Consistency Training with Virtual Adversarial Discrete Perturbation

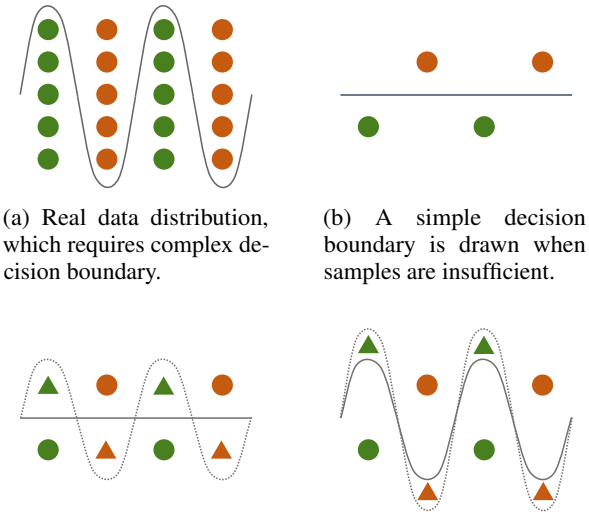
Jungsoo Park[1,3*] **Gyuwan Kim**[2*†] **Jaewoo Kang**[3]
[1]Clova AI, NAVER Corp. [2] University of California, Santa Barbara [3]Korea University
{jungsoo_park,kangj}@korea.ac.kr
gyuwankim@ucsb.edu

Abstract

Consistency training regularizes a model by enforcing predictions of original and perturbed inputs to be similar. Previous studies have proposed various augmentation methods for the perturbation but are limited in that they are agnostic to the training model. Thus, the perturbed samples may not aid in regularization due to their ease of classification from the model. In this context, we propose an augmentation method of adding a discrete noise that would incur the highest divergence between predictions. This virtual adversarial discrete noise obtained by replacing a small portion of tokens while keeping original semantics as much as possible efficiently pushes a training model's decision boundary. Experimental results show that our proposed method outperforms other consistency training baselines with text editing, paraphrasing, or a continuous noise on semi-supervised text classification tasks and a robustness benchmark[1].

1 Introduction

Building a natural language processing (NLP) system often requires an expensive process to collect a massive amount of labeled text data. Semi-supervised learning (SSL) (Chapelle et al., 2009) mitigates the requirement for such labeled data by exploiting the structure of unlabeled data. Among the SSL methods, the consistency training framework (Laine and Aila, 2017; Sajjadi et al., 2016) enforces a model to produce similar predictions of original and perturbed inputs. This method has several advantages over other training algorithms such as naively adding augmented samples into the training set (Wei and Zou, 2019; Ng et al., 2020) in that it provides a richer training signal than a one-hot label, and also applies to both labeled and unlabeled data (Xie et al., 2020).



(a) Real data distribution, which requires complex decision boundary.

(b) A simple decision boundary is drawn when samples are insufficient.

(c) Augmentations can push the decision boundary (dotted line) from the current one (bold line).

(d) Augmentations which are outside the current decision boundary enable further pushing it.

Figure 1: A simple illustration of the intuition behind our method is visualized in the two-dimensional space, where the augmented samples (triangle) would aid the training given a limited number of data (circle).

For perturbing a text while preserving its semantics, some approaches inject continuous noise to embedding vectors (Xie et al., 2017; Miyato et al., 2018), and others modify text itself in discrete fashion by edit operations (Kobayashi, 2018; Wei and Zou, 2019) or paraphrasing with back-translation (Sennrich et al., 2016; Edunov et al., 2018; Xie et al., 2020). However, adding continuous noise might not strongly regularize the training model, compared to diverse discrete noise-based augmentation methods (Ebrahimi et al., 2017; Cheng et al., 2019). Also, the augmentations with discrete noise are mostly black-box approaches based on simple rules or fixed models without access to the training model's internal states, having no control over output augmentations that would aid in the regularization of the training model. As seen in Fig. 1 (d), the

*Equal contribution.
†Work done while working at NAVER Clova & AI Lab
[1]Code repo: https://github.com/clovaai/vat-d

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5646 - 5656

augmented samples with similar semantics but that are outside the training model's decision boundary (*i.e.* adversarial) are the ones that would effectively regularize the model to fit into the complex real data distribution.

