# Virtual Data Augmentation: A Robust and General Framework for Fine-tuning Pre-trained Models

**Kun Zhou**[2,4], **Wayne Xin Zhao**[1,4†], **Sirui Wang**[3], **Fuzheng Zhang**[3],
**Wei Wu**[3] and **Ji-Rong Wen**[1,2,4]

[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]School of Information, Renmin University of China. [3]Meituan Inc., Beijing, China
[4]Beijing Key Laboratory of Big Data Management and Analysis Methods
`francis_kun_zhou@163.com, batmanfly@gmail.com`
`{wangsirui, zhangfuzheng, wuwei30}@meituan.com, jrwen@ruc.edu.cn`

## Abstract

Recent works have shown that powerful pre-trained language models (PLM) can be fooled by small perturbations or intentional attacks. To solve this issue, various data augmentation techniques are proposed to improve the robustness of PLMs. However, it is still challenging to augment semantically relevant examples with sufficient diversity. In this work, we present Virtual Data Augmentation (VDA), a general framework for robustly fine-tuning PLMs. Based on the original token embeddings, we construct a multinomial mixture for augmenting virtual data embeddings, where a masked language model guarantees the semantic relevance and the Gaussian noise provides the augmentation diversity. Furthermore, a regularized training strategy is proposed to balance the two aspects. Extensive experiments on six datasets show that our approach is able to improve the robustness of PLMs and alleviate the performance degradation under adversarial attacks. Our codes and data are publicly available at `https://github.com/RUCAIBox/VDA`.

## 1 Introduction

Recently, pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have achieved remarkable success in various natural language processing (NLP) tasks (Rajpurkar et al., 2016; Wang et al., 2019; Zhou et al., 2020b). As a general and effective approach, fine-tuning PLMs on specific datasets has become the mainstream paradigm for developing NLP applications. Despite the success, researchers have found that PLMs can be easily fooled by adversarial attacks (Jin et al., 2020; Li et al., 2020b). Although encapsulated into a black box, these attack strategies can detect the vulnerabilities of a PLM via intentional queries (He et al., 2021; Li et al., 2020a), and then add small perturbations (*e.g.,* synonyms substitution) into the input texts for misleading PLMs to incorrect predictions.

As found in previous works (Schmidt et al., 2018; Yin et al., 2019; Jiang et al., 2020), a possible reason of the vulnerability is that these PLMs do not generalize well on semantic neighborhood around each example in the representation space. To solve this issue, adversarial data augmentation (ADA) methods (Jia and Liang, 2017; Wang and Bansal, 2018; Michel et al., 2019) have been proposed by revising original data to augment attack-related data for training. However, due to the discrete nature of language, it is challenging to generate *semantically relevant* and *sufficiently diverse* augmentations. Although attempts by leveraging expert knowledge (Ren et al., 2019; Li et al., 2019b) and victim models (Jin et al., 2020; Li et al., 2020b) have achieved better performance, their generalizability and flexibility is highly limited.

Recently, virtual adversarial training (Miyato et al., 2017; Madry et al., 2018) is applied to various NLP models for improving the performance and robustness (Zhu et al., 2020; Jiang et al., 2020), which usually generates gradient-based perturbation on the embedding space as virtual adversarial samples. However, it is hard to explicitly constrain the gradient-based perturbation within the same semantic space as the original sample. In addition, unlike attacks in computer vision (Zheng et al., 2016; Miyato et al., 2019), textual adversarial attacks are discrete (*e.g.,* word replacement) and are hard to be captured by gradient-based perturbations.

To solve these challenges, we propose Virtual Data Augmentation (VDA), a robust and general framework for fine-tuning pre-trained models. Our idea is to generate data augmentations at the embedding layer of PLMs. To guarantee semantic relevance, we consider a multinomial mixture of the original token embeddings as the augmented embedding for each position of the input. In the
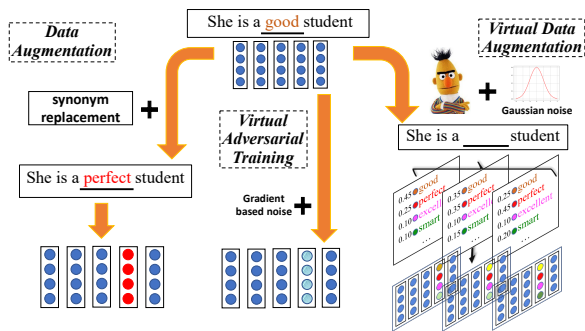
---

† Corresponding author

Figure 1: A comparison among synonym replacement based data augmentation methods, virtual adversarial training and our VDA framework.

mixture, each token embedding is weighted according to its likelihood estimated by a masked language model conditioned on the input. To provide sufficient diversity, we further incorporate Gaussian noise in the above multinomial mixture, which enhances the randomness of the augmentations. As shown in Figure 1, for a target token *"good"*, we first predict the substitution probabilities of candidate tokens via a masked language model, then inject the Gaussian noise to produce multiple multinomial mixtures. After that, we aggregate the candidate embeddings with the multinomial mixtures to generate new embeddings (virtual data embeddings) to replace the original embedding of *"good"*.

There are two major advantages to our VDA approach. First, with the original token embeddings as the representation basis, the augmented embeddings stay close to the existing embeddings, which avoids the unexpected drift of semantic space. Second, with the injected Gaussian noise, we are able to generate diverse variations for augmentations. In order to enhance the relevance with the given injected Gaussian noise, we further design a regularized training strategy that guides the learning of the augmented virtual data towards the original predictions of PLMs. In this way, our approach has considered both semantic relevance and sufficient diversity. Besides, since VDA only revises the input embeddings, it is agnostic to downstream tasks, model architectures and learning strategies.

