Title of your project

Stanford CS224N Default Project

Danhua Yan
Department of Computer Science
Stanford University
dhyan@stanford.edu

1 Key Information to include

No external collaborators | Not sharing projects

2 Research paper summary

Title	Unsupervised Data Augmentation for Consistency Training
Authors	Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, Quoc V. Le
Venue	Conference on Neural Information Processing Systems (NeurIPS)
Year	2020
URL	`https://arxiv.org/abs/1904.12848v6`

Table 1: Bibliographical information [1].

Background. Semi-supervised learning (SSL) is a widely used approach that leverages unlabeled data to enhance the performance of supervised learning tasks. This is a field that has received considerable attention in deep learning, as high-performing models generally require a substantial amount of high-quality labeled data, which is often costly to acquire. SSL could be particularly beneficial for building robust and generalizable multi-task NLU models. These models, designed to handle a variety of linguistic tasks, typically require abundant data from each task. However, labeled data for some tasks are notably scarcer than for others.

Among the SSL methods, consistency training - which involves ensuring a model produces consistent outputs for an unlabeled example even when it is slightly altered (*i.e.* injecting small noise) - has proven to be effective across numerous benchmarks. In this study, the authors strive to improve SSL consistency training by leveraging advanced data augmentation techniques to inject noise into the input data.

Summary of contributions. This paper demonstrates that effective data augmentation can significantly enhance SSL. The **Unsupervised Data Augmentation (UDA)** framework employs advanced data augmentation techniques, such as back-translation for NLP and RandAugment for vision, as noise injection methods for SSL. UDA achieves top performance on various benchmarks and integrates well with transfer learning. For instance, fine-tuning from BERT with limited labeled examples can match or even surpass a model trained on a larger dataset. On the IMDb text classification dataset, with only 20 labeled examples, UDA achieves an error rate of 4.20, outperforming the state-of-the-art model trained on 25,000 labeled examples.

Utilizing advanced data augmentation requires ensuring the augmented data is of high quality and consistent with the original data distribution. Additionally, it's crucial to address situations where a significant gap exists between the amounts of labeled and unlabeled data, as the model could quickly overfit the labeled data. The authors considered these factors when designing the loss function.

Formally, given a mix of labeled dataset L and unlabeled dataset U, we're trying to learn a classification model p_θ with parameters θ, that maps a given input x to a class distribution $\hat{y} = p_\theta(x)$. Let $q(\cdot)$ denote the data augmentation process, where $\hat{x} = q(x)$ represents the augmented input. We are trying to find θ that minimizes the following loss function: $\boldsymbol{\min_\theta J(\theta) = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{unsup}}}$, where $\mathcal{L}_{\text{sup}}$ is the supervised loss, $\mathcal{L}_{\text{unsup}}$ is the unsupervised loss, λ is the regularization coefficient, usually set to $\lambda = 1$ in this paper.

The supervised loss primarily uses cross-entropy, but the authors propose a technique called Training Signal Annealing (TSA), which dynamically selects a subset of the labeled dataset, L_t, at each training step t. The idea is to exclude easy predictions at each step, helping the model learn from difficult samples and avoid overfitting the labeled dataset too quickly. At each step t, the subset $L_t = \{x \in L \mid p_\theta(x) < \eta(t)\}$. The threshold function $\eta(\cdot)$ sets a dynamic threshold at each step t. Depending on the dataset characteristics, one can choose log, linear, or exponential functions for $\eta(\cdot)$. In conclusion, the supervised loss is:

$$\mathcal{L}_{\text{sup}} = -\sum_{i=1}^{|L_t|} y_i \log p_\theta(x_i)$$

The unsupervised loss is defined as the consistency loss between the unlabeled and augmented data inputs. It's the KL-divergence between the predicted class distributions on unlabeled data, denoted as $\hat{y}_{<ul>}$, and the augmented version, denoted as $\hat{y}_{<aug>}$. To focus on producing consistent data augmentation, the authors propose confidence-based masking. Here, loss is only calculated on a subset of the unlabeled data U_t at step t, where $U_t = \{x \in U \mid p_\theta(x) > \beta\}$. β is a high constant (typically 0.8), including only samples where the model is confident, so that the consistency loss can concentrate more on the distribution discrepancies resulting from suboptimal data augmentation. Moreover, the authors claimed that sharpened predictions for unlabeled data will further improve performance, by adjusting Softmax temperature τ to 0.4. Concretely, $\hat{y}_{<ul>} = p_{\tilde{\theta}}^{\text{sharp}}(x), \hat{y}_{<aug>} = p_\theta(q(x))$, the unsupervised loss is:

$$\mathcal{L}_{\text{unsup}} = -\sum_{j=1}^{|U_t|} p_{\tilde{\theta}}^{\text{sharp}}(x_j) \log \frac{p_\theta(q(x))}{p_{\tilde{\theta}}^{\text{sharp}}(x_j)}$$

Note that the term $p_{\tilde{\theta}}(x_j)^{\text{sharp}}$ uses a *fixed* copy of the current parameters, denoted as $\tilde{\theta}$, to indicate that the gradient is not propagated through $\tilde{\theta}$. This is to ensure the loss is minimizing the divergence against a stable reference against current model parameters.