To this end, we explore virtual adversarial training with discrete token replacements (*VAT-D*). Our framework (1) first perturbs a given input text by replacing a small subset of tokens to *maximize the divergence between the original and the perturbed samples' model predictions (i.e., virtual adversarial)* while filtering tokens to replace for constraining the semantic similarity, and (2) train a model to minimize the divergence of the predictions of original and perturbed inputs.

VAT-D shares the advantages of virtual adversarial training (VAT) with continuous noise (Miyato et al., 2018) in that the perturbation is model-dependent, changing over the training time to approximate the augmented samples that would effectively push the decision boundary. On the other hand, VAT-D differs from VAT in that the search space is discrete rather than continuous, thus not constrained by the pre-defined norm on the embedding space. Our method relies on the training model's predictions which do not require label information, hence being the first work to successfully apply the adversarial training with perturbation on discrete space to the SSL framework.

Our proposed method empirically outperforms previous state-of-the-art methods on topic classification datasets (Chang et al., 2008; Mendes et al., 2012; Zhang et al., 2015) under various SSL scenarios. We additionally conduct experiments on ANLI robustness benchmark dataset (Nie et al., 2020) for testing the robustness when only labeled samples are given where the method improves over the RoBERTa-Large (Liu et al., 2019) by 8 points.

2 Background

We explain the concept of consistency training and VAT that our framework relies on.

Consistency Training Consistency training (Laine and Aila, 2017; Sajjadi et al., 2016) enforces models' predictions to be invariant when the input is perturbed. This regularization pushes the decision boundary to traverse a low-density region (Verma et al., 2019). The consistency loss is formally defined as

$$\mathcal{L}(\mathbf{x}, \mathbf{x}') = D[p(\cdot \mid \mathbf{x}), p(\cdot \mid \mathbf{x}')] \quad (1)$$

where D is a non-negative divergence metric between two probability distributions (*e.g.*, KL-divergence), $\mathbf{x}'$ is a perturbed sample from an input $\mathbf{x}$ by any transformation.

Virtual Adversarial Training VAT (Miyato et al., 2017, 2018) is a consistency training method, which perturbs a given input with continuous noise to maximize the divergence from the model's prediction of the original input. Such virtually adversarial examples effectively smooth the decision boundary compared to the random perturbation (Miyato et al., 2018). The formal definition of virtual adversarial samples is $\hat{\mathbf{x}} = \underset{\mathbf{x}' \in Neighbor(\mathbf{x})}{\operatorname{argmax}} \mathcal{L}(\mathbf{x}, \mathbf{x}')$ where the training objective is to minimize the $\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}})$. Miyato et al. (2017) perturbs input by injecting noise to the embedding space, where the constraint of the perturbation is ϵ-ball in L^p norm centered at $\mathbf{x}$, *i.e.* $Neighbor(\mathbf{x}) = \{\mathbf{x}' \mid \|\mathbf{x}' - \mathbf{x}\|_p \leq \epsilon\}$.

3 Method

We aim to generate a perturbed sample by adding discrete noise that incurs the highest divergence of the model's prediction logits from the original one without significant changes in its semantics. Our augmentation is made *on-the-fly* depending on the current model to push the decision boundary during training effectively.

Virtual Adversarial Discrete Noise We develop the consistency training framework by perturbing inputs with virtual adversarial discrete noise, called VAT-D. We want to perturb a given sentence $\mathbf{x} = (x_1, \ldots, x_M) \in V^M$ of sequence length M into a new sentence $\mathbf{x}' = (x'_1, \ldots, x'_M) \in V^M$ of the same length, where V is the word vocabulary. In contrast with the continuous case, we constrain that $\mathbf{x}'$ differs from $\mathbf{x}$ in only small portion of positions changing their surface forms, *i.e.* $Neighbor(x) = \{\|\mathbf{x}' - \mathbf{x}\|_H / M \leq \tau\}$ where H denotes hamming distance in the token-level and τ is the replacement ratio. In this work, we only focus on the replacement for simplicity.

