CS 839 Scribing

Liang Shang, Siyang Chen

1 Introduction

We are introducing unsupervised data augmentation (UDA), an augmentation method that focus on the quality of injected noise, which delivers substantial improvements in semi-supervised training results. UDA substitutes simple noising operation (such as simple Gaussian or dropout noise) with advanced data augmentation methods (such as RandAugment and back-translation). UDA performs better on six classification tasks: IMDb, Yelp-2, Yelp-5, Amazon-2, Amazon-5 for text classification and CIFAR-10, SVHN for image classification. Semi-supervised learning has shown promising improvements in deep learning models when labeled data is scarce. Common recent approaches involve using of consistent training on large amount of unlabeled data to constraint model predictions to be invariant to input noise.

2 Unsupervised Data Augmentation (UDA)

Consistency Training

Consistency training regularizes model predictions to be invariant to small noises to either input examples or hidden states. (This make the model robust to any small changes). Most methods under this framework differs in how and where the noise injection is applied. Advanced data augmentation methods used in supervised learning also perform well in semi-supervised learning. (Strong correlation present).

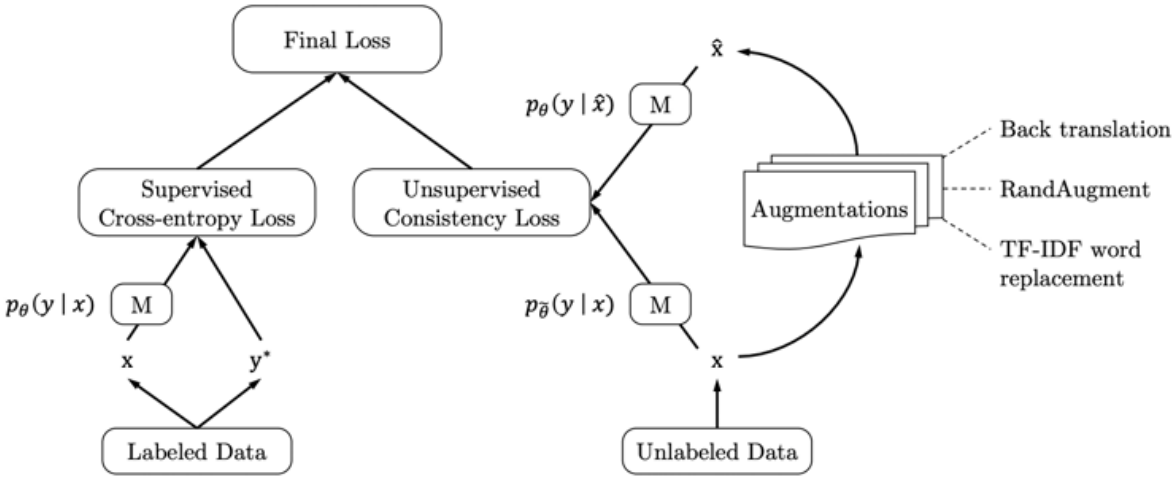
Supervised Data Augmentation

let $q(\hat{x}|x)$ be the augmentation transformation from which one can draw augmented examples $\hat{x}$ based on an original example x. It is required that any example $\hat{x} \sim q(\hat{x}|x)$ drawn from the distribution shares the same ground-truth label as x. Equivalent to constructing an augmented labeled set from the original supervised set and then training the model on the augmented set. (The augmented set needs to provide additional inductive biases to be more effective). Despite promising results, data augmentation only provides a steady but limited performance boost because these augmentations has only been applied to a set of small-size labeled examples. This limitation motivated semi-supervised learning where abundant data is available.

Unsupervised Data Augmentation

This procedure enforces the model to be insensitive to the noise. This is essentially minimizing the consistency loss gradually propagates label information from labeled examples to unlabeled ones. The UDA presented in this paper focus on the ‘quality’ of

the noise operation and its influence on performance of consistency training network. The mechanism is explained in figure below:



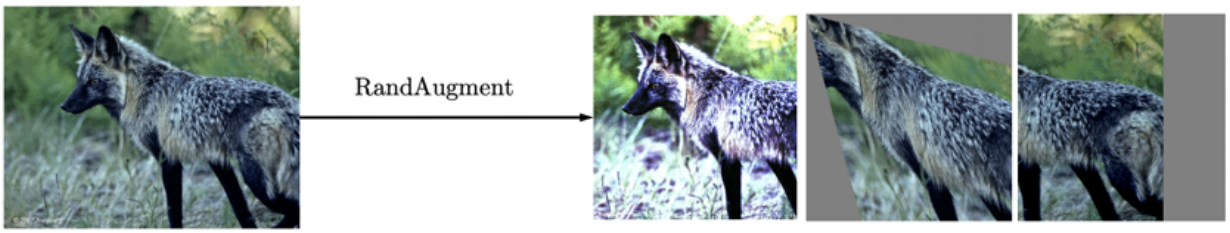
The UDA mechanism utilize a weighting factor λ when trained with labeled examples. This is used to balance the supervised cross entropy and the unsupervised consistency training loss.

Advantage of advanced data augmentation

- Valid noise: Advanced data augmentation methods generates realistic augmented examples that share the same ground-truth labels with the original example.
- Diverse noise: Advanced data augmentation can generate a diverse set of examples since it can make large modifications of the input example without changing its label.
- Targeted inductive biases: Data augmentation operations that work well in supervised training essentially provides the missing inductive biases.

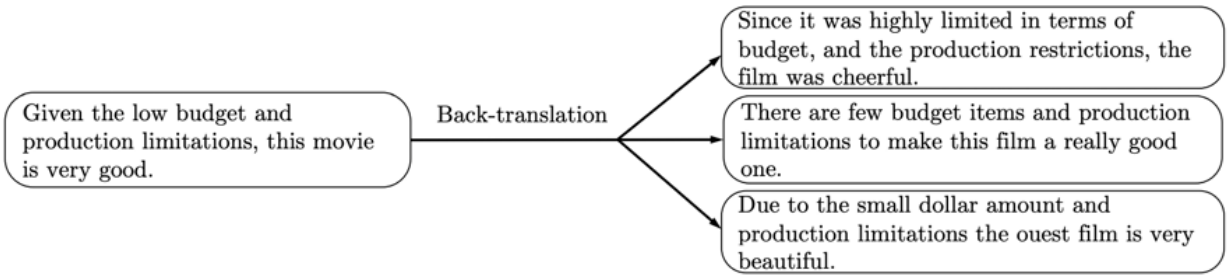
Augmentation Strategies – Image Classification

- RandAugment is used for image data augmentation.
- Instead of searching, RandAugment sample uniformly from the Python Image Library (PIL).
- This makes RandAugment simpler and requires no labeled data as there is no need to search for optimal policies.



Augmentation Strategies - Text Classification

- Back-Translation is used for text classification.
- The procedure is translating an existing example x in language A into another language B and then translating it back into A to obtain an augmented example $\hat{x}$.
- Back-translation can generate diverse paraphrases while preserving the semantics of the original sentence, which improves performance.
- A random sampling with a tunable temperature is used for the generation.



Word replacing with TF-IDF for text classification

- Simple back-translation has little control over which words will be retained, but this requirement is important for topic classification tasks (some key words are more informative than others).
- To address this problem, UDA replaces uninformative words with low TF-IDF scores while keeping those with high TF-IDF values.

Additional Training Techniques - Confidence based masking

- Examples that the current model is not confident about is masked.
- This is done by controlling the calculation of consistency loss in each minibatch.
- Specifically, consistency loss is computed only on examples whose highest probability among classification categories is greater than a threshold β.
- This threshold β is set to a high value to avoid calculating unsure models.

Additional Training Techniques - Sharpening Predictions

- Regularizing predictions to have low entropy is beneficial, thus prediction sharpening is done when computing the target distribution on unlabeled examples by using low temperature τ.

Additional Training Techniques - Domain Relevance Data Filtering

- Class distributions of out-of-domain data are mismatched with those of in-domain data, so simply use out-of domain unlabeled data is not sufficient.
- To obtain data relevant to the domain for task at hand, the baseline model trained on the in-domain data is used to infer the labels of data in a large out-of-domain dataset and the examples our model is most confident are picked out.
- This is essentially sorting all examples based on classified probability (for each category) and select the examples with the highest probabilities of being in that category.