To evaluate the effectiveness of our proposed VDA framework, we construct extensive experiments on six datasets. Results show that VDA can boost the robustness of all the baseline models without performance degradation. We also find that our approach can be further improved by combining it with traditional adversarial data augmentation.

Our contributions are summarized as follows:

• We propose a new data augmentation framework for resisting discrete adversarial attacks on PLMs, which is general to improve the robustness of various PLMs on downstream tasks.

• Our approach utilizes a masked language model with Gaussian noise to augment virtual examples for improving the robustness, and also adopts regularized training to further guarantee the semantic relevance and diversity.

• Extensive experiments on six datasets have demonstrated that the proposed approach is able to effectively improve the robustness of PLMs, which can be further improved by combining with existing adversarial data augmentation strategies.

## 2 Related Work

We review the related work in the following three aspects.

### 2.1 Adversarial Attack in NLP

Inspired by the success in compute vision (Goodfellow et al., 2015; Kurakin et al., 2017), adversarial attack in NLP tasks has become an emerging research topic in recent years (Gao et al., 2018; Yang et al., 2020; Chen et al., 2020). Early works usually adopt heuristic rules to revise the input text for producing adversarial samples, including character modification (Ebrahimi et al., 2018), synonyms replacement (Alzantot et al., 2018), word insertion or deletion (Zhang et al., 2019). However, with the revolution of large-scale PLMs, these attack strategies can be defended (Jones et al., 2020; Gui et al., 2021; Zhou et al., 2020a) to some extent. To attack PLMs, TextFooler (Jin et al., 2020) designs an attack algorithm to revise the input data and queries the PLM several times to find important words for replacement, which greatly reduces the accuracy of BERT. Following it, recent works (Li et al., 2020b; He et al., 2021) continuously improve the quality of the adversarial samples and the attack success ratio. In our approach, we consider improving the robustness of PLMs against these adversarial attack methods via a new fine-tuning framework VDA.

### 2.2 Data Augmentation

Data augmentation has been extensively studied in NLP tasks for improving the robustness (Wang and Yang, 2015; Fadaee et al., 2017; Wei and Zou, 2019). Similar to adversarial attack, early works mostly try heuristic rules to revise the in-

put data for augmentation, such as synonym replacement (Wang and Bansal, 2018), grammar induction (Min et al., 2020), word insert and delete (Wei and Zou, 2019). With the development of text generation techniques, back translation (Xie et al., 2020; Ribeiro et al., 2018) and variant autoencoder (Wang et al., 2020; Li et al., 2019c) are used to augment new data. Besides, a surge of works (Hou et al., 2018; Li et al., 2019a; Zhou et al., 2019) focus on augmentation for specific tasks with special rules or models. Although they perform well, these methods have lost the generality. In this paper, we propose a new data augmentation framework VDA that utilizes a masked language model with Gaussian noise to augment virtual examples for improving the robustness. our VDA is agnostic to downstream tasks, model architectures and learning strategies.

## 2.3 Virtual Adversarial Training

To improve the robustness of neural networks against adversarial examples, virtual adversarial training (VAT) (Miyato et al., 2015; Kurakin et al., 2017; Qin et al., 2019) has been widely used in compute vision. It formulates a class of adversarial training algorithms into solving a minimax problem, which can be achieved reliably through multiple projected gradient ascent steps (Qin et al., 2019). Recently, VAT has shown its effectiveness in NLP tasks, where the gradient-based noise is able to improve the performance and smoothness of the pre-trained models (Zhu et al., 2020; Jiang et al., 2020). However, due to the discrete nature of language, it has been shown that VAT methods are not very effective in defending against adversarial attacks (Si et al., 2020; Li and Qiu, 2021).

## 3 Preliminary

This work seeks to improve the fine-tuning performance of pre-trained language models (PLM), in that the fine-tuned model will become more robust to data permutations or attacks. Specially, we take the text classification task as an example task to illustrate our approach, where a set of $n$ labeled texts $\{\langle x_i, y_i \rangle\}$ are available. Each labeled text consists of a text $x_i$ and a label $y_i$ from the label set $\mathcal{Y}$. We refer to the adversarial example generated from a text $x_i$ as adversarial text, denoted by $\hat{x}_i$. The purpose of adversarial examples is to enhance the model robustness in resisting intentional data perturbations or attacks.

Let $f$ denote a PLM parameterized by $\theta$. Following (Jia and Liang, 2017; Michel et al., 2019), we incorporate adversarial examples to improve the fine-tuning of PLMs. To conduct the adversarial learning, we formulate the learning objective as follows:

$$\arg\min_\theta \sum_{i=1}^n L_{\mathrm{c}}(f(x_i), y_i) + \lambda \sum_{j=1}^m L_{\mathrm{reg}}(f(x_j), f(\hat{x}_j)),$$

$$(1)$$

where $m$ is the number of adversarial texts that we use, $\lambda$ is a trade-off parameter, $L_{\mathrm{c}}$ and $L_{\mathrm{reg}}$ try to minimize the classification loss and reduce the prediction difference between original and adversarial texts, respectively.

