Efficient Semi-supervised Consistency Training for Natural Language Understanding

George Leung
Amazon Alexa AI, USA
leu@amazon.com

Joshua Tan
Amazon Alexa AI, USA
jshtan@amazon.com

Abstract

Manually labeled training data is expensive, noisy, and often scarce, such as when developing new features or localizing existing features for a new region. In cases where labeled data is limited but unlabeled data is abundant, semi-supervised learning methods such as consistency training can be used to improve model performance, by training models to output consistent predictions between original and augmented versions of unlabeled data.

In this work, we explore different data augmentation methods for consistency training (CT) on Natural Language Understanding (NLU) domain classification (DC) in the limited labeled-data regime. We explore three types of augmentation techniques (human paraphrasing, back-translation, and dropout) for unlabeled data and train DC models to jointly minimize both the supervised loss and the consistency loss on unlabeled data. Our results demonstrate that DC models trained with CT methods and dropout-based augmentation on only 0.1% (2,998 instances) of labeled data with the remainder as unlabeled can achieve a top-1 relative accuracy reduction of 12.25% compare to fully supervised model trained with 100% of labeled data, outperforming fully supervised models trained on 10x that amount of labeled data. The dropout-based augmentation achieves similar performance compare to back-translation-based augmentation with much less computational resources. This paves the way for applications of using large scale unlabeled data for semi-supervised learning in production NLU systems.

1 Introduction

Deep learning, especially transformer-based language models (Vaswani et al., 2017), have achieved state-of-the-art performance in many tasks and are widely used in NLU systems. A challenge in deep learning is that it often requires large amounts of labeled training data in order to reach a desirable performance level. This is especially a problem for NLU systems in commercial production as the cost of labeling data scales with the expanding number of supported features and languages.

Recent research in semi-supervised learning (SSL) demonstrated that it is possible to combine a small amount of labeled data and a large amount of unlabeled data to match or even outperform purely supervised learning (Xie et al., 2020; Gao et al., 2021). One of the most promising approaches in SSL is called consistency training (Bachman et al., 2014; Rasmus et al., 2015; Tarvainen and Valpola, 2017; Verma et al., 2019). In short, consistency training is a technique that regularizes model predictions to be invariant to augmentations of unlabeled data. Examples of augmentations include applying noise to input features (Sajjadi et al., 2016; Miyato et al., 2018) or hidden states (Bachman et al., 2014).

In this paper, we experimented with consistency training in a major NLU task: Domain Classification (DC). We tested three different types of data augmentations: paraphrasing by user feedback, back-translation, and dropout. As a testbed for our approach, we applied our experiments to BERT (Devlin et al., 2019)-based models using a real-world dataset of voice-controlled assistant in Portuguese. We found that all three types of augmentations can be effectively used alongside consistency training to improve model performance compared to a baseline model trained without consistency training. For the scenario where labeled data was limited to only 0.1% of all available labeled data, the best top-1 accuracy, which is -9.14% compare to fully supervised model trained with 100% labeled data, was achieved by consistency train-

ing on data augmented using back-translation. If we use dropout-only augmentation, the relative top-1 accuracy change was -12.25%. Lastly, we observed a relationship between the amount of labeled data used for training and the size of CT benefits, with larger benefits for smaller sets of labeled data. Our results demonstrate the possibility of using consistency training to drastically reduce the amount of labeled data needed for an NLU system while retaining a reasonable accuracy. This can be done on large unlabeled datasets without using computationally expensive back-translation or financially costly human-authored augmentation.

