Unsupervised Data Augmentation Experimentation

Joanna Yu
UC Berkeley
joannayu@ischool.berkeley.edu
Spring 2020

Abstract

While there has been rapid development in the field of Natural Language Processing in the last decade, the scarcity of labeled data remains a problem and researchers are looking for better ways to make use of the abundant unlabeled data for semi-supervised learning. Research has shown that unlabeled data can improve adversarial robustness and consistency training is among those that show great promises. The recent work done by Google, titled "Unsupervised Data Augmentation for Consistency Training" (UDA) [1], shows that back-translation, as an advanced data augmentation technique, is an effective way to improve model performance. This work focuses on expanding the framework from the UDA paper to further investigate the tradeoff between labeled and unlabeled data and the role of domain relevance in semi-supervised learning using unsupervised data augmentation. Experiment results show that large amount of unlabeled data can in fact compensate for the lack of labeled data and that domain relevance actually plays a rather small role on model performance in this framework.

1 Introduction

Even though there is an explosive amount of data being generated by the minute from many sources, very little can be leveraged for machine learning purposes. Most of the deep learning architectures rely on labeled data, which is scarce and requires a lot of time and money to compile. This gives rise to the research around semi-supervised learning, where supervised training can benefit from fine-tuning by unlabeled data.

The idea behind consistency training [2] is to regularize model predictions to increase model robustness. Different types of noise have been added at different points in order to achieve this purpose. A few examples of noise injection methods include additive Gaussian noise, dropout noise, and adversarial noise.

- **Gaussian noise:** Statistical noise having a probability density function equal to that of the normal distribution, which is also known as the Gaussian distribution.
- **Dropout noise:** Noise injection to hidden units as a way of regularizing a neural network.
- **Adversarial noise:** A small amount of carefully constructed noise is added with the goal of causing the neural network to misclassify the label. Research has shown that unlabeled data can improve adversarial robustness [3].

The UDA paper establishes that using advanced data augmentation methods to perform consistency training is a better source of noise in semi-supervised learning. UDA is shown to outperform purely supervised model that uses orders of magnitude more labeled data. Specifically, in the IMDb movie review dataset, with only 20 labeled examples, UDA outperforms the state-of-the-art purely supervised model trained on 25,000 labeled examples. UDA achieves similar success on both text as well as image tasks.

This work aims to extend the investigation done by the work in the UDA paper in the following areas:

1. Track model performance with respect to the proportion of labeled vs. unlabeled data, i.e. how much unlabeled data may potentially compensate for the lack of labeled data? The experiments in the UDA paper take advantage of the entire set of augmented unlabeled data while varying the amount of labeled data. This work will analyze the effect of varying the amount of augmented unlabeled data while keeping amount of labeled data constant. Since this experiment is designed to complement the experiments done in the UDA paper, the same IMDb movie review dataset will be used.
2. Determine how domain relevance of unlabeled data can affect model performance. One natural question about semi-supervised learning is what type of unlabeled data can maximize model performance and how much domain relevance plays a role. As completely in-domain unlabeled data may not be readily available, it is useful to know whether semi-in-domain and out-of-domain augmented unlabeled data can fill in the gap. This work will investigate the role of domain relevance by using four additional datasets from Amazon, Twitter, and Kaggle.

2 Background and Methods

2.1 Data Augmentation and Back-translation

The purpose of data augmentation is to increase and enrich the training data by performing various transformations to the original training data. The UDA paper uses back-translation [3, 4, 5] as a way to produce a rich and diverse set of paraphrases from the original dataset. Back-translation is the process of translating a data sample from language A to B and then back to A again. The UDA paper finds that diversity of the paraphrases provides more benefit than quality or validity. In this work, to be consistent with the work done in the UDA experiments, the WMT'14 English-French translation model is used in both directions to perform back-translation on each sentence. This is a sentence-level back-translation, which can maintain the overall semantics of the sentence. One other alternative is to use TF-IDF to perform word-level back-translation, which has the advantage of controlling the model to replace uninformative words with low TF-IDF scores.