For the PLM $f$, we assume that it is already pre-trained on general-purpose large-scale text data, we would like to fine-tune its parameter $\theta$ based on some downstream task. The PLMs are usually developed based on multi-layered Transformer architecture such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), where a sequence of tokens will be encoded into a sequence of contextual representations. Here, we take the representation of the first token (*i.e.,* `[CLS]`) as the input of the classifier, and optimize the classification performance with the cross-entropy loss.

## 4 Our Approach

In this section, we describe our proposed framework *Virtual Data Augmentation (VDA)* for robustly fine-tuning PLMs. Our framework consists of two important ingredients, namely embedding augmentation and regularized training.

## 4.1 Embedding Augmentation

To improve the model robustness, a good adversarial example should adhere to the original semantic space, as well as incorporate sufficient variations in meanings. However, existing studies cannot make a good trade-off between the two aspects.

Considering this difficulty, we generate adversarial texts at the embedding layer of PLMs. For adversarial training, continuous embeddings are easier to optimize and can encode more semantic variations than discrete tokens. The key idea of embedding augmentation is inspired by the word replacement strategy in previous data augmentation methods (Kobayashi, 2018; Wei and Zou, 2019). Instead of selecting some tokens for replacement, we use an augmented embedding to replace the original contextual embedding of a specific token
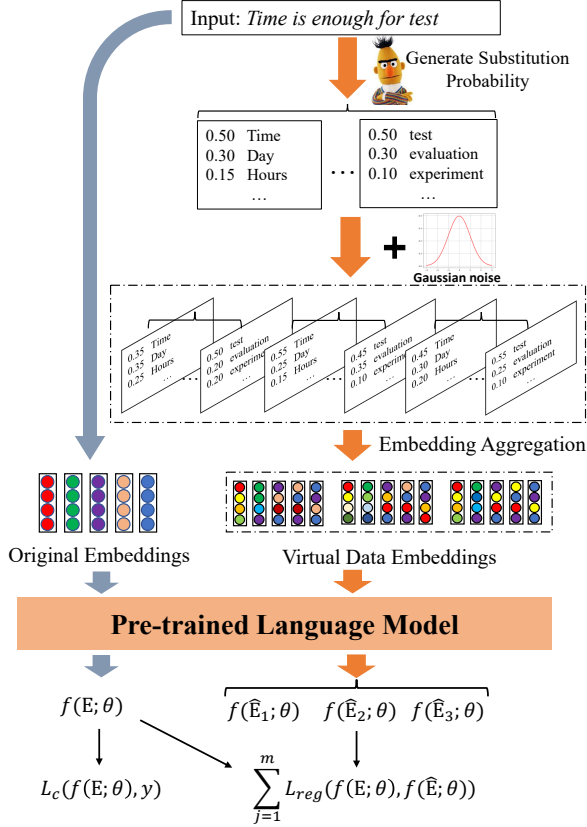
Figure 2: Illustration of our framework VDA. We show the case that we generate three virtual examples for the input sentence.

in the input sentence. To adhere to the original semantic space, the augmented embedding is derived by a probabilistic mixture of the embeddings of the vocabulary terms, where each term is weighted according to its *substitution probability* (*i.e.,* replacing the original token with the candidate term) calculated by a masked language model (MLM).

To simplify our presentation, we only discuss the augmentation for a specific token $\tilde{w}$ from an input sentence $S$. The same procedure will be applied to each position of the original sentence $S$. Specially, we utilize the MLM to evaluate the substitution probabilities of all the terms in the vocabulary. For each chosen token, we predict its probability to be replaced by other words in the whole vocabulary via MLM, denoted as $p(\hat{w}_i|S)$. Finally, we obtain the substitution probabilities of all the terms as

$$\{p(\hat{w}_1|S), \ldots, p(\hat{w}_V|S)\}, \quad (2)$$

where $V$ is the vocabulary size. Different from previous masked prediction (Devlin et al., 2019), we do not mask the chosen token but also keep it as the input to compute the substitution probabilities. In this way, we aim to generate very relevant

embeddings for augmentation. Such a strategy is also very efficient in practice, since it no longer performs the costly mask-and-completion operations for each token.

To augment diverse virtual data embeddings, we further draw a random noise from the Gaussian distribution as

$$\epsilon \sim \mathcal{N}(0, \sigma^2), \quad (3)$$

where the randomness can be controlled by the standard variance $\sigma$. By mixing the random noise with the substitution probabilities, we can produce multiple different probability distributions for each instance as

$$p'(\hat{w}_i|S) = \texttt{softmax}(p(\hat{w}_i|S) + \epsilon). \quad (4)$$

Then, for each target token $\tilde{w}$, we obtain its corresponding substituted embedding by aggregating the token embedding matrix according to the noised substitution probability as

$$\hat{\mathbf{e}}_{\tilde{w}} = \mathbf{p}_{\tilde{w}} \cdot \mathbf{M}_E, \quad (5)$$

where $\mathbf{p}_{\tilde{w}} = \{p'(\hat{w}_i|S)\}_{i=1}^V$, and $\mathbf{M}_E \in \mathbb{R}^{V \times d}$ is the token embedding matrix from the MLM. Note that by using the output of MLM, our approach can augment more "real" embeddings from the semantic space spanned by original token embeddings. Besides, mixing Gaussian noise brings additional semantic diversity for augmentation.