Gradient Information The white-box approaches having an access to the training model's internal states, mostly rely on the gradient vectors of the loss function with respect to the input embeddings for finding adversarial discrete noise (Ebrahimi et al., 2017). However, for acquiring such gradient information under the framework of

consistency training as in Eq. 1, naively resorting to the linear approximation of the loss function with respect to the input embeddings like in previous works (Ebrahimi et al., 2017; Michel et al., 2019; Cheng et al., 2019) does not hold since the first-order term from Taylor expansion is zero when the label information is substituted to model's predictions (Miyato et al., 2018).

We bypass the obstacle by sharpening the distribution of original examples' predictions to enable the linear approximation. Sharpening the distribution makes high probabilities higher and lower probabilities lower while not changing their relative order. By sharpening the distribution of the original inputs' predictions, the first-order term does not result in zero, hence can be utilized for the approximation. This is because the modified divergence loss is not zero when $\mathbf{x}' = \mathbf{x}$ indicating the non-negative divergence is not necessarily minimum at $r = \mathbf{x}' - \mathbf{x} = 0$ (Note that the derivative of $f(x)$ is zero when the $f(x)$ is minimum at x). The optimizing objective of Eq. 1 is modified to

$$\tilde{\mathcal{L}}(\mathbf{x}, \mathbf{x}') = D[p^{sharp}(\cdot \mid \mathbf{x})), p(\cdot \mid \mathbf{x}')] \quad (2)$$

by sharpening the predicted distribution given an original input by the pre-defined temperature T as $p^{sharp}(\cdot \mid \mathbf{x}) = p(\cdot|\mathbf{x})^{\frac{1}{T}} / \left\| p(\cdot|\mathbf{x})^{\frac{1}{T}} \right\|_1$.

Virtual Adversarial Token Replacement Consequently, the optimization problem to find a virtual adversarial discrete perturbation changes to

$$\hat{\mathbf{x}} = \underset{\mathbf{x}' \in Neighbor(\mathbf{x})}{\operatorname{argmax}} \tilde{\mathcal{L}}(\mathbf{x}, \mathbf{x}').$$

Finally, we train the modified consistency loss function from Eq. 2 with obtained discrete perturbation. The replacement operation of m-th token x_m to the arbitrary token x can be written as $\delta(x_m, x) := e(x) - e(x_m)$, where $e(\cdot)$ denotes embedding look-up. We induce a virtual adversarial token by the following criteria (Ebrahimi et al., 2017; Michel et al., 2019; Cheng et al., 2019; Wallace et al., 2019; Park et al., 2020):

$$\hat{x}_m = \underset{x \in \text{top\_k}(x_m, V)}{\operatorname{argmax}} \delta(x_m, x)^\top \cdot g_{x_m} \quad (3)$$

$$\text{where } g_{x_m} = \nabla_{e(x_m)} \tilde{\mathcal{L}}(\mathbf{x}, \mathbf{x}')|_{\mathbf{x}'=\mathbf{x}}$$

$\mathbf{g}_{x_m}$ is the gradient vector of the sharpened consistency loss from Eq. 2 with respect to the m-th token. In brief, we replace the m-th original token x_m with one of the candidates x that approximately maximizes the consistency loss. We randomly select token indexes to perturb and replace them simultaneously. To bound the semantics similarity between the original sentence and the perturbed one, we use a masked language model (MLM) (Devlin et al., 2019; Liu et al., 2019) to restrict a set of possible candidates to replace x_m. We filter top-k candidates (Cheng et al., 2019), denoted as $\text{top\_k}(x_m, V)$, from the vocabulary having the highest MLM probability at position m when an original sentence x is given to the MLM. More training details are in Appendix A.

4 Experimental Setup

4.1 Dataset

We experiment on three topic classification datasets and Adversarial NLI (ANLI) (Nie et al., 2020). The former evaluate our method's effectiveness in SSL and the latter is for evaluating the robustness of the models under the standard supervised training framework. The three topic classification benchmarks consist of AG News (Zhang et al., 2015), DBpedia (Mendes et al., 2012), and YAHOO! Answers (Chang et al., 2008). We follow the experimental setting from Chen et al. (2020a), where we train with a limited number of labeled data in diverse settings, namely, 10, 200, 2500 per class. We randomly sample the labeled, unlabeled, and development set and report the performance on the official test set. For producing the confident results, we report the average of five different seeds' distinct runs.

