representation of text, and BERT creates this representation using an attention mechanism that captures the context both before and after a word (Gao, Feng, Song, & Wu, 2019). BERT captures long term word dependencies and contexts that neural networks fail to recognize, allowing it to represent text with finer detail and granularity (Sun et al., 2019). Although BERT solves the text representation problem well, it does not actually complete a given NLP task successfully without fine-tuning. BERT can be fine-tuned to improve accuracy for a single task and for multiple tasks within a single domain of knowledge (Yu et al., 2019). Generalized BERT models trained on large datasets work well for initial text-to-vector embeddings, but augmenting BERT with neural networks improves the accuracy for text classification tasks (Li et al., 2019). Likewise, additional training data from the problem domain enhances the effectiveness of BERT while removing some of its generality (Yu et al., 2019). Overall, BERT performs better than current neural networks in the first step of representing text. When paired with fine-tuning and other neural models, BERT can boost accuracy of tasks significantly.

These performance improvements clearly indicate that BERT should be used for any new NLP problems that arise, but BERT should also replace current NLP solutions. Technology companies, such as Amazon, Apple, Microsoft, and many others, should adopt BERT models to save computational costs as well as improve the accuracy of their NLP services. Due to the novelty of BERT, many NLP solutions most likely use some variant of neural networks to create

text embeddings, so this transition to BERT will require plenty of effort, but the rewards far outweigh the costs. Therefore, NLP problems should be solved using BERT to represent the text and by combining other neural models into the task specific implementations.

Gao, Z., Feng A., Song X., & Wu X. (2019). Target-Dependent Sentiment Classification With BERT. *IEEE Access*, 7, 154290-154299. https://doi.org/10.1109/ACCESS.2019.2946594

Sentiment analysis is the application of machine learning models to analyze text in various media and determine its overall sentiment (e.g. positive, negative, neutral, conflicting, etc.). Traditional methods involve using various forms of neural networks coupled with feature embeddings. Neural network architectures have continuously improved over time, moving from conventional neural networks to convolutional and eventually Long Short-Term Memory units. These models pose two types of issues in training: memory and computational limits. Neural networks require massive amounts of data to train, often storing vast amounts of "empty" information (i.e. zero values). These models boast heavy computation requirements, resulting in extremely lengthy training times. Using a rather new architecture, Bidirectional Encoder Representations from Transformers (BERT), has already proven superior. BERT works drastically different from traditional neural networks via its attention mechanism. Attention computation allows the model to effectively capture the relevant context for each word in a text, even linking long-term dependencies. This study directly compares the accuracy of BERT against common algorithms on multiple datasets including Twitter, restaurant reviews, and laptop reviews. BERT performed at a 5% to 10% higher accuracy than current state-of-the-art models. This shows that BERT can achieve sufficient accuracy on sentiment analysis tasks, effectively capturing the opinion and intents of user generated text. BERT outperforms the common neural network models in sentiment analysis tasks by a statistically significant margin. The high accuracy results from BERT tests have proven it to be the new baseline of sentiment analysis, allowing real world application to more effectively model user sentiment.

Li, W., Gao, S., Zhou, H., Huang, Z., Zhang, K., & Li, W. (2019). The Automatic Text Classification Method Based on BERT and Feature Union. Proceedings from ICPADS 2019: *IEEE 25th International Conference on Parallel and Distributed Systems*. Tianjin, China. https://doi.org/10.1109/ICPADS47876.2019.00114

In natural language text classification problems, text inputs are mapped to mathematical vectors which then run through another model to reduce the vectors to a single category, yielding the desired classification. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) cells are the most common ways to generate embeddings, but these models incur heavy costs. CNNs, RNNs, and LSTM can all create the text embedding as well as classify the resulting vector. RNNs show moderate success in generating useful text embeddings, and CNNs perform the classification aspect well, reducing the embedding vector to a single category from a predetermined set. These previous methods are computationally expensive and take long amounts of time to both train and run classification tasks. Furthermore, RNNs fail to capture long term text dependencies and can thus lose essential context when trying to embed large text examples. This study proposes a new system in which a Bidirectional Encoder Representations from Transformers (BERT) model augments current methods by overhauling the text embedding generation step. Instead of RNNs, BERT is used to generate the text embedding mappings. This study then attempts to improve the classification step by combining the CNN classification model with an LSTM model. The performance of various combinations of BERT with CNNs and LSTMs were tested on a dataset of text examples from a job recruitment website. The experimentation indicates that using BERT for the text embedding step far outperforms any methods utilizing RNNs or CNNs to generate embeddings. Another essential finding shows that BERT performs more accurately when combined with both a CNN and LSTM model, achieving up to 96% accuracy over the 92% accuracy when used alone. BERT can enhance current strategies of text classification by altering the upstream text embedding generation. Using BERT over RNNs and LSTM to generate embeddings reduces computational costs as well, saving time and money.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification?. Proceedings from CCL 2019: *Chinese Computational Linguistics 2019: Lecture Notes in Computer Science*. Kunming, China. https://doi.org/10.1007/978-3-030-32381-3_16