### 4.2 Regularized Training

The above augmentation strategy is able to enhance the semantic variations by continuous embeddings. However, augmented data is likely to incorporate unexpected semantic drift in representations. To further improve the model robustness, instead of directly using the augmented embeddings as positive examples, we propose a regularized training strategy to prevent large changes between the predictions given real and augmented embeddings. Formally, given the original data point $(\mathbf{E}_i, y_i)$ and the augmented virtual data $\hat{\mathbf{E}}_i$, where $\mathbf{E}_i$ and $\hat{\mathbf{E}}_i$ denote the original embeddings and augmented embeddings of the instance respectively, we set the regularization loss in Equation 1:

$$L_{\text{reg}}(\theta) = \frac{1}{k} \sum_{i=1}^{k} D_{sKL}\bigg(f(\mathbf{E}_i; \theta), f(\hat{\mathbf{E}}_i; \theta)\bigg),$$

$$(6)$$

where $D_{sKL}$ is the symmetric KL-divergence, $k$ denotes the number of augmented examples. The

3878

**Algorithm 1** The framework of VDA.

---

**Require:** Pre-trained model $f(.)$, a pre-trained masked language model $f_{MLM}(.)$, Gaussian distribution $\mathcal{N}(0, \sigma^2)$, virtual data sampling number $k$, training epoch $m$.

1: **Input:** The training data $\{\langle x_i, y_i \rangle\}$.
2: **Output:** The fine-tuned parameters $\theta$.
3: **for** $i = 1 \ldots m$ **do**
4:     **for** minibatch $B \in \{\langle x_i, y_i \rangle\}$ **do**
5:         Tokenize input sentences in $B$ into $\{w_1, ..., w_m\}$.
6:         Generate the substitution probability of all tokens.
7:         **for** $j = 1 \ldots k$ **do**
8:             Sample $\epsilon$ from $N(0, \sigma^2)$.
9:             Produce $p'(\hat{w}_i|S)$ using Eq. 4.
10:            Augment virtual data using Eq. 5.
11:            Optimize $\theta$ using Eq. 1.
12:         **end for**
13:     **end for**
14: **end for**
15: **return** $\theta$

---

| Task | Data | Train | Dev | Test |
|---|---|---|---|---|
| **Sentence Classification** | Yelp | 25000 | 5000 | 5000 |
| | IMDB | 25000 | 5000 | 5000 |
| | AG | 120000 | 3000 | 3000 |
| | MR | 8595 | 1000 | 1000 |
| **Sentence Pair Classification** | QNLI | 100000 | 4743 | 5463 |
| | MRPC | 3668 | 408 | 1725 |

Table 1: Statistics of the datasets.

regularizer enforces the model $f$ to produce similar scores for the original data and augmented data, which lies in the semantic neighborhood of original embeddings. Furthermore, we instantiate the classification loss in Equation 1 as follows:

$$L_c(\theta) = \frac{1}{n} \sum_{i=1}^{n} \text{CE}\left( f(\mathbf{E}_i; \theta), y_i \right), \qquad (7)$$

where $CE(\cdot, \cdot)$ is the cross-entropy loss function, which can be changed according to specific tasks.

### 4.3 Overview and Discussion

In this part, we present the overview and discussions of our VDA approach.

**Overview** The overall framework of VDA consists of two important parts, namely embedding augmentation and regularized training. We present the overall training algorithm in Algorithm 1. For embedding augmentation, we utilize the output of a MLM as the multinomial mixtures to augment new embeddings for each token in the input sentence. It is called *virtual data augmentation*, since the augmented embeddings do not correspond to actual text or tokens, but a probabilistic mixture of all the token embeddings. Then, for regularized training, we leverage the original predictions to guide the learning of the augmented embeddings, which reduces the influence from noisy or incorrect perturbations in the augmentations.

**Discussion** In the background of machine learning (Schmidt et al., 2018; Yin et al., 2019), robustness corresponds to the ability to resist data drift, perturbation and attack. To improve the robustness, a key point is that the model is able to generalize

to the semantic neighborhood of training data instances (Schmidt et al., 2018). However, discrete augmentation methods (Wei and Zou, 2019; Wang and Bansal, 2018) (*e.g.,* insert, delete or replace tokens) do not have good generalization ability for model optimization. While, virtual adversarial training methods (Zhu et al., 2020; Jiang et al., 2020) cannot well constrain the augmentations in the original semantic space. As a comparison, our approach utilizes original token embeddings to augment new embeddings, so that the augmentations will stay close to the existing embeddings in the same semantic space. For relevance, we adopt a MLM to generate the multinomial mixture according to the likelihood of each candidate given the input. For diversity, we inject Gaussian noise to enhance the randomness. To further balance the two aspects, we design a regularized strategy to guide the augmentation learning towards the original predictions. By only revising the embeddings, our approach is model-agnostic and domain-agnostic, which is general to apply to various PLMs on different downstream tasks.

## 5 Experiment - Main Results

We demonstrate the effectiveness of VDA for fine-tuning PLMs in the text classification task.

### 5.1 Experimental Setup

#### 5.1.1 Dataset

We conduct experiments on the sentence classification task and the sentence-pair classification task. The dataset statistics are summarized in Table 1.

**Sentence Classification** We use four sentence classification datasets for evaluation.

    • **Yelp** (Zhang et al., 2015) [1]: a binary sentiment classification dataset based on restaurant reviews.

    • **IMDB** [2]: a binary document-level sentiment classification dataset on movie reviews.