As for ANLI, we train the model with two different settings, training with only the ANLI dataset or additionally training with other NLI datasets, including SNLI (Bowman et al., 2015), MNLI (Williams et al., 2017), and FEVER (Thorne et al., 2018) following the original work (Nie et al., 2020). Further details are in Appendix B.

4.2 Baseline

We compare our method with various baselines of the perturbation methods including EDA (Wei and Zou, 2019), UDA (Xie et al., 2020), VAT (Miyato et al., 2017, 2018) for the topic classification SSL task. For the ANLI dataset, we compare with the baselines (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Jiang et al., 2020) that have reported numbers on the official validation and test set. More details are in Appendix C.

Method	AG_NEWS			YAHOO!			DBpedia		
	10	*200*	*2500*	*10*	*200*	*2500*	*10*	*200*	*2500*
BERT (Devlin et al., 2019)	79.4	88.6	91.6	58.2	70.1	73.9	97.8	98.8	99.1
EDA (Wei and Zou, 2019)	83.8	88.9	91.8	62.0	70.6	73.8	98.4	98.8	99.1
UDA (Xie et al., 2020)	83.8	88.5	91.6	62.0	70.4	73.7	98.2	98.8	99.1
VAT (Miyato et al., 2017)	82.3	88.9	91.8	62.4	70.7	74.1	98.4	98.8	99.1
VAT-D	**86.2**	**90.0**	**92.3**	**65.3**	**71.7**	74.1	98.4	**99.0**	**99.2**

Table 1: Accuracy on topic classification datasets under the various SSL settings. 10, 200, 2500 denote the number of labeled samples per class used during training. We average five different runs with a differently indexed dataset to show the significance (Dror et al., 2018). The numbers in the bold denote the best score.

Method	Dev				Test			
	A1	A2	A3	ALL	A1	A2	A3	ALL
MNLI + SNLI + ANLI + FEVER								
BERT(Nie et al., 2020)	57.4	48.3	43.5	49.3	-	-	-	44.2
XLnet (Nie et al., 2020)	67.6	50.7	48.3	55.1	-	-	-	52.0
RoBERTa (Nie et al., 2020)	73.8	48.9	44.4	53.7	-	-	-	49.7
SMART (Jiang et al., 2020)	**74.5**	50.9	47.6	57.1	**72.4**	49.8	**50.3**	57.1
VAT-D	**74.5**	**54.2**	**50.8**	**59.2**	**72.4**	**51.8**	49.5	**57.4**
ANLI								
RoBERTa (Nie et al., 2020)	71.3	43.3	43.0	51.9	-	-	-	-
SMART (Jiang et al., 2020)	74.2	49.5	49.2	57.1	**72.4**	50.3	49.5	56.9
VAT-D	**74.8**	**52.1**	**51.1**	**58.8**	72.1	**51.4**	**51.7**	**57.9**

Table 2: Accuracy on the ANLI benchmark. The numbers of the baselines are from the original papers (Nie et al., 2020; Jiang et al., 2020). The upper section is for training with all the NLI datasets, and the bottom is for training with only the ANLI.

4.3 Training Details

We exploit the unlabeled data from the topic classification datasets and the labeled data from the ANLI for consistency loss. Throughout the experiments, we set the replacement ratio τ as 0.25 and top-k as 10. We sharpen the predictions with T as 0.5 for topic classification datasets (including baselines) and 0.75 for the ANLI.

5 Experimental Results

5.1 Semi Supervised Text Classification

Table 1 shows the experimental results on topic classification datasets under SSL setup. Our method outperforms the baselines by up to 7.1 points from the BERT model finetuned with standard cross-entropy loss and 2.9 points from other methods utilizing the consistency regularization loss. The accuracy gained from the proposed method from the baselines, especially when the number of labeled samples is limited. However, since all the methods have already achieved high accuracy in the DBpedia, the difference among methods is not significant.