2.2 Base Model

In the UDA paper, the model achieves excellent performance in the binary sentiment analysis task using BERT with only 20 supervised data examples. This serves as the baseline model comparison for the additional experiments this work will perform. Due to resource constraints, $\text{BERT}_{\text{base}}$ will be used. The $\text{BERT}_{\text{base}}$ uncased model contains 12 layers, 12 attention heads, 768 hidden layers, and 110 million model parameters [7]. Since both the labeled and unlabeled data come from the IMDb dataset, this semi-supervised learning is considered completely in-domain and serves as a good starting point for domain relevance exploration as well.

2.3 About the Data

	Positive	**Negative**	**Total**
Labeled Training Data	12,500	12,500	25,000
Test Data	-	-	25,000
Unlabeled Training Data	-	-	50,000

Table 1: Data distribution of the IMDb supervised training set.

The main dataset, IMDb movie review dataset, is an ideal dataset for the proposed experiments since it contains a good amount of labeled and unlabeled examples. In addition, using a movie review dataset allows for the possibility of appending additional movie review data if one wishes to experiment with data beyond the size of the IMDb.

Additional datasets, as shown in Table 2, are selected for the domain relevance experiments. The Amazon movie and TV review dataset is considered in-domain with IMDb. The Amazon office product review and Twitter airline sentiment datasets are considered semi-in-domain since, even though the reviews are not related to the movie industry, they are still sentiment expressions about products and services. Finally, the Kaggle natural disaster dataset serves as an out-of-domain comparison to the IMDb. Since tweets are often very short, the Twitter and Kaggle datasets can also serve as a good contrast against the Amazon datasets in terms of text length and model performance.

	Domain Relevance
IMDb Movie Reviews	In-Domain
Amazon Movie and TV Reviews	In-Domain
Amazon Office Product Reviews	Semi-in-Domain
Twitter Airline Sentiment	Semi-in-Domain
Kaggle Natural Disaster Tweets	Out-of-Domain

Table 2: Dataset selected for the domain relevance experiments and their corresponding domain relevance to the supervised training set (IMDb).

3 Experiments

The experimentation performed in this work is leveraged from the work done in the UDA paper while adapting the framework to additional datasets. A series of notebooks and scripts are run on Google Colaboratory Pro using GPU/TPU environment. A large amount of Google Cloud Storage is used for this project due to the size of BERT and model checkpoint files[1]. Depending on the total size of data being fine-tuned and trained, a single model takes anywhere from 30 minutes to over 6 hours to run.

The metrics used in model evaluation is the error rate, as consistent with the UDA paper. All models are run for at least 10 epochs by adjusting the training steps given the amount of labeled and unlabeled data being used. Due to resource constraints, hyper-parameters are tuned to be consistent with the use of $\text{BERT}_{\text{base}}$. In particular, the sequence length has been reduced from 512 to 128 due to memory constraints. Such reduction in sequence length is expected to degrade model performance since sentences with more than 128 tokens will be truncated and the model cannot fine-tune on text longer than that length. The error rate of the baseline model using 20 labeled examples and the full augmented unlabeled dataset is 8.2%.[2]

For back-translation, the hyper-parameter *sampling_temp* is used to control the balance between diversity and quality of the paraphrasing. Experiments from the UDA paper show that keeping it at 0.9 leads to optimal results for this dataset. In addition, consistent with the UDA paper, data examples under 500 words are eliminated, which reduces the actual total unlabeled examples from 75,000 to 69,972. Additional restriction is placed on back-translation where translation

[1] Model checkpoint files are over 5GB per model for each of the 70+ models ran in the experiments.

[2] The best performing model from the UDA paper uses max sequence length of 512 on $\text{BERT}_{\text{LARGE}}$ to achieve an error rate of 4.20%.

would not be performed if the length of the resulting paraphrase differs significantly from that of the original sentence.

For the exploration of the tradeoff between labeled and augmented unlabeled data, seven different amounts[3] of labeled data are selected. For each of them, model performances are tracked for every increase of 4,000 augmented unlabeled examples from 0 to 16,000, at which point the error rate begins to level off so additional models are run at 24,000, 48,000, and the full dataset of 69,972 examples. Back-translation is performed on the unlabeled data for each experiment and the result is then passed to the model for fine-tuning and training.

The experiments for domain relevance follow similar procedure. Four different amounts[4] of labeled data are selected. The amount of unlabeled data is kept constant at 16,000 examples[5] so the results are comparable across all datasets. When feasible, only examples over 128 tokens are selected so training can be done on longer text. Back-translation is performed to augment the unlabeled data. The previously explained 500-word minimum requirement for back-translation is removed since two of the datasets are tweets, which are often very short. All other hyper-parameters are unchanged.