3 Theoretical Analysis

Theoretical Assumptions

- **In-domain** augmentation: data examples generated by data augmentation have non-zero probability under p_U, i.e., $p_U(\hat{x}) > 0$ for $\hat{x} \sim q(\hat{x}|x), x \sim p_U(x)$
- **Label-preserving** augmentation: data augmentation preserves the label of the original example, i.e., $f^*(x) = f^*(\hat{x})$ for $\hat{x} \sim q(\hat{x}|x), x \sim p_U(x)$
- **Reversible** augmentation: the data augmentation operation can be reversed, i.e., $q(\hat{x}|x) > 0 \Leftrightarrow q(x|\hat{x}) > 0$

Theoretical Intuition

For a graph G_{p_U}, where each node corresponds to a data sample $x \in X$ and an edge $(\hat{x}, x)$ exists iff $q(\hat{x}|x) > 0$, if we have an N-category classification problem. Then by an ideal data augmentation method, G_{p_U} should have exactly N components.

And for each component C_i of the graph, as long as we have one labeled data, by traversing C_i via augmentation operation $q(\hat{x}|x)$, we can propagate the label over all data in C_i.

So, in order to find a perfect classifier via such label propagation, there should exist at least one labeled example in each component, which means the number of components is

the lower bound the minimum amount of labeled examples needed. Then, since with a better augmentation method the number of components can be decreased, the minimum amount of labeled examples needed can also be decreased.

Theoretical Analysis

Theorem 1. Under UDA, let P_i be the total probability that a labeled data point fall into the i-th components, i.e., $P_i = \sum_{x \in C_i} P_L(x)$. Let $Pr(A)$ donate the probability that the algorithm cannot infer the label of a new test example given m labeled examples from P_L. $Pr\ (A)$ is given by

$$Pr(A) = \sum_i P_i\ (1 - P_i)^m$$

In addition, $O(k/\epsilon)$ labeled examples can guarantee an error rate of $O(\epsilon)$, i.i.e.,

$$m = O\left(\frac{k}{\epsilon}\right) \Rightarrow Pr(A) = O(\epsilon)$$

Where k is the number of components in G_{p_U}.

4 Experiment Results

Step I: Correlation between supervised and semi-supervised performance

1. Stronger data augmentations found in supervised learning can always lead to more gains when applied to the semi-supervised learning settings.

Step II: Algorithm comparison on vision semi-supervised learning benchmarks – vary the size

1. UDA consistently outperforms the two baselines given different sizes of labeled data.
2. The performance difference between UDA and VAT shows the superiority of data augmentation based noise. The difference of UDA and VAT is essentially the noise process. While the noise produced by VAT often contain high-frequency artifacts that do not exist in real images, data augmentation mostly generates diverse and realistic images.

Step III: Algorithm comparison on vision semi-supervised learning benchmarks – vary the model

1. UDA outperforms all published results by significant margins and nearly matches the fully supervised performance, which uses 10x more labeled examples, which shows the huge potential of state-of-the-art data augmentations under the consistency training framework in the vision domain.

Step IV: Evaluation on test classification datasets

1. Even with very few labeled examples, UDA can offer decent or even competitive performances compared to the SOTA model trained with full supervised data. Particularly, on binary sentiment analysis tasks, with only 20 supervised examples, UDA outperforms the previous SOTA trained with full supervised data on IMDb and is competitive on Yelp-2 and Amazon-2.
2. UDA is complementary to transfer learning / representation learning. As we can see, when initialized with BERT and further finetuned on in-domain data, UDA can still significantly reduce the error rate from 6.50 to 4.20 on IMDb.
3. For five-category sentiment classification tasks, there still exists a clear gap between UDA with 500 labeled examples per class and BERT trained on the entire supervised 8 set. Intuitively, five-category sentiment classifications are much more difficult than their binary counterparts. This suggests a room for further improvement in the future.

Step V: Scalability test on the ImageNet dataset

1. In both 10% and the full data settings, UDA consistently brings significant gains compared to the supervised baseline. This shows UDA is not only able to scale but also able to utilize out-of-domain unlabeled examples to improve model performance.

5 Conclusion

- Data augmentation and semi-supervised learning are well connected, better data augmentation can lead to significantly better semi-supervised learning.
- UDA generate diverse and realistic noise and enforces the model to be consistent with respect to these noises.
- UDA combines well with representation learning & transfer learning, and nearly matches the performance of fully supervised models trained on full labeled sets which are larger by one magnitude in size.
- UDA are able to scale and utilize out-of-domain data to improve the performance