# Unsupervised Data Augmentation with Naive Augmentation and without Unlabeled Data
(2020)

David Lowell, Brian E. Howard, Zachary C. Lipton, Byron C. Wallace

**Summary**

## Contributions

UDA or Unsupervised Data Augmentation, a popular semi-supervised method that consists of adding a unsupervised loss to force the model to have consistent predictions over the unlabeled data point by minimizing the consistency loss between the clean and augmented unlabeled inputs. In this paper, this authors investigate this method and question the effectiveness and utility of various design choices. Specifically, the authors demonstrate that clever data augmentation such as back-translation might not be necessary and simple augmentation such as random substitution is sufficient. Additionally, they show that the loss can also be very beneficial for standard supervised case and can be applied over the same labeled set and still be highly effective.

## Method

The original UDA loss function consists of computing the cross-entropy loss over the labeled set $L$, and then the consistency loss over the unlabeled set $\mathcal{U}$,

$$\sum_{x,y \in L}[-\log(\hat{p}(y \mid x))] + \lambda \sum_{x \in \mathcal{U}}[\mathcal{L}(x)]$$

In this case, the consistency loss $\mathcal{L}$ is defined as the KL-Divergence between model predictions for the original and augmented examples

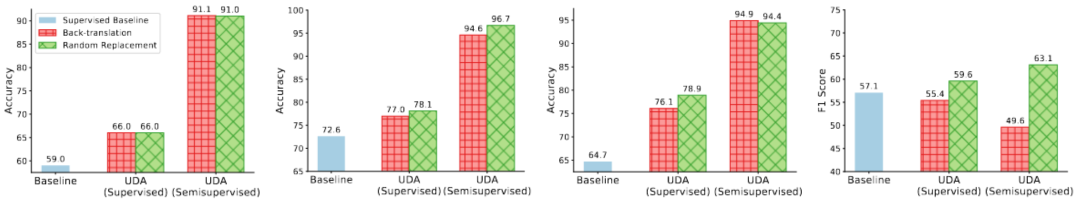
$$\mathcal{L}(x) = \mathcal{D}_{\mathrm{KL}}\{\hat{p}(y \mid x) \| \hat{p}(y \mid q(x))\}$$

where $q$ is a data augmentation operation and $\hat{p}(y \mid q(x))$ is the output probability distribution over the labels produced by the model given an input $x$.

For the augmentation $q$, instead of an elaborate technique such as back-translation where the input in first translate in to a given intermediate language then back into the original language, thus producing a similar but different input. The authors proposed a simple *Uniform Random Word Replacement* augmentation, where for a given word in the input sequence $x$, we either copy the same word with probability $p$ or replace it with a random word with probability $1 - p$. However, this augmentation is only applicable for sentence level tasks, and for work level tasks such as sequence tagging tasks where an work to label alignment is necessary, instead of choosing a word at random, the authors propose to use a pre-trained language model to replace the word, where the replacement is chose of the top 10 predicted words.

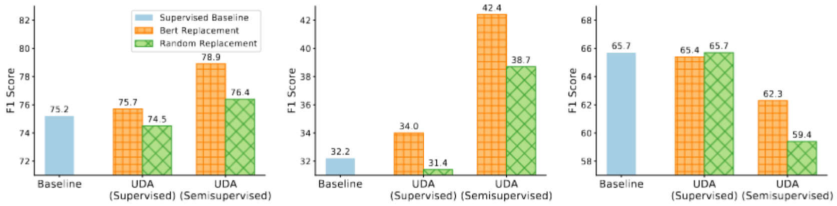
In addition to simplifying the augmentation technique, the authors also simplify the loss function, where in case where the unlabeled set $\mathcal{U}$ is not available, we can simply compute the consistency loss $\mathcal{L}(x)$ over the labeled set itself by applying the augmentation $q$ over the labeled examples and with different original to augmented ratios to simulate a larger unlabeled set (eg, augmenting each labeled example 20 time for a 1:20 ratio).