---

[1] https://www.yelp.com/
[2] https://datasets.imdbws.com/

| Datasets | Yelp | | | | IMDB | | | |
|---|---|---|---|---|---|---|---|---|
| Metrics | Ori Acc | Att Acc | Q Num | Perturb | Ori Acc | Att Acc | Q Num | Perturb |
| BERT | 0.957 | 0.487 | 645.878 | 12.291 | **0.938** | **0.310** | 954.852 | 9.745 |
| BERT$_{VDA}$ | **0.959** | **0.533** | **687.681** | **17.720** | **0.938** | 0.307 | **1061.656** | **12.483** |
| FreeLB | 0.960 | 0.400 | 604.952 | 17.559 | 0.936 | 0.137 | 784.833 | 10.727 |
| FreeLB$_{VDA}$ | **0.962** | **0.507** | **694.910** | **18.869** | **0.939** | **0.250** | **1015.518** | **14.400** |
| SMART | 0.960 | 0.437 | 636.523 | 20.356 | **0.940** | 0.110 | **611.279** | 9.232 |
| SMART$_{VDA}$ | **0.962** | **0.527** | **691.941** | **20.372** | 0.938 | **0.143** | 824.192 | **12.785** |
| SMix | 0.957 | 0.557 | 705.042 | 15.305 | **0.935** | 0.210 | 833.642 | 9.816 |
| SMix$_{VDA}$ | **0.960** | **0.600** | **765.517** | **17.491** | 0.934 | **0.383** | **1105.819** | **12.733** |
| RoBERTa | **0.977** | 0.500 | 730.741 | **11.991** | 0.957 | **0.233** | 911.250 | 9.588 |
| RoBERTa$_{VDA}$ | 0.972 | **0.643** | **780.300** | 9.323 | **0.959** | 0.210 | **960.996** | **11.469** |
| Datasets | AG | | | | MR | | | |
| Metrics | Ori Acc | Att Acc | Q Num | Perturb | Ori Acc | Att Acc | Q Num | Perturb |
| BERT | 0.944 | 0.327 | 223.271 | **22.309** | 0.868 | 0.210 | 58.217 | 20.291 |
| BERT$_{VDA}$ | **0.946** | **0.450** | **268.981** | 21.757 | **0.878** | **0.339** | **70.519** | **22.775** |
| FreeLB | **0.945** | 0.301 | 215.178 | 21.175 | 0.879 | 0.240 | 60.498 | 20.266 |
| FreeLB$_{VDA}$ | **0.945** | **0.473** | **271.282** | **22.634** | **0.883** | **0.302** | **69.109** | **21.557** |
| SMART | 0.944 | 0.403 | 251.643 | 21.953 | 0.880 | 0.226 | 56.067 | 20.514 |
| SMART$_{VDA}$ | **0.945** | **0.484** | **273.724** | **23.572** | **0.885** | **0.298** | **63.965** | **23.386** |
| SMix | 0.944 | 0.425 | 269.616 | 21.981 | 0.880 | 0.251 | 61.289 | **21.842** |
| SMix$_{VDA}$ | **0.947** | **0.513** | **278.833** | **24.422** | **0.883** | **0.319** | **68.893** | 20.874 |
| RoBERTa | 0.951 | 0.464 | 301.749 | 19.716 | 0.919 | 0.344 | 81.432 | **27.128** |
| RoBERTa$_{VDA}$ | **0.952** | **0.497** | **303.964** | **23.127** | **0.925** | **0.439** | **89.427** | 24.310 |

Table 2: Main results on the sentence classification task. Ori Acc, Att acc, Q Num and Perturb denote the original accuracy, attack accuracy, query number and perturbed percentage per sample. "$_{VDA}$" denotes that the model is trained with our proposed VDA framework. The best results in each group are highlighted in bold.

• **AG's News** (Zhang et al., 2015): a news-type classification dataset, containing 4 types of news: World, Sports, Business, and Science.

• **MR** (Pang and Lee, 2005): a binary sentiment classification dataset based on movie reviews.

**Sentence-Pair Classification** We also use two sentence-pair classification datasets for evaluation.

• **QNLI** (Demszky et al., 2018): a question-answering dataset consisting of question-paragraph pairs. The task is to determine whether the context sentence contains the answer to the question.

• **MRPC** (Dolan and Brockett, 2005): a corpus of sentence pairs with human annotations about the semantic equivalence.

### 5.1.2 Baselines

To evaluate the generalization of our framework, we implement VDA on the following models.

• **BERT-Base** (Devlin et al., 2019) is the 12-layer BERT model with 768 hidden units and 12 heads, totally 110M parameters.

• **FreeLB** (Zhu et al., 2020) is an adversarial

training approach for fine-tuning PLMs, which adds gradient-based perturbations to token embeddings. We implement it on BERT-Base.

• **SMART** (Jiang et al., 2020) is a robust and efficient computation framework for fine-tuning PLMs. Limited by the GPU resource, we can only implement the smooth-inducing adversarial regularization on BERT-Base but remove the Bregman Proximal Point Optimization.

• **SMix** (Si et al., 2020) uses mixup on [CLS] tokens of the PLM to cover larger attack space. We implement it on BERT-Base. For a fair comparison, we remove the adversarial data augmentation strategy here, and leave it on Section 6.2.

• **RoBERTa-Large** (Liu et al., 2019) is a robustly optimized BERT model with more training data and time. It owns 24 layers, 1024 hidden units and 16 heads, totally 355M parameters.