2 Background

2.1 Consistency training

Consistency training (Bachman et al., 2014; Rasmus et al., 2015; Tarvainen and Valpola, 2017; Verma et al., 2019) is a Semi-Supervised Learning technique that utilizes unlabeled data to enforce consistency of the model output given similar inputs. The general schematic of this method is shown in Figure 1. In summary, consistency training is a multitask learning with two objectives: minimizing the supervised loss for labeled data and the consistency loss for unlabeled data. The supervised loss is a regular cross-entropy loss for the labeled data. For the consistency loss, the unlabeled data is first paraphrased with data augmentation methods. Then the original data x and the augmented data x' will be passed through the same encoder model M to generate two output distributions respectively $p_M(y|x)$ and $p_M(y|x')$. The consistency loss is defined by the Kullback–Leibler divergence between the two output distributions $D(p_M(y|x)||p_M(y|x'))$. Finally the consistency loss and supervised loss are combined and back-propagated to update the model parameters. In this way consistency training forces the model to be insensitive to the noise introduced by data augmentation.

2.1.1 Paraphrasing by user feedback

MARUPA (Falke et al., 2020) (Mining Annotations from User Paraphrasing) is a tool to leverage real-world user implicit feedback to collect paraphrased utterances. Sometimes when a user is having a failed interaction with the system, the user will paraphrase the utterance

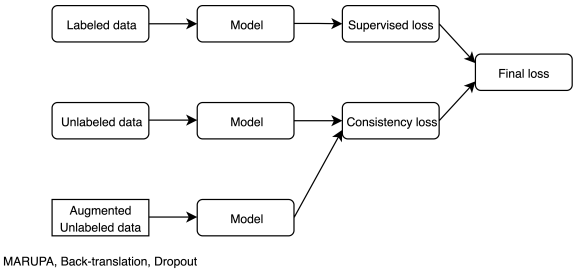

Figure 1: Training objective for consistency training. Note that the three *model* blocks in this figure represent the same encoder model with the same set of parameter.

and retry. MARUPA collects these utterances fully autonomously without the need for human annotators using paraphrase detection, friction detection and label projection models. This dataset is filtered to make sure it is relevant to the main task (Domain classification). In our experiment, we use the MARUPA dataset without the labels as the augmented unlabeled dataset for the consistency training.

2.1.2 Paraphrasing by back-translation

Back-translation a common approach for data augmentation in NLP (Xie et al., 2020; Edunov et al., 2018). Recent development of Neural Machine Translation (NMT) (Vaswani et al., 2017), which are trained on extremely large corpora, have impressive accuracy in translating text. Back-translation leverages this to generate augmented data by translating example text sequences from original language to an intermediate language and then back to original language. This method allow us to generate paraphrases while retaining the semantic structure, and has shown to improve performances in question answering tasks (Yu et al., 2018; Dong et al., 2017). In our experiment, we leverage a commercially available cloud-based translate service to paraphrase the unlabeled dataset using back-translation.

2.1.3 Dropout as data augmentation

Dropout (Srivastava et al., 2014) is a technique to prevent overfitting in training deep neural networks by randomly dropping units inside the network. In recent researches (Bouthillier et al., 2015; Gao et al., 2021) dropout is also shown to be an effective method for data augmentation. The underlying idea is to pass the same input sequence to the encoder twice with different

dropout mask. The two resulting embeddings are then used to compute the consistency loss. This method outperforms several deterministic augmentation approaches such as word deletion and replacement (Gao et al., 2021). Another advantage of this method is no extra paraphrase process is needed and we can directly use the unlabeled data for consistency learning.

3 Experiment

We designed our experiments to explore the performance impact of incorporating consistency training using each type of data augmentation. We also investigate how performance changes as the amount of labeled data or unlabeled data used for training is varied.

3.1 Consistency-training (CT) models

All the models are based on a distilled (Hinton et al., 2015) Portuguese BERT (Devlin et al., 2019) language model. This model has 4 transformer layers and feedforward hidden dimension of 1200 compare to 12 and 3072 in the BERT-Base model. All experiments were trained on Amazon Web Services EC2 p3.16xlarge instances. We implemented CT using a multi-task learning framework that trained models to jointly minimize the sum of supervised cross-entropy error on labeled data and the consistency loss on unlabeled data. All models were configured to train for up to 20 epochs. During training, CT models alternate between computing loss on the supervised task and the consistency-loss task. The task sampling rates were set such that both tasks would finish iterating through their associated data at approximately the same time. We compare the CT models against a set of baseline models that only performed supervised training.