3.1 Results

As expected, as more labeled data is used, the error rate begins at a better starting point (Figure 1). Increasing the amount of augmented unlabeled data helps model performance, especially during the first 10,000 unlabeled examples. The model sees the most benefit from the use of augmented unlabeled data when the amount of labeled data is extremely low (20 examples). In contrast, the improvement in model performance is marginal when there is ample amount of labeled data (10,000 data examples). When the full unlabeled dataset is used, all seven models converge to a similar error rate, in the range of 0.073 to 0.084. It is unclear whether the models will converge further to an even smaller range of error rate if there are additional augmented unlabeled data.

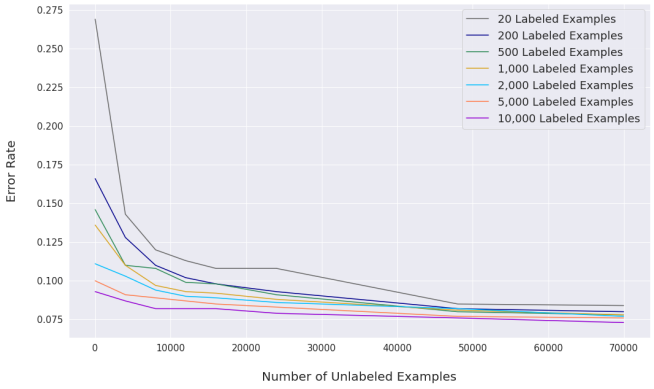

Figure 1. Error Rate Comparison

3.2 Relationship between Labeled and Unlabeled Data

When the proportion of labeled and augmented unlabeled data are plotted against each other, one can clearly see that fine-tuning with large amount of augmented unlabeled data does in fact compensate for the lack of labeled data. For example, a 0.09 error rate can be achieved with two options: 1) a combination of 10,000 labeled example and 4,000 augmented unlabeled examples; 2) a combination of 200 labeled examples and 69,972 (the full augmented unlabeled dataset). The

[3] To investigate the tradeoff between labeled and augmented unlabeled data, experiments are run on samples of 20, 200, 500, 1,000, 2,000, 5,000, and 10,000 labeled examples.

[4] The domain relevance experiments are run on samples of 20, 200, 2,000, and 5,000 labeled examples.

[5] The Twitter airline sentiment dataset only has 14,640 examples. The Kaggle natural disaster dataset only has 10,875 examples.

reduction of labeled data from 10,000 to 200 while achieving the same error rate illustrates the advantage of semi-supervised learning using data augmentation.

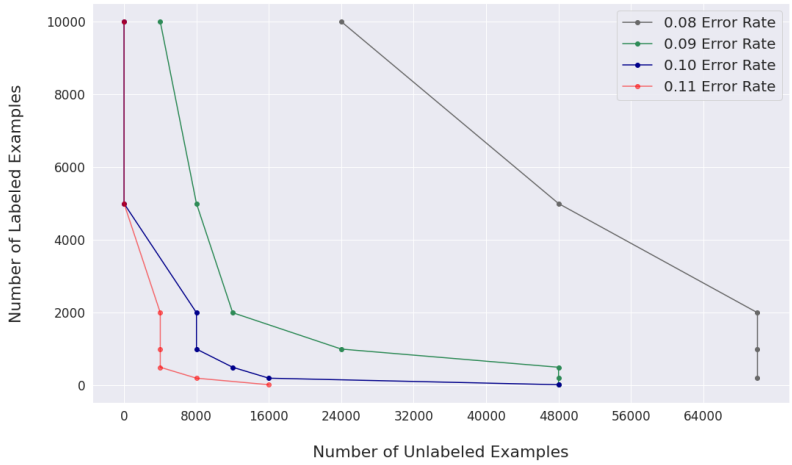

Figure 2: Trade-off between labeled and augmented unlabeled data at various error rate levels.

3.3 Domain Relevance

A common belief is that domain relevance should play a role in semi-supervised learning and that fine-tuning with out-of-domain data will not benefit the model as much as in-domain data would. The results are surprising in two ways.