### 5.1.3 Evaluation Metrics

We set up various metrics for measuring accuracy and robustness. Original accuracy, is the accuracy

| Datasets | MRPC | | | | QNLI | | | |
|---|---|---|---|---|---|---|---|---|
| Metrics | Ori Acc | Att Acc | Q Num | Perturb | Ori Acc | Att Acc | Q Num | Perturb |
| BERT | 0.826 | 0.163 | 77.276 | 9.601 | 0.909 | 0.342 | 93.515 | 13.451 |
| BERT$_{VDA}$ | **0.831** | **0.206** | **90.686** | **10.617** | **0.913** | **0.410** | **112.088** | **14.816** |
| FreeLB | 0.827 | 0.154 | 82.372 | 10.193 | 0.910 | 0.363 | 98.703 | 14.283 |
| FreeLB$_{VDA}$ | **0.838** | **0.205** | **87.379** | **10.566** | **0.915** | **0.428** | **111.812** | **15.953** |
| SMART | 0.831 | 0.139 | 81.114 | 10.270 | 0.909 | 0.309 | 91.435 | **14.654** |
| SMART$_{VDA}$ | **0.832** | **0.179** | **85.628** | **11.546** | **0.911** | **0.388** | **105.918** | 14.611 |
| SMix | 0.824 | 0.249 | **116.660** | 11.224 | 0.886 | 0.171 | 71.849 | 10.171 |
| SMix$_{VDA}$ | **0.833** | **0.258** | 97.380 | **11.448** | **0.915** | **0.389** | **109.562** | **14.537** |
| RoBERTa | 0.850 | 0.179 | 78.905 | 9.357 | 0.934 | 0.408 | 101.977 | **12.112** |
| RoBERTa$_{VDA}$ | **0.859** | **0.255** | **97.485** | **11.623** | **0.941** | **0.411** | **103.405** | 11.901 |

Table 3: Main results on the sentence-pair classification task. "$_{VDA}$" denotes that the model is trained with our proposed VDA framework. The best results in each group are highlighted in bold.

of models on the original test set. While attack accuracy is the counter-part of after attack accuracy, which is the core metric measuring the robustness. Larger attack accuracy reflects better robustness. In this paper, we adopt BERT-Attack (Li et al., 2020b) as the attack method, since it can generate fluent and semantically preserved samples. For AG, MR, QNLI and MRPC datasets, we follow previous works (Jin et al., 2020; Li et al., 2020b) to randomly sample 1000 instances for robustness evaluation. For Yelp and IMDB, we randomly sample 300 instances since the long sentences in the two datasets are more time-consuming. Note that for sentence-pair classification datasets (*i.e.,* QNLI and MRPC), we attack the second sentence in evaluation. Besides, we also apply the query number and perturbed percentage per sample for evaluation. Under the black-box setting, queries of the target model are the only way of attack methods to access information. The larger query number indicates that the vulnerability of the target model is harder to be detected, which reflects better robustness. The perturbed percentage is the ratio of perturbed words number to the text length, a larger percentage also reveals more difficulty to successfully attack the model.

### 5.1.4 Implementation Details

We implement all baseline models based on HuggingFace Transformers [3], and their hyper-parameters are set following the suggestions from the original papers. For our proposed VDA, we reuse the same hyper-parameter setting as the original baseline model. All models are trained on a

[3]https://huggingface.co/transformers/

GeForce RTX 3090.

For hyper-parameters in VDA, the sampling number $m$ is set as 1, the learning rate is $1e^{-5}$. We use 5% steps to warm up PLMs during training. The variance of Gaussian noise is mostly set as $1e^{-2}$ and tuned in $\{1e^{-3}, 4e^{-3}, 1e^{-2}, 4e^{-2}\}$, the weight $\lambda$ is mostly set as 1.0 and tuned in $\{0.04, 0.1, 0.4, 1.0, 4.0\}$.

### 5.2 Main Results

Table 2 reports the evaluation results of our proposed VDA framework and the baseline models on sentence classification datasets. And the results on sentence-pair classification datasets are shown in Table 3. Based on these results, we can find:

First, FreeLB and SMART mostly outperform BERT-base model on the original accuracy metric, but perform not well on robustness-related metrics, especially on Yelp and IMDB datasets. These methods adopt gradient-based perturbations and smoothness-inducing regularization, respectively, which are able to improve the classification accuracy but may be not effective in defending against adversarial attacks. A potential reason may be that textual adversarial attacks are discrete, which can not be captured by virtual adversarial training.

Second, SMix improves the robustness of BERT-base in all datasets, but performs not well in original accuracy. It mixes hidden representations of the BERT-base model, which increases the coverage of the attack space for PLMs but may augment noised examples into training data. Besides, RoBERTa-large outperforms all other baselines in performance and robustness metrics. The reason is that RoBERTa-large is pre-trained on more training

| AG | | | | |
|---|---|---|---|---|
| Method | Ori Acc | Att Acc | Q Num | Perturb |
| BERT | 0.944 | 0.360 | 241.758 | 22.416 |
| +$VDA$ | **0.949** | 0.468 | 284.946 | 22.470 |
| +$VDA-\epsilon$ | 0.945 | 0.445 | 280.672 | 21.642 |
| +$CEVDA$ | 0.945 | 0.451 | 275.034 | 20.621 |
| +$Argmax$ | 0.943 | **0.478** | **298.186** | 19.515 |
| +$Sample$ | 0.946 | 0.459 | 274.553 | **22.479** |

| QNLI | | | | |
|---|---|---|---|---|
| Method | Ori Acc | Att Acc | Q Num | Perturb |
| BERT | 0.834 | 0.163 | 77.276 | 9.601 |
| +$VDA$ | 0.838 | **0.206** | **90.686** | 10.617 |
| +$VDA-\epsilon$ | 0.833 | 0.184 | 86.914 | **11.075** |
| +$CEVDA$ | 0.832 | 0.173 | 80.389 | 10.492 |
| +$Argmax$ | 0.834 | 0.160 | 71.074 | 9.224 |
| +$Sample$ | **0.840** | 0.149 | 74.197 | 9.263 |