# Results



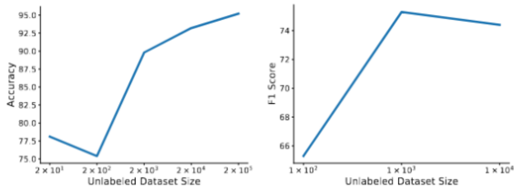
(a) IMDB (b) Yelp (c) Amazon (d) Evidence Inference

Figure 1: Comparison of performance achieved on classification tasks using different variants of UDA. Each bar represents the average performance across five sets of labeled data (labeled data quantity is noted parenthetically). The supervised baseline represents standard ML on the supervised data set only, without any consistency loss. Supervised with consistency loss represents use of consistency loss, but only over the labeled data, with the unlabeled data discarded. Semisupervised with consistency loss represents use of consistency loss over the entire dataset, both labeled and unlabeled.



(a) CoNLL (b) TAC (c) EBMNLP

Figure 2: Comparison of performance achieved on sequence tagging tasks using different variants of UDA. Each bar represents the average performance across ten sets of labeled data (labeled data quantity is noted in parenthesis). The supervised baseline represents standard ML on the supervised data set only, without any consistency loss. Supervised with consistency loss represents use of consistency loss, but only over the labeled data, with the unlabeled data discarded. Semisupervised with consistency loss represents use of consistency loss over the entire dataset, both labeled and unlabeled.



(a) Yelp (b) CoNLL

Figure 3: Comparison of performance achieved using varying quantities of unlabeled data. Curves are averaged across 5 experiments for Yelp, and 10 for CoNLL. Each labeled dataset consists of 20 labeled examples for Yelp and 100 examples for CoNLL. Random replacement is used as the augmentation method for Yelp and BERT-based replacement is used as the augmentation method for CoNLL.

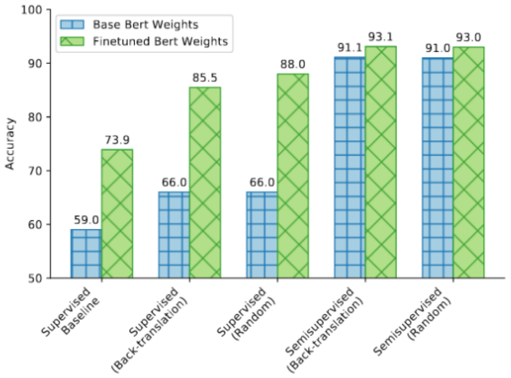


Figure 5: Performance on IMDB using off-the-shelf pretrained BERT weights compared to BERT weights finetuned to IMDB (i.e., after continuing pretraining BERT on IMDB.

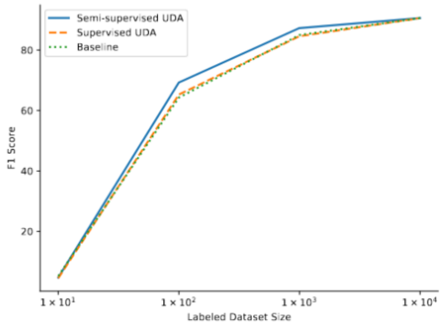


Figure 4: Analysis of performance achieved using varying quantities of labeled data on the CoNLL set. The blue solid line represents the case that UDA is used and all unlabeled data is incorporated in training. The orange dashed line represents the case that UDA is used, but with only the labeled data. The green dotted line represents training without any consistency loss. Each curve is averaged across ten experiments. BERT-based replacement is used as the augmentation method.

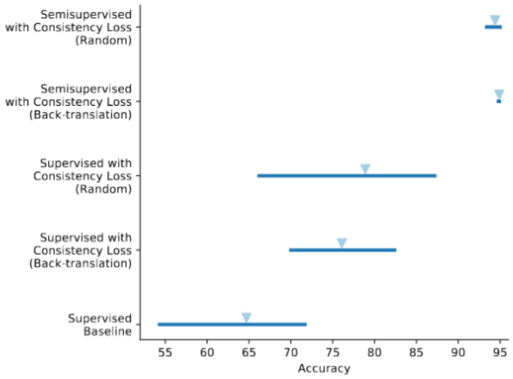


Figure 6: Performance ranges on the Amazon dataset with spans indicating the minimum-to-maximum performance over 5 independent samples (of the labeled subset). Triangles indicate means.