3.2 Augmentations

We experimented with a total of five CT models varying in type of data augmentation used for consistency regularization: paraphrase by humans (MARUPA), back-translation, and dropout.

For *MARUPA CT* models, augmentations were comprised of paraphrase data. We leveraged the MARUPA paraphrase dataset as unlabeled pairs of augmented data. This dataset consisted of 2,258,828 utterance pairs (4,517,656 total).

For *Back-translate CT* models, augmentations were comprised of back-translated utterances. We used a cloud-based Translate service to translate from Portuguese to an intermediate language and back to Portuguese, generating a total of 2,998,782 pairs. For some pairs the original and back-translated utterances are the same, and in that case we switched to a different intermediate language until a different back-translated utterance is obtained. The list of intermediate language is *English, French, Japanese, Korean, Chinese, Hindi* and *Hungarian*.

For *Dropout CT* models, we used dropout to generate an equivalent of data augmentation on the embedding space. Our dropout augmentation involved applying dropout to the same data instance twice with different dropout masks using the same dropout probability. Dropout layers were located in each BERT transformer blocks and fully connected layer with dropout probability set to 0.1. The unlabeled data used in *Dropout CT* is the same as the original data in the back-translation dataset.

We also test two combinations of augmentations. In *Dropout+MARUPA CT* models, we combine dropout and paraphrase augmentations. Specifically, we applied independently sampled dropout to both utterances in a paraphrase pair, and then compute the consistency loss between the dropout-augmented pair. For *Dropout+Back-translate CT* models, we combine dropout with back-translation pairs in similar fashion.

3.3 Training data

We experimented with six different labeled-data sizes: 0.1%, 1%, 2%, 5%, 25%, and 100% of the available training data. We randomly sampled three sets of data for each labeled-data size less than 100%. Within each sample, we use a randomly selected 90% as the training data and use the remaining 10% as the validation set. Unless otherwise stated, for each model we experimented with we trained three separate instances, each using a different data split.

We also experimented with different unlabeled data sizes. For this set of experiments we limited our exploration to Dropout CT models that were trained with 0.1% of the available labeled data. For all Dropout CT models, we treat the remaining labeled data as the set

of available unlabeled data (i.e., for a model trained using 0.1% of the labeled data, we take the remaining 99.9% and remove the label). We experimented with models that used 25%, 50%, 75%, and 100% of the available unlabeled data. As before, we created three random samples for each unlabeled-set size less than 100% and trained a separate model on each split.

3.4 Evaluation

We evaluated our models using a held-out test set. We considered two different types of testing scenarios. In the first, we tested against the full test set of 191,762 utterances, approximating the distribution of a real-world application scenario. In the second, we tested against a test set that had been filtered to remove all utterances appearing in the training set. This filtered set contained 46,211 utterances and was intended to examine how well our models are able to generalize to unseen utterances.

Our experiments were performed using a production BERT-based domain classification model. Models with differing architectures or for different ML tasks may not yield the same results. Similarly, our results may not generalize to industry applications of NLU in other domain areas, using different spoken languages, or with access to substantially larger amounts of labeled training data.

4 Results

Here we present the results of our consistency-training experiments and illustrate how model performance changed as we varied the underlying training data.

4.1 Metrics definition

All metrics are reported as relative change, including Top-1 accuracy, Top-1-Unseen accuracy, false accept rate and false reject rate. The relative change is defined by

$$(\mu - \mu_r)/\mu_r$$

where μ is the experiment metric and μ_r is the reference metric achieved by the fully supervised model trained on 100% of labeled data.