Table 4: Ablation and variation study of our approach on the developed set of AG and QNLI datasets. BERT indicates the BERT-base model.

| MR | | | | |
|---|---|---|---|---|
| Metrics | Ori Acc | Att Acc | Q Num | Perturb |
| BERT-base | 0.866 | 0.215 | 61.886 | 21.243 |
| +$VDA$ | **0.874** | 0.326 | 70.681 | 21.577 |
| +$ADA$ | 0.862 | 0.287 | 72.635 | **22.627** |
| +$VDA+ADA$ | 0.869 | **0.386** | **77.586** | 20.781 |

| MRPC | | | | |
|---|---|---|---|---|
| Metrics | Ori Acc | Att Acc | Q Num | Perturb |
| BERT-base | 0.834 | 0.163 | 77.276 | 9.601 |
| +$VDA$ | **0.838** | 0.206 | 90.686 | 10.617 |
| +$ADA$ | 0.828 | 0.214 | 89.721 | 10.317 |
| +$VDA+ADA$ | 0.837 | **0.215** | **95.786** | **10.623** |

Table 5: The study of combining VDA and ADA on the developed set of MR and MRPC datasets, BERT indicates the BERT-base model.

data with more training time, which can directly improve the generalization and robustness to adversarial attack samples.

Finally, we compare our proposed framework with these baseline models. After being combined with VDA, it is clear to see a significant improvement in robustness metrics on most of datasets. Our VDA utilizes a masked language model to generate substitution probabilities, and then add a Gaussian noise. In this way, we can augment diverse and semantic-consistent examples, which are able to improve the robustness of PLMs. Furthermore, we can also see that the most of baseline models combined with VDA achieve a marginal improvement in original accuracy. It indicates that our approach can better balance the performance and robustness of PLMs. Among them, we can see that our VDA can bring more improvement in MRPC and QNLI. The reason may be that the two tasks are more difficult and require more data for training. The virtual augmented data via our approach is semantic-consistent and diverse, hence it can be more helpful for these tasks.

## 6 Experiment - Analysis and Extension

In this section, we continue to study and analyze the effectiveness of our proposed VDA.

### 6.1 Ablation and Variation Study

We devise four variations for exploring the effectiveness of key components in our proposed VDA. BERT+$VDA-\epsilon$ is the variation by removing the Gaussian noise $\epsilon$ in Eq. 4. BERT+$CEVDA$ replaces the symmetric KL-divergence by cross-entropy loss. BERT+$Argmax$ and BERT+$Sample$ adopt argmax and sample operators to select the substituted token according to the substitution probability, respectively. We conduct the experiments on AG and QNLI datasets.

As shown in Table 4, most of the variations perform better than BERT in robustness metrics, since they all augment virtual data for improving the robustness. Among them, BERT+$VDA$ outperforms most of the variations in both accuracy and robustness metrics. It indicates that the Gaussian noise, symmetric KL-divergence loss and weighted aggregated embeddings are all useful to improve the robustness and stabilize the accuracy. However, we can see BERT+$Argmax$ and BERT+$Sample$ achieve better results than BERT+$VDA$ in part of metrics, but cause a dramatic drop in other metrics. It indicates that the two variations can not balance the trade-off between accuracy and robustness well.

### 6.2 Virtual Data Augmentation with Adversarial Data Augmentation

Our proposed VDA is general to various methods, including conventional adversarial data augmentation (ADA). In this part, we collect the adversarial examples curated from the MR and MRPC training sets, and add them to the original training set, respectively. Then we test the accuracy and the robustness of BERT-base model and our VDA after training with the adversarial data. As seen in Table 5, although augmented adversarial data improves the robustness of BERT, the performance on original accuracy also drops. The reason may be that there are noised instances in the adversarial data. As a comparison, our proposed VDA can augment diverse and semantic-consistent virtual data, which better balances accuracy and robustness. Be-
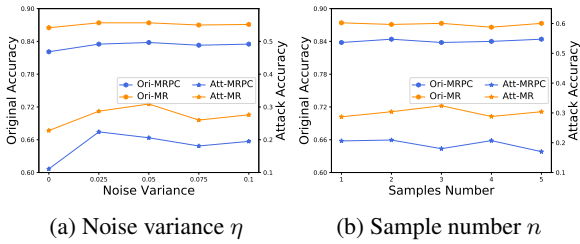
(a) Noise variance $\eta$     (b) Sample number $n$

Figure 3: Performance comparison w.r.t. noise variance and sample number on the developed set of MR and MRPC datasets.



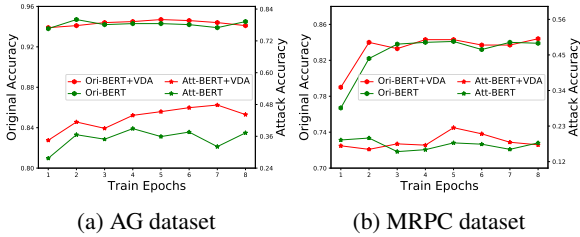(a) AG dataset     (b) MRPC dataset

Figure 4: Performance comparison of BERT and BERT$_{VDA}$ w.r.t. training epochs on the developed set of AG and MRPC datasets.

sides, after combining with ADA, our VDA can be further improved on accuracy and robustness metrics. It indicates that our approach is also general to ADA methods.