Among the baselines, VAT (Miyato et al., 2017, 2018) performs reasonably well. The finding supports the claim that a transformation during consistency training should be done with regard to the training model.

5.2 Adversarial Natural Language Inference

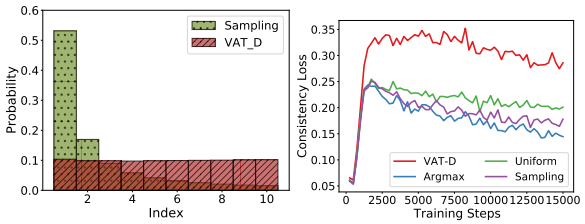
Table 2 shows the experimental results on the ANLI dataset with different training settings: training with all the NLI datasets, or training with only the ANLI dataset. Our method improves over baselines, including RoBERTa-Large (Liu et al., 2019) and SMART (Jiang et al., 2020) in both settings. Specifically, our method improves on an average of 8.0 points in the test set from training with cross-entropy loss only. Compared to SMART, which combines smoothness regularization, i.e., a variation of VAT, and Bregman proximal point optimization for finetuning, our method outperforms it on an average of 1 point from the test set without using other techniques such as Bregman proximal point optimization.

6 Effectiveness of the White-box Search

Our central intuition behind the proposed method is to generate the augmented samples concerning the model, i.e., vulnerable to the model. This section further conducts an ablation study on whether such virtual adversarial search is crucial in discrete

Method	AG_NEWS			YAHOO!		
	10	*200*	*2500*	*10*	*200*	*2500*
VAT-D	**86.2**	**89.8**	**92.3**	**65.3**	**71.7**	**74.1**
Uniform	83.8	89.3	91.8	63.2	70.8	73.8
Argmax	83.2	89.0	91.9	63.7	70.9	73.7
Sampling	84.8	89.3	91.8	63.5	70.9	73.8

Table 3: Accuracy according to different sampling strategies from top-K candidates.



(a) Top-k Distribution (b) Consistency Loss

Figure 2: Index distribution from the top-k candidates sorted by MLM scores (a) and consistency loss of different sampling strategies during training (b).

space. We select the token among the top-K candidates that would incur the highest divergence from the model's prediction. We compare with other sampling strategies among the top-K candidates, namely, uniform sampling (Uniform), selecting the token with maximum MLM probability (Argmax), and sampling from MLM probabilities (Sampling). We match other training details except for the sampling strategy for a fair comparison.

Table 3 illustrates the result of comparisons. The virtual adversarial search among candidates outperforms other search strategies in discrete space, especially when the number of labeled samples is limited. The result demonstrates that the virtual adversarial search is indeed the crucial component during perturbation. Furthermore, Fig. 2 shows the indexes that our method selected from top-k distribution (sampling from YAHOO! dataset) during training. The distribution of indexes selected from our method resembles the uniform distribution; however, as in the loss plot from Fig. 2, our method searches for the diverse yet adversarial candidates to the model, i.e., incurring high divergence.

7 Related Works

Consistency Regularization Consistency regularization (Laine and Aila, 2017; Sajjadi et al., 2016) has been mainly explored in the context of SSL (Chapelle et al., 2009; Oliver et al., 2018). A line of research in text-domain (Miyato et al., 2017; Clark et al., 2018; Xie et al., 2020; Miyato et al., 2018; Jiang et al., 2020; Asai and Hajishirzi, 2020) explored the idea. Existing studies explored varying perturbation methods. Injecting norm-constrained continuous noise to the embedding space (Miyato et al., 2017; Jiang et al., 2020; Liu et al., 2020; Chen et al., 2020b; Sato et al., 2019) and directly perturbing the text (Clark et al., 2018; Minervini and Riedel, 2019; Li et al., 2019; Xie et al., 2020; Asai and Hajishirzi, 2020) via discrete noise are the primary approaches for the perturbation. Our method perturbs the sentence by the discrete noise, yet the noise is generated concerning the training model.