4.2 Size of labeled data

Our results show that consistency training on augmented data can lead to significant improvements in performance in limited-data settings. As shown in Table 1, when restricting models to use only 1% of the available labeled data as training data, the baseline supervised model achieves a top-1 accuracy of -67%. For the Dropout CT model trained on the same 1% of labeled data, we see a top-1 accuracy of -4%. The difference in performance is even more apparent in models trained using only 0.1% of the labeled data. For models trained with 0.1% of the labeled data, the baseline model achieved an top-1 accuracy of only -99%. The Dropout CT model trained on the same amount of labeled data achieve a top-1 accuracy of -12.25%. This improvement in top-1 accuracy demonstrates the utility of consistency training on unlabeled data when labeled data is extremely limited. Table 1 also compares the top-1 accuracy of the baseline and Dropout CT model when tested on utterances not seen during training. As expected, given the same model the top-1-unseen accuracy is lower than the top-1 accuracy, as this represents a more difficult task. However, we still see a performance improvement in top-1-unseen accuracy when applying consistency training.

In Figure 2 we plot the top-1 accuracy of the baseline and Dropout CT model as we vary the amount of labeled training data. While both the baseline and Dropout CT models benefit from training with additional labeled data, the benefit is much greater for the baseline model. Figure 2 also sheds light on the difficulty of the domain classification task. We see that a baseline model trained on 2% of the labeled data has comparable perfomance to a baseline model trained on all the labeled data.

4.3 Size of unlabeled data

Results on varying the size of the unlabeled training data our Dropout CT model trained with 0.1% of the available labeled data are shown in Figure 3. We see that even when using only 25% of the unlabeled data (742k instances), consistency training with dropout-based augmentations achieves a top-1 accuracy of -23%. Increasing the amount of unlabeled data generally led to improved performance.

4.4 Types of augmentation

Table 2 shows our experiments comparing CT models that use different types of data augmentations, where each model was trained on only

		Top-1		Top-1-Unseen	
% Labeled data	Count	Baseline	Dropout CT	Baseline	Dropout CT
0.1%	2998	-98.96%	-12.25%	-98.16%	-26.66%
1%	26989	-67.33%	-4.16%	-67.67%	-9.09%
2%	53978	-2.40%	-2.71%	-14.73%	-5.62%
5%	134945	-1.52%	-2.50%	-3.12%	-4.64%
25%	674725	-0.60%	-0.64%	-1.39%	-1.39%
100%	2698903	0%	-	0%	-

Table 1: Top-1 accuracy relative change for baseline models trained on different amounts of labeled data.

		FAR			FRR		
	Top-1	Video	Shopping	Music	Video	Shopping	Music
Baseline	-98.96%	-100%	-100%	-100%	137%	766%	2877%
Dropout CT	-12.25%	308%	344%	71%	59%	346%	543%
MARUPA CT	-22.42%	1145%	2844%	14%	64%	191%	1760%
Back-translate CT	-9.14%	370%	733%	106%	27%	20%	132%
Dropout+MARUPA CT	-21.79%	839%	3372%	14%	73%	236%	1695%
Dropout+Back-translate CT	-9.66%	267%	567%	131%	32%	14%	91%

Table 2: Top-1 accuracy, false acceptance rate (FAR), and false rejection rate (FRR) relative change for the supervised baseline model and the consistency-training models using different underlying data augmentations. All models are trained with 0.1% labeled data. Metrics are reported as relative change compare to fully supervised model trained using 100% of labeled data. The ground truth test data included 44,221 Music utterances, 2,145 Shopping utterances, and 904 Video utterances.

0.1% of the labeled data. Overall, every data augmentation method helps CT to perform better than the baseline model. Out of all the augmentation methods we tested, Back-translate CT performed best. The Back-translate CT model achieved an average top-1 accuracy of -9.14%, followed by the Dropout+Back-translate CT model with a top-1 accuracy of -9.66%. MARUPA models in general performed worse than Back-translate models, but still had significant improvement over the baseline.