### 6.3 Hyper-parameter Analysis

Our framework includes a few parameters to tune. Here, we report the tuning results of two parameters on MR and MRPC datasets, *i.e.,* the variance of the Gaussian noise $\eta$ and the number of argumented virtual data. We show the change curves of original accuracy and attack accuracy in Figure 3. We can see that our model achieves the best performance when the variance is nearby 0.05. It indicates that too small or too large noise may influence the quality of the augmented virtual data. Besides, our model also achieves the best performance when the sampling number is nearby 3. It shows that augmenting 3 examples per sample is enough to improve the robustness.

### 6.4 Performance Change during Regularizing Fine-tuning

In this part, we investigate how the accuracy and robustness change during regularizing fine-tuning with our VDA. We conduct experiments on AG and MRPC datasets and report the original accuracy and attack accuracy metrics. As shown in Figure 4, the original and attack accuracy of the model can be improved with the increasing of train-

ing epochs. When reaching the optimal point, the accuracy and robustness start to shock, and even decrease to some extent. The reason may be that the model has overfitted. An interesting finding is that the optimal points of the original accuracy and attack accuracy are usually not the same one. A possible reason is that accuracy and robustness are not always consistent objectives for deep models. Besides, we can see that after combined with our VDA, BERT is able to achieve a better optimal point with higher original and attack accuracy. It indicates that VDA is an effective regularization approach for BERT.

## 7 Conclusion

In this work, we proposed the framework *virtual data augmentation (VDA)*, for robustly fine-tuning pre-trained language models. It is a general framework agnostic to downstream tasks, model architectures and learning strategies. In VDA, we augmented new embeddings by making weighted aggregation on token embedding matrix according to a multinomial mixture distribution. To construct the mixture distribution, we utilized a masked language model to generate the substitution probability for guaranteeing semantic consistency, and a Gaussian noise to provide diversity. And we also adopted a regularized training strategy to further enhance the robustness. Extensive experiments on six datasets have demonstrated that the proposed approach can effectively improve the robustness of various PLMs.

# References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2890–2896.

Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu. 2020. Seqvat: Virtual adversarial training for semi-supervised sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8801–8811. Association for Computational Linguistics.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *CoRR*, abs/1809.02922.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing, IWP@IJCNLP 2005, Jeju Island, Korea, October 2005, 2005*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 31–36.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 567–573.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 50–56.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tao Gui, Xiao Wang, Qi Zhang, Qin Liu, Yicheng Zou, Xin Zhou, Rui Zheng, Chong Zhang, Qinzhuo Wu, Jiacheng Ye, Zexiong Pang, Yongxin Zhang, Zhengyan Li, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xinwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Bolin Zhu, Shan Qin, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. *CoRR*, abs/2103.11441.

Xuanli He, Lingjuan Lyu, Qiongkai Xu, and Lichao Sun. 2021. Model extraction and adversarial transferability, your BERT is vulnerable! *CoRR*, abs/2103.10013.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1234–1245.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2021–2031.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: robust and efficient fine-tuning for pretrained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2177–2190.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2752–2765.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the*

*North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 452–457.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2020a. Contextualized perturbation for textual adversarial attack. *CoRR*, abs/2009.07502.

Guanlin Li, Lemao Liu, Guoping Huang, Conghui Zhu, and Tiejun Zhao. 2019a. Understanding data augmentation in neural machine translation: Two perspectives towards generalization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5688–5694.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019b. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*.

Juntao Li, Lisong Qiu, Bo Tang, Dongmin Chen, Dongyan Zhao, and Rui Yan. 2019c. Insufficient data can also rock! learning to converse using smaller data with augmentation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6698–6705.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020b. BERT-ATTACK: adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6193–6202.

Linyang Li and Xipeng Qiu. 2021. Tavat: Token-aware virtual adversarial training for language understanding. In *AAAI 2021*.

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. *CoRR*, abs/1907.11692.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018.

Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3103–3114.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2339–2352.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1979–1993.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. 2015. Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 115–124.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. 2019. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13824–13833.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin,*

*Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1085–1097.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 856–865.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. 2018. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5019–5031.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Better robustness by more coverage: Adversarial training with mixup augmentation for robust fine-tuning. *CoRR*, abs/2012.15699.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Boxin Wang, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang, and Bo Li. 2020. T3: treeautoencoder constrained adversarial text generation for targeted attack. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6134–6150.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2557–2563.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 575–581.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I. Jordan. 2020. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. *J. Mach. Learn. Res.*, 21:43:1–43:36.

Dong Yin, Kannan Ramchandran, and Peter L. Bartlett. 2019. Rademacher complexity for adversarially robust generalization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 7085–7094.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019. Generating fluent adversarial examples for natural languages. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5564–5569.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Stephan Zheng, Yang Song, Thomas Leung, and Ian J. Goodfellow. 2016. Improving the robustness of deep neural networks via stability training. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4480–4488.

Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020a. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1893–1902. ACM.

Kun Zhou, Kai Zhang, Yu Wu, Shujie Liu, and Jingsong Yu. 2019. Unsupervised context rewriting for open domain conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

*Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1834–1844. Association for Computational Linguistics.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020b. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020,*

*Barcelona, Spain (Online), December 8-13, 2020*, pages 4128–4139. International Committee on Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.