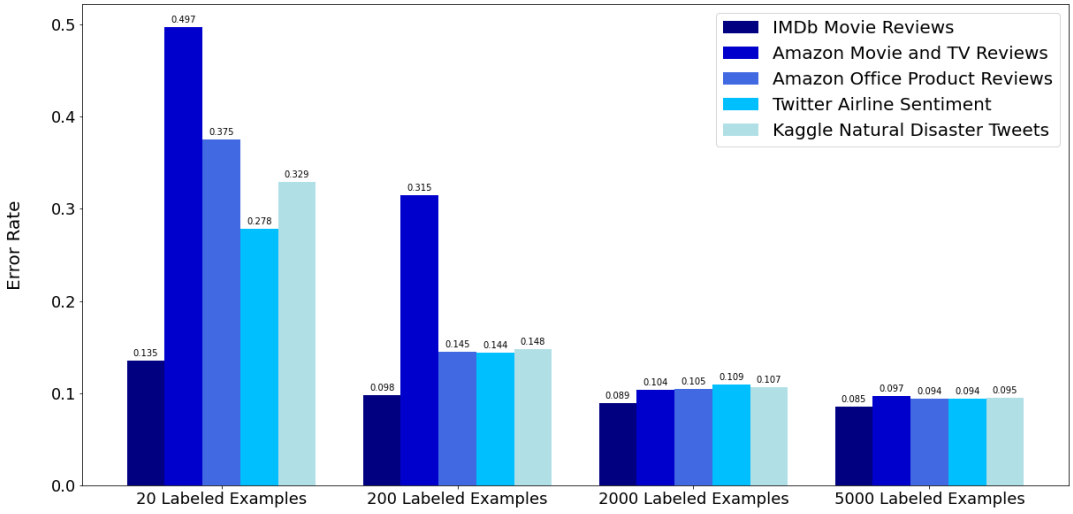

Figure 3: Model performance comparison for different domains.

First, as shown in Figure 3, the performance of the model seems to rely much more heavily on the amount of labeled data than the domain relevance of the unlabeled data. When there are only 20 labeled examples, model performances seem rather random. In particular, it is quite puzzling why the Amazon movie and TV review dataset performs so badly given the fact that it is considered in-domain with IMDb and one would expect the performance to be similar to that of IMDb. At

the 200 labeled example level, with the exception of the Amazon movie data, the performance of the other 3 datasets are extremely similar and such similarity is exhibited as the amount of unlabeled data is increased to 2,000 and then 5,000.

Second, longer text in the unlabeled dataset does not lead to any noticeable benefit to model performance. The Amazon datasets contain considerably more words than the Twitter and Kaggle datasets. Many tweets are under 10 words. Even though more text means more data for fine-tuning, there is no noticeable benefit in terms of model performance.

4 Future Work

In section 3.1, it is mentioned that the models converged to a range of 0.073 to 0.084 when the full unlabeled dataset is used. To investigate the effect of having additional data for fine-tuning, one can either combine other movie review datasets to the IMDb or use a larger sentiment dataset altogether. The latter choice has the benefit of using another dataset to validate the behavior observed on the IMDb data.

The domain relevance experiment can be extended by using more augmented unlabeled data. It would be informative to see whether domain relevance plays a role at much higher amount of augmented unlabeled data. This means the Twitter and Kaggle datasets will need to be replaced since they are rather small.

The back-translation in this work is performed with the WMT'14 English-French translation model in both directions. Since back-translation is a critical part of the model, exploring other translation models may shed light on whether other language translation models lead to better model performance.

5 Conclusion

There is a general trend that a large amount of augmented unlabeled data can compensate for the lack of labeled data. The domain relevance of the augmented unlabeled data does not seem to play an important role. Although additional experimentation on more data will help determine if the observed patterns are generalizable, the results give new light to the way semi-supervised learning can be done. If domain relevance is proved to play a small role, one can improve model performance by potentially fine-tuning on any available data with unsupervised data augmentation. The results also give more motivation for further research and experimentations to uncover better ways to perform consistency training.

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*. September, 2019

2. Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Advances in neural information processing systems, pp. 1195–1204, 2017.

3. Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736*, 2019.

4. 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*, 2018.

5. Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

6. Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. arXiv preprint arXiv:1808.09381, 2018.

7. 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*, 2018.