We find mixed results on the performance benefit of combining types of augmentations together for consistency training. While the Dropout+MARUPA CT model had a slightly higher top-1 accuracy than the MARUPA CT model (-21.79% vs. -22.42%), the Dropout+Back-translate CT model performed slightly worse than Back-translate CT (-9.66% vs. -9.14%).

We note that the Dropout CT methods, although slightly less performant than Back-translate CT models, have a greater advantage from an operations perspective. Dropout augmentation does not require any kind of domain expertise, pre-computation, or external translation models, which can greatly reduce data-preprocessing time and operational costs.

In addition to top-1 accuracy, Table 2 shows false acceptance and false reject rates for three differently sized domains. The baseline model incorrectly rejects all utterances for which the ground truth domain was one of Video, Shopping, or Music. More interestingly, for a pair of models the better performing model in terms of top-1 accuracy was not always the better performing model in terms of false acceptance or rejection rates for a given domain. For example, although the Dropout CT model had a higher top-1 accuracy than the MARUPA CT model (-12.25% vs. -22.42%), if lowering the false reject rate for the Shopping domain is the highest priority, then the MARUPA CT model may be more appropriate.

5 Related work

5.1 Data Augmentation in NLP

Hedderich et al. (2021) provide a survey of NLP techniques for training models in low-resource

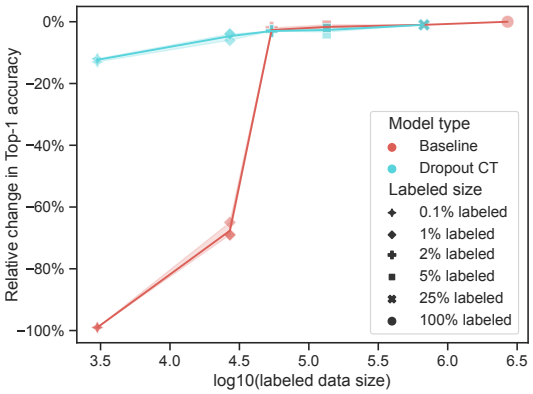

Figure 2: Comparison of top-1 accuracy relative change for baseline and Dropout CT models trained on different amounts of labeled data. Data points are shown for all three experiments run for a given model differing only in training sample (often overlapping).

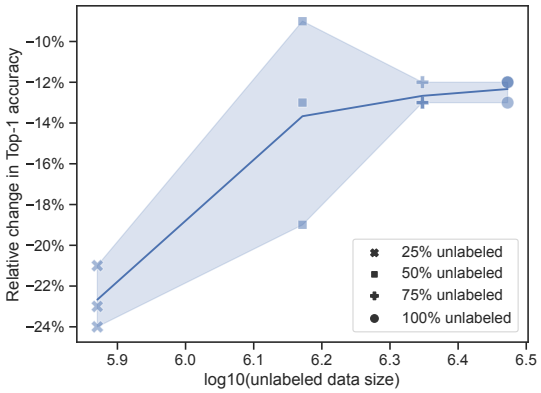

Figure 3: Comparison of top-1 accuracy relative change for Dropout CT models trained on different amounts of unlabeled data. All models were trained using 0.1% of the labeled data.

scenarios. One of the most common techniques to address this is data augmentation, which produces new input instances by applying transformations to existing data.

In our study, we applied hidden-space augmentations by using independently sampled dropout masks for the same instance. Prior work has also proposed dropout as a data augmentation technique. Bouthillier et al. (2015) demonstrate that the effect of dropout on a neural network can be replicated by projecting dropout noise back into the input space and training a model on the generated data. Zhao et al. (2019) show that dropout can be viewed as equivalent to data augmentation whenever the input space dimension is equal to or higher than the output space.

5.2 Consistency training

Consistency regularization, also known as *consistency training* (Chen et al., 2021), is a popular technique in Semi-Supervised Learning. The underlying idea is that model predictions should not change much when data instance is perturbed. Xie et al. (2020) proposed UDA, a framework for leveraging data augmentation in SSL settings by jointly minimizing a standard supervised loss with consistency-based loss on data and its augmentations.