Adversarial Training Our method extends the white-box-based adversarial training framework (Goodfellow et al., 2014; Madry et al., 2018), which has recently been explored widely in NLP (Miyato et al., 2017; Ebrahimi et al., 2017; Michel et al., 2019; Wang et al., 2019; Zhu et al., 2020; Jiang et al., 2020; Liu et al., 2020). Cheng et al. (2019) use adversarial training on machine translation by discrete word replacements relying on the label information, so not applicable to SSL different from ours. There are also black-box approaches for generating the adversarial attacks or test sets (Jia and Liang, 2017; Alzantot et al., 2018; Ribeiro et al., 2018, 2019; Gardner et al., 2020) to evaluate the vulnerability of the NLP models, unlike our method, which utilizes gradient information during training. Li et al. (2020); Garg and Ramakrishnan (2020); Li et al. (2021) perturb input using MLMs similar to ours but designed for an attack so inefficient for adversarial training.

Data Augmentation Synthetically generated training examples are utilized to augment an existing dataset (Feng et al., 2021). Existing word-level augmentation methods (Zhang et al., 2015; Xie et al., 2017; Wei and Zou, 2019) are based on heuristics. Mixup-based methods (Zhang et al., 2018) interpolate input texts in hidden embeddings (Chen et al., 2020a; Guo et al., 2019) or input-level (Yoon et al., 2021; Kim et al., 2021). Other methods include utilizing back-translation models (Sennrich et al., 2016; Xie et al., 2020), contextual language models (Kobayashi, 2018; Wu et al., 2019), or generative models (Anaby-Tavor et al., 2020; Yang et al., 2020). Unlike previous works, our method is subject to the training model, thus approximating the augmented points, efficiently filling in gaps from the training data.

Acknowledgements

We thank Jinhyuk Lee, Jaewook Kang, and Sungdong Kim for the discussion and feedback on the paper. We also thank the members of the Conversation team in Naver CLOVA for active discussion. This research was supported by National Research Foundation of Korea (NRF-2020R1A2C3010638) and the Ministry of Science and ICT, Korea, under the ICT Creative Consilience program (IITP-2022-2020-0-01819).

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *EMNLP*.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *AAAI*.

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *ACL*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *AAAI*.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3).

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020a. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *ACL*.

Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu. 2020b. Seqvat: Virtual adversarial training for semi-supervised sequence labeling. In *ACL*.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *ACL*.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *ACL*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. Hotflip: White-box adversarial examples for text classification. In *ACL*.

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

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *ACL-Findings*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *EMNLP-Findings*.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *EMNLP*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. In *ICLR*.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *ACL*.

Yekyung Kim, Seohyeong Jeong, and Kyunghyun Cho. 2021. Linda: Unsupervised learning to interpolate in natural language processing. *arXiv preprint arXiv:2112.13969*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *NAACL*.

Samuli Laine and Timo Aila. 2017. Temporal ensembling for semi-supervised learning. In *ICLR*.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. Contextualized perturbation for textual adversarial attack. In *ACL*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. In *EMNLP*.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *EMNLP*.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Pablo N Mendes, Max Jakob, and Christian Bizer. 2012. Dbpedia: A multilingual cross-domain knowledge base. In *LREC*. Citeseer.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *NAACL*.

Pasquale Minervini and Sebastian Riedel. 2019. Adversarially regularising neural nli models to integrate logical background knowledge. In *CONLL*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. *ICLR*.

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).

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *EMNLP*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *ACL*.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL-HLT: Demonstrations*.

Jungsoo Park, Mujeen Sung, Jinhyuk Lee, and Jaewoo Kang. 2020. Adversarial subword regularization for robust neural machine translation. In *EMNLP*.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *ACL*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *ACL*.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*.

Motoki Sato, Jun Suzuki, and Shun Kiyono. 2019. Effective adversarial regularization for neural machine translation. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *ACL*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *NAACL*.

Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. 2019. Interpolation consistency training for semi-supervised learning. *IJICAI*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *EMNLP*.

Dilin Wang, Chengyue Gong, and Qiang Liu. 2019. Improving neural language modeling via adversarial training. In *ICLR*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*. Springer.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *NeurIPS*, 33.