In natural language processing, text classification aims to assign a given sequence of text to a member of a set of predefined categories. The Bidirectional Encoder Representations from Transformers (BERT) model has proven effective in learning mathematical language representations (i.e. embeddings), but there has been little experimentation with its use in classification tasks. Previous work indicates the success of using pre-trained word embeddings over embeddings learned from scratch, regardless of the model used to learn the embeddings. Recent breakthroughs with BERT architecture have improved the success of creating text-to-vector embeddings. However, those studies fail to address how BERT can be fine-tuned for specific applications and problems. Most natural language processing problems essentially have two steps: learn how to represent text as a mathematical vector and then train the vector representation for a specific use (e.g. classification, sentiment analysis). Previous research focused on the first step of creating the embeddings with little work showing how BERT may improve the second step of training for a specific use. This study experiments with various fine-tuning strategies for BERT to see how these variations affect BERT's effectiveness in a specific task—text classification. Various hyperparameters such as learning rate, decay factor, hidden size, and number of transformer blocks were altered and analyzed for their effects on accuracy of classifying eight different datasets. These experiments concluded that a layer-wise decreasing learning rate assists BERT in overcoming the catastrophic forgetting problem, in which training on application specific data caused it to "forget" information learned. The results of testing also indicate that pre-training BERT with additional data both in-domain and within-task offer significant boosts in accuracy of the model, while multi-task fine-tuning offers little to no improvement. Therefore, the generalized BERT model can be fine-tuned for specific tasks such as text classification. These fine-tuning strategies allow BERT models to reach accuracies that are acceptable for real-world use cases like classifying Yelp reviews by category.

Yu, S., Su, J., & Luo, D. (2019). Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge. *IEEE Access*, 7, 176600-176612. https://doi.org/10.1109/ACCESS.2019.2953990

Utilizing Bidirectional Encoder Representations from Transformers (BERT) models in natural language processing tasks has proven more effective that traditional neural networks, but BERT still struggles with domain-specific knowledge and training on datasets with short text entries. BERT's generalized pre-training misrepresents contexts of words with multiple meanings. Recent advances with BERT have already proven superiority over feature-based models for general natural language tasks. Some research also shows how to fine-tune BERT for text classification problems. BERT has been proven successful in categorizing datasets with text examples containing many words. Previous studies do not address issues BERT faces with datasets that contain text with few words (i.e. less than about 500 words per example). Little investigation has been done with pre-training BERT with domain-specific knowledge, but these studies fail to describe how to pre-train BERT for a domain that does not have a readily available dataset. This study aims to improve how BERT can be used for short-text datasets in the task of text classification. The researchers propose an altered BERT model that can construct auxiliary sentences from an input sentence. These auxiliary sentences will emulate longer input texts which are then the inputs to BERT classification models. The experiments demonstrate that the BERT model with auxiliary sentence construction performs better for classification of short texts than other feature-based and fine-tuning methods proposed elsewhere, but it offers no improvements for domain-specific knowledge. The auxiliary approach was more accurate than a regular fine-tuned BERT model in both binary classification and multi-class classification applications. Utilizing auxiliary sentence generation grants BERT more flexibility in training datasets as well as better performance in all types of classification tasks. Strong performance on various types of datasets make this application of BERT more effective computationally than older neural models and generalized BERT models without fine-tuning.

References

Gao, Z., Feng A., Song X., & Wu X. (2019). Target-Dependent Sentiment Classification With BERT. *IEEE Access*, 7, 154290-154299. https://doi.org/10.1109/ACCESS.2019.2946594

Li, W., Gao, S., Zhou, H., Huang, Z., Zhang, K., & Li, W. (2019). The Automatic Text Classification Method Based on BERT and Feature Union. Proceedings from ICPADS 2019: *IEEE 25th International Conference on Parallel and Distributed Systems*. Tianjin, China. https://doi.org/10.1109/ICPADS47876.2019.00114

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification?. Proceedings from CCL 2019: *Chinese Computational Linguistics 2019: Lecture Notes in Computer Science*. Kunming, China. https://doi.org/10.1007/978-3-030-32381-3_16

Yu, S., Su, J., & Luo, D. (2019). Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge. *IEEE Access*, 7, 176600-176612. https://doi.org/10.1109/ACCESS.2019.2953990