In conclusion, the authors demonstrate that advanced data augmentation techniques provide a superior source of noise for consistency enforcing SSL. UDA performs exceptionally well in text and vision tasks, rivaling fully supervised models trained on larger datasets. This sheds light on future research opportunities to transfer advanced supervised augmentation techniques to the SSL setting for various tasks.

Limitations and discussion. While UDA excels in binary classifications and NLP tasks with limited training samples, it falls short in multi-class tasks like Yelp-5 and Amazon-5 datasets. The paper relies heavily on empirical results, lacking in-depth mathematical reasoning and research into the benefits of the adjusted loss function. This leaves future researchers unsure of where to focus. Despite these limitations, UDA's design is sound and applicable to a variety of tasks in both language and vision, inspiring the transformation of supervised tasks into a semi-supervised manner.

Why this paper? Many top-performing solutions on multi-task NLP benchmarks like GLUE and SuperGLUE have adopted UDA's framework. They utilize the powerful knowledge retrieval ability of recent large language models (LLMs) and gain from semi-supervised learning, which offers more training data and robust generalization. In numerous real-world scenarios, companies often lack high-quality annotated data, and online open-source datasets for research are not tailored for specific use cases. UDA's method highlights opportunities for using generator model as a data augmentation approach for few-shot or even zero-shot learning. It's inspiring to see the approach and results of this early paper focusing on this topic.

Wider research context. The UDA framework significantly contributes to NLP research by introducing a novel approach to language representation. It leverages advanced data augmentation techniques like back-translation, enhancing performance on tasks like text classification and promising wider applicability. Techniques like confidence-based masking and Training Signal Annealing could extend to other NLP tasks with labeling concerns. UDA's success in semi-supervised learning scenarios highlights its potential for few-shot or zero-shot learning, a promising future NLP research direction.

3 Project description

Goal. The primary goal of this project is to boost supervised NLP tasks performance by integrating BERT with UDA techniques, effectively transforming a supervised problem into a semi-supervised setting for robust and generalizable model training. By implementing key aspects of the original BERT and utilizing its pre-trained weights, we aim to establish a strong baseline in supervised NLP tasks. We then transition to a semi-supervised setting using back-translation for noise injection, as suggested by UDA. Additionally, we plan to experiment with large language models (LLMs) to generate synthetic examples, advancing our understanding of their potential in few-shot and zero-shot learning scenarios. This project seeks to answer whether BERT, combined with innovative semi-supervised techniques, can significantly improve performance in NLP tasks with limited labeled data, leveraging the strengths of both supervised and unsupervised approaches.

Task. After implementing key aspects of the original BERT model, load pre-trained weights into the BERT model. Perform sentiment analysis on the SST and CFIMDB datasets. Extend the BERT embeddings to simultaneously perform three tasks: sentiment analysis, paraphrase detection, and semantic textual similarity.

Data. For the baseline NLP tasks, we will use the BERT model to perform sentiment analysis using the Stanford Sentiment Treebank (SST) dataset (5-classes) and the CFIMDB movie reviews dataset (binary classes). For the extended downstream tasks, we will use the SST for sentiment analysis, the Quora dataset for paraphrase detection, and the SemEval SST Benchmark dataset for semantic textual similarity tasks. Moreover, we will use data augmentation data generated by LLMs as an unlabeled dataset to improve the generalization of the BERT model.

Methods. Initially, we will adhere to the default project handouts to implement minBERT and conduct baseline tasks on SST and CFIMDB datasets. Additionally, we plan to leverage UDA techniques, utilizing the proposed loss function, and creating augmented examples through back-translation. This process involves translating an existing example in language A into another language B, then translating it back to A to obtain an augmented example. We aim to leverage TogetherAI to create such augmented examples. Subsequently, BERT will be fine-tuned in a semi-supervised manner, aiming to boost performance on multiple downstream tasks.

Baselines. We will use pre-trained weights from the BERT model without finetuning as embeddings. The pooled output will be projected to a linear layer for classification, as described in the default project handout. This will serve as our baseline. All extensions will be evaluated against this baseline.

Evaluation. Specify at least one well-defined, numerical, automatic evaluation metric you will use for quantitative evaluation. What existing scores will you be comparing against for this metric? For example, if you're reimplementing or extending a method, state what score(s) the original method achieved; if you're applying an existing method to a new task, mention the state-of-the-art performance on the new task, and say something about how you expect your method to perform compared to other approaches. If you have any particular ideas about the qualitative evaluation you will do, you can describe that too.

Ethics. What are the ethical challenges and possible societal risks of your project, and what are mitigation strategies? Please provide a 1 paragraph description of A) the ethical challenges and possible negative societal risks regarding your project, identifying at least 2 valid ethical concerns or potential societal risks, and B) some practical, specific strategies to mitigate those risks (you are not required to implement these mitigation strategies). Try to be a little specific and intentional in describing ethical concerns (e.g. more depth than ‘there will be issues with user data privacy’), and especially around issues specific to your project. Writing clear limitations and identifying potential risks is something that we’re encouraging as part of a broader trend in the research community to include such statements as part of published research. This includes Stanford’s Ethics and Society Review (ESR) and the EMNLP Impact Statement, which are great resources for getting started. This section will be graded primarily for effort and intentionality, and feedback may be provided.

References

[1] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.