5.3 Contrastive learning

The goal of contrastive learning (Chopra et al., 2005), which is very similar to consistency learning, is to learn a data representation such that similar data instances are located near to each other in and dissimilar instances are pushed apart. Wang and Isola (2020) showed that optimizing contrastive metric can lead to better *alignment* and *uniformity* of features in the embedding space. Gao et al. (2021) show that standard dropout noise can outperform other types of data augmentation for contrastive learning of sentence embeddings.

6 Conclusion

With the aim of developing a strategy to efficiently leverage large amounts of unlabeled data in deployed NLU systems, we examined three different augmentation techniques for consistency training using real-world data. Back-translation performed the best, dropout was slightly behind and paraphrase by human users was the worst-performing technique. From an operations perspective dropout is more favorable because it doesn't require any extra system resources and is quick to compute. Paraphrasing by back-translation requires a machine-translation model that can translate to an intermediate language and back. This adds extra cost and processing time for unlabeled data which scales linearly with the amount of unlabeled data. For industry-scale NLU applications with massive amounts of data, dropout-based consistency training can provide performance gains over purely supervised methods with minimal additional resource overhead.

References

Philip Bachman, Ouais Alsharif, and Doina Precup. 2014. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27:3365–3373.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. MixMatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32.

Xavier Bouthillier, Kishore Konda, Pascal Vincent, and Roland Memisevic. 2015. Dropout as data augmentation. *arXiv preprint arXiv:1506.08700*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in NLP. *arXiv preprint arXiv:2106.07499*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1597–1607. PMLR.

Sumit. Chopra, Raia Hadsell, and Yan LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Tobias Falke, Markus Boese, Daniil Sorokin, Caglar Tirkaz, and Patrick Lehnen. 2020. Leveraging user paraphrasing behavior in dialog systems to automatically collect annotations for long-tail utterances. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 21–32, Online. International Committee on Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, 28:3546–3554.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, pages 5998–6008.

Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. 2019. Interpolation consistency training for semi-supervised learning. In *International Joint Conference on Artificial Intelligence*, pages 3635–3641.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of

Proceedings of Machine Learning Research, pages 9929–9939. PMLR.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In *International Conference on Learning Representations*.

Dazhi Zhao, Guozhu Yu, Peng Xu, and Maokang Luo. 2019. Equivalence between dropout and data augmentation: A mathematical check. *Neural networks*, 115:82—89.

A Appendix

A.1 Ablation studies

The training of our CT models depends on a few hyperparameters, including: training signal annealing (TSA) schedule, softmax temperature control, and a confidence threshold for computing consistency loss. We explored the impact of each hyperparameter on resulting model performance. For these experiments, we used the Dropout CT model trained 0.1% of labeled data. We did not train multiple models for each random data split.

	Top-1 relative change
Dropout CT*	-11.84%
confidence thresh= 0.6	-11.01%
confidence thresh = 0.3	-11.42%
confidence thresh = none	-32.37%
TSA schedule = log	-13.70%
TSA schedule = exp	-85.69%
TSA schedule = none	-14.22%
softmax temp = 0.7	-13.70%
softmax temp = 0.9	-12.87%
softmax temp = none	-11.94%

Table 3: Ablation studies related to confidence-based thresholding (confidence thresh), training-signal-annealing (TSA) schedule, and softmax temperature. In this table Dropout CT is the base model that each subsequent model modifies. We report the Dropout CT score only for the model trained on the same 0.1% data sample as used for the ablation-study experiments. All numbers are Top-1 accuracy relative changes compare to performance of baseline model trained with 100% labeled data. *For the base Dropout CT configuration, we used a linear TSA schedule, a consistency-loss softmax temperature of 0.85, and consistency-loss confidence threshold of 0.45.