**Project Proposal**

W266 Spring 2020

Joanna Yu

Mark Butler, Friday 4-5:30pm

While there has been rapid development in the field of NLP in the last decade, the scarcity of labeled data remains a problem and researchers are looking into ways to make use of unlabeled data. Research has shown that unlabeled data can improve adversarial robustness.[2] Back-translation[3,4] has also been used as a way to augment the data. This project will focus on expanding the recent work done by Google Brain on Unsupervised Data Augmentation (UDA) [1] to perform Semi-Supervised Learning. The paper [1] shows that UDA improves performance for both image and text classification. The base model used in the paper is a pre-trained BERT[5] model fine-tuned with IMDb dataset and augmented using Amazon-2, Amazon-5, Yelp-2, and Yelp-5. This project proposes to use the model from the paper and investigate UDA performance improvement in the following areas with additional datasets:

1. Model performance with respect to the proportion of labeled vs unlabeled data, ie. how much unlabeled data may potentially make up for the lack of labeled data? This project will analyze the effect of varying the amount of unlabeled data while keeping labeled data constant. Since this experiment is designed to complement the experiment done in the paper, I will use the same datasets as the paper. I have not downloaded and run the model yet.
2. Determine how domain relevance of unlabeled data can affect this Semi-Supervised model’s performance. The base model will be a BERTbase model finetuned by the labeled US Airline Sentiment dataset from Kaggle. The Yelp or Amazon dataset will serve as a slightly out-of-domain comparison. And lastly, the Enron Email dataset (or Wikipedia) will serve as a completely out-of-domain comparison.

Proposed datasets:

* Twitter US Airline Sentiment from Kaggle
* Amazon and Yelp datasets
* Enron Email Dataset (longer text and may not perform well with BERT)
* Wikipedia Dataset

References:

1. Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised Data Augmentation for Consistency Training. *arXiv preprint* [*arXiv:1904.12848v4*](https://arxiv.org/abs/1904.12848v4). September, 2019
2. Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *arXiv preprint* [*arXiv:1905.13736*](https://arxiv.org/pdf/1905.13736.pdf), 2019.
3. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint* [*arXiv:1804.09541*](https://arxiv.org/pdf/1804.09541.pdf), 2018.
4. Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint* [*arXiv:1511.06709*](https://arxiv.org/pdf/1511.06709.pdf), 2015.
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* [*arXiv:1810.04805*](https://www.aclweb.org/anthology/N19-1423.pdf), 2018.