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. In *ICLR*.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. G-daug: Generative data augmentation for commonsense reasoning. In *EMNLP-Findings*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Soyoung Yoon, Gyuwan Kim, and Kyumin Park. 2021. Ssmix: Saliency-based span mixup for text classification. In *ACL-Findings*, pages 3225–3234.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *ICLR*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NeurIPS*.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for language understanding. In *ICLR*.

Algorithm 1: VAT_D Module

Input : input sentence $\mathbf{x}$, index to perturb $\mathbf{I}$
Output : perturbed sentence $\hat{\mathbf{x}}$
Function *VAT_D*($\mathbf{x}$, $\mathbf{I}$):
  $\hat{\mathbf{x}} \leftarrow \mathbf{x}$
  for $m \in \mathbf{I}$ **do**
    $g_{x_m} \leftarrow \nabla_{e(x_m)} \tilde{\mathcal{L}}(\mathbf{x}, \mathbf{x}')|_{\mathbf{x}'=\mathbf{x}}$
    $\hat{x}_m \leftarrow \underset{x \in \text{top\_k}(x_m, V)}{\text{argmax}} \; \delta(x_m, x)^\top \cdot g_{x_m}$
    Replace m-th token of $\hat{\mathbf{x}}$ to $\hat{x}_m$
return $\hat{\mathbf{x}}$

A Training Details

Alg. 1 illustrates the procedure (VAT_D) to acquire virtual adversarial tokens with the modified consistency loss. We randomly select token indexes to perturb I, subject to the length of the sentence. Considering multiple substitutions, an exhaustive search over all possible combinations to find the optimal one is computationally intractable. For efficient generation during each training step, we replace multiple tokens simultaneously instead of greedy search or beam search, which has shown to work considerably well in previous works (Ebrahimi et al., 2017; Cheng et al., 2019). During training, the models are optimized with standard cross-entropy and consistency loss with an equal weight where we utilize KL-Divergence as the divergence D. Our method takes approximately 2.5 times the standard training whereas other baselines (e.g., EDA, Back-translation) take about 1.7 times the standard training. We utilize P40 for training the SSL experiments and V100 for the ANLI task.

In our preliminary experiment, utilizing the MLM with masking was worse than that without masking, similar to Li et al. (2020). While utilizing the MLM for filtering top-k candidates, we empirically verified that not applying masking operations to the sentence achieved better performance than doing so. We conjecture that the loss of information when applying masking operation has evoked the perturbed samples to significantly deviate from the original ones, resulting in a degradation in performance. The finding matches that of Li et al. (2020). Thus we do not apply masking operations throughout the experiments. Moreover, we do not fine-tune the *off-the-shelf* MLM on the training corpus but only the classification model, which is to ensure a fair comparison with other augmentation baselines.

Dataset	Genre	Class	Unlabel	Dev	Test
AG_NEWS	News	4	20k	20k	19k
YAHOO!	QA	10	50k	20k	60k
DBPedia	Wikipedia	14	70k	20k	50k

Table B.1: Data statistics for the topic classification datasets following the experimental setting from Chen et al. (2020a).

Dataset	Genre	Train	Dev	Test
A1	Wikipedia	17k	1k	1k
A2	Wikipedia	45k	1k	1k
A3	Various	100k	1.2k	1.2k
ANLI	Various	162k	3.2k	3.2k
MNLI	Various	392k	-	-
Fever	Wikipedia	208k	-	-
SNLI	Image Captions	549k	-	-

Table B.2: Data statistics for the ANLI with three rounds (A1-A3) and concerning NLI datasets for the training.

B Further Details on Data

The dataset statistics and split information regarding topic classification tasks and ANLI is presented in Table B.1 and Table B.2.

ANLI (Nie et al., 2020) is an NLI testbed recently introduced for evaluating the robustness of the models in natural language understanding. The dataset consists of three rounds (A1-A3), each consisting of a train-dev-test set with increasing difficulty, where the data is generated by human-and-model-in-the-loop fashion to fool the strong pre-trained models (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019).

C Further Details on Baselines

For the SSL setup, we use the following baselines:

BERT (Devlin et al., 2019) We use the pre-trained BERT-base-uncased model and finetune it for the classification dataset using only standard cross-entropy loss.

EDA (Wei and Zou, 2019) EDA is a simple data augmentation strategy based on word unit operations such as synonym replacement or deletion. We perturb the unlabeled samples using EDA[2] and exploit them for consistency training.

UDA (Xie et al., 2020) UDA paraphrases the sentence using the back-translation. We employ the WMT-19 DE↔EN model from fairseq[3] (Ott et al.,

[2] https://github.com/jasonwei20/eda_nlp
[3] https://github.com/pytorch/fairseq

2019) to do the back-translation on unlabeled samples, and exploit them for consistency training.

VAT (Miyato et al., 2017, 2018) We re-implement VAT where we apply the consistency loss to the unlabeled samples.

D Augmentation Quality

We present some augmentation samples in Table D.1 from three topic-classification datasets. As presented in the table, the augmentation samples moderately modify some tokens from the original sentence following the original context.

However, since we are decoding multiple tokens at a same time, some samples are shown to be ungrammatical (e.g., *is* $\rightarrow$ *will* instead of *will be*). Moreover, if the chosen token to be modified are entities, the augmentation sample can sometimes change the information presented in the sentence (e.g., *Patryk Dominik* $\rightarrow$ *Patryk Deinik*). However, since we are solving the task of the closed-domain topic classification task, the problems didn't matter much in this setting. If we are to solve the knowledge-intensive task, we would have to consider other filtering modules for not changing the entities.

Source	Sample
AG_NEWS	
Original	Turkey agonized over pressure to recognize cyprus in the last hurdle to an historic agreement
Augmentation	Turkey agonized over pressure to recognize cyprus in the final hurdle of an historic deal
Original	Rockets struck a baghdad hotel housing foreign contractors and journalists late thursday
Augmentation	Rockets hit a baghdad hotel housing visiting contractors and journalists late thursday
Original	Ten people were injured yesterday when a bomb exploded outside the Indonesian embassy in Paris
Augmentation	Ten civilians were injured yesterday when a bomb exploded outside the Jakarta embassy in Paris
Original	Pakistan authorities are putting the city of Karachi on ... for an al - qaida strike after its forces killed a top terror suspect
Augmentation	Pakistan governments are putting the city of karachi of ... for an al - qaida bomb after its members killed one top terrorism suspect
YAHOO!	
Original	How can guests get sound security under wireless internet environment at hotel ?
Augmentation	How can visitors get sound security under wireless internet environment at hotel ...
Original	Can you find some ones screen name by using there real name ? ? ? yes
Augmentation	Could you find other ones screen name by using there real surname ? : ? yes
Original	What is the perfect gift for my girlfriends b - day ? ? ? (information in here about her) ? she loves to : ride your black sport bike
Augmentation	What will the perfect gift for my girlfriends b - day ? ? ? (information here here of her) ? she wants to : ride your black racing bike
Original	Purpose of administration and it department to a business ? i work in it ... without us , companies would be at a standstill
Augmentation	Purpose of administration and it department to a corporation ? i work in that ... without us of companies would be near a stands market
DBpedia	
Original	Patryk Dominik Sztyber (born 4 august 1979 in Opoczno) stage name seth is a Polish heavy metal musician
Augmentation	Patryk Deinik Sztybor (birth 8 august 1979 in Opoczno) stage name seth is a Warsaw heavy metal musician
Original	Twill is a quarterly magazine published between Paris and Milan. It has an international readership
Augmentation	Twill is the quarterly magazine printed between Paris and Milan. It has an international readers range
Original	The pond creek station located east of Wallace Kansas ... is a two - story frame building that was a stagecoach station built 1865
Augmentation	The lake branch station built outside to Wallace Kansas ...is a two - story frame building that been a stagecoach station designed 1865
Original	Until we have wings is an album by randy stonehill released in 1990 on myrrh records
Augmenatation	Until we have wings is an album by randy stonehill published mid 1989 on myrrh records

Table D.1: Generated augmentation examples from our method along with original samples