Textual Manifold-based Defense Against Natural Language Adversarial Examples

Dang Minh Nguyen
VinAI Research, Vietnam
v.dangnm12@vinai.io

Luu Anh Tuan*
Nanyang Technological University, Singapore
anhtuan.luu@ntu.edu.sg

Abstract

Recent studies on adversarial images have shown that they tend to leave the underlying low-dimensional data manifold, making them significantly more challenging for current models to make correct predictions. This so-called off-manifold conjecture has inspired a novel line of defenses against adversarial attacks on images. In this study, we find a similar phenomenon occurs in the contextualized embedding space induced by pretrained language models, in which adversarial texts tend to have their embeddings diverge from the manifold of natural ones. Based on this finding, we propose Textual Manifold-based Defense (TMD), a defense mechanism that projects text embeddings onto an approximated embedding manifold before classification. It reduces the complexity of potential adversarial examples, which ultimately enhances the robustness of the protected model. Through extensive experiments, our method consistently and significantly outperforms previous defenses under various attack settings without trading off clean accuracy. To the best of our knowledge, this is the first NLP defense that leverages the manifold structure against adversarial attacks. Our code is available at https://github.com/dangne/tmd.

1 Introduction

The field of NLP has achieved remarkable success in recent years, thanks to the development of large pretrained language models (PLMs). However, multiple studies have shown that these models are vulnerable to adversarial examples - carefully optimized inputs that cause erroneous predictions while remaining imperceptible to humans (Jin et al., 2020; Li et al., 2020; Garg and Ramakrishnan, 2020). This problem raises serious security concerns as PLMs are widely deployed in many modern NLP applications.

Various defenses have been proposed to counter adversarial attacks, which can be broadly categorized into empirical (Si et al., 2021; Dong et al., 2021b; Miyato et al., 2017) and certified (Jia et al., 2019; Ye et al., 2020; Zeng et al., 2021; Zhao et al., 2022) defenses. Adversarial training is currently the most successful empirical method (Athalye et al., 2018; Dong et al., 2021a; Zhou et al., 2021). It operates by jointly training the victim model on clean and adversarial examples to improve the robustness. However, one major drawback of this approach is the prohibitively expensive computational cost. Schmidt et al. (2018) has theoretically shown that with just simple models, the sample complexity of adversarial training already grows substantially compared to standard training. Alternatively, certified defenses aim to achieve a theoretical robustness guarantee for victim models under specific adversarial settings. However, most certified defenses for NLP are based on strong assumptions on the network architecture (Jia et al., 2019; Shi et al., 2020), and the synonym set used by attackers is often assumed to be accessible to the defenders (Ye et al., 2020). Li et al. (2021) has shown that when using different synonym set during the attack, the effectiveness of these methods can drop significantly.

Concurrent with the streams of attack and defense research, numerous efforts have been made to understand the characteristics of adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015; Tanay and Griffin, 2016; Gilmer et al., 2018; Ilyas et al., 2019; Tsipras et al., 2019). One rising hypothesis is the off-manifold conjecture, which states that adversarial examples leave the underlying low-dimensional manifold of natural data (Tanay and Griffin, 2016; Gilmer et al., 2018; Stutz et al., 2019; Shamir et al., 2021). This observation has inspired a new line of defenses that leverage the data manifold to defend against adversarial examples, namely manifold-based defenses (Saman-

*Corresponding author

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6612 - 6625
December 7-11, 2022

gouei et al., 2018; Meng and Chen, 2017; Song et al., 2017). Despite the early signs of success, such methods have only focused on images. It remains unclear if the off-manifold conjecture also generalizes to other data domains such as texts and how one can utilize this property to improve models' robustness.

In this study, we empirically show that the off-manifold conjecture indeed holds in the contextualized embedding space of textual data. Based on this finding, we develop Textual Manifold-based Defense (TMD), a novel method that leverages the manifold of text embeddings to improve NLP robustness. Our approach consists of two key steps: (1) approximating the contextualized embedding manifold by training a generative model on the continuous representations of natural texts, and (2) given an unseen input at inference, we first extract its embedding, then use a sampling-based reconstruction method to project the embedding onto the learned manifold before performing standard classification. TMD has several benefits compared to previous defenses: our method improves robustness without heavily compromising the clean accuracy, and our method is structure-free, i.e., it can be easily adapted to different model architectures. The results of extensive experiments under diverse adversarial settings show that our method consistently outperforms previous defenses by a large margin.

In summary, the key contributions in this paper are as follows:

- We show empirical evidence that the off-manifold conjecture holds in the contextualized embedding space induced by PLMs.
- We propose TMD, a novel manifold-based defense that utilizes the off-manifold conjecture against textual adversarial examples.
- We perform extensive experiments under various settings, and the results show that our method consistently outperforms previous defenses.

2 Related Work

2.1 Adversarial Attacks and Defenses

Textual adversarial examples can be generated at different granularity levels. One can add, delete, replace characters (Gao et al., 2018; Ebrahimi et al., 2018a; Li et al., 2019) or words (Ren et al., 2019; Jin et al., 2020; Li et al., 2020; Garg and Ramakrishnan, 2020; Alzantot et al., 2018), or manipulate the entire sentence (Iyyer et al., 2018; Ribeiro et al., 2018; Zhao et al., 2018) to maximize the prediction error without changing the original semantics. Among the different attack strategies above, word substitution-based attacks are the most popular and well-studied methods in the literature (Ebrahimi et al., 2018b; Alzantot et al., 2018; Ren et al., 2019; Jin et al., 2020; Li et al., 2020). Ebrahimi et al. (2018b) is the first work to propose a white box gradient-based attack on textual data. Follow-up works (Alzantot et al., 2018; Ren et al., 2019; Jin et al., 2020; Li et al., 2020) introduce additional constraints to the perturbation space such as using synonyms for substitution, part-of-speech or semantic similarity checking to ensure the generated samples are semantically closed to the original ones.

Regarding the defenses against NLP attacks, adversarial training is one of the most successful defenses. The first adversarial training method is introduced in Goodfellow et al. (2015) for image data. The authors show that training on both clean and adversarial examples can improve the model's robustness. Miyato et al. (2017) develops a similar approach for the textual domain with L_2-bounded perturbation in the embedding space. Jia et al. (2019) and Huang et al. (2019) propose using axis-aligned bounds to restrict adversarial perturbation. However, Dong et al. (2021a) later argues that these bounds are not sufficiently inclusive or exclusive. Therefore, the authors propose to instead model the perturbation space as the convex hull of word synonyms. They also propose an entropy-based regularization to encourage perturbations to point to actual valid words. Zhu et al. (2020) proposes FreeLB, a novel adversarial training approach that addresses the computational inefficiency of previous methods. However, their original work focuses on improving generalization rather than robustness.

2.2 Off-manifold Conjecture and Manifold-based Defenses

The off-manifold conjecture was first proposed in Tanay and Griffin (2016) as an alternative explanation for the existence of adversarial examples to previous ones (Goodfellow et al., 2015; Szegedy et al., 2014). Although it has been shown in previous works that on-manifold adversarial examples do exist (Gilmer et al., 2018), Stutz et al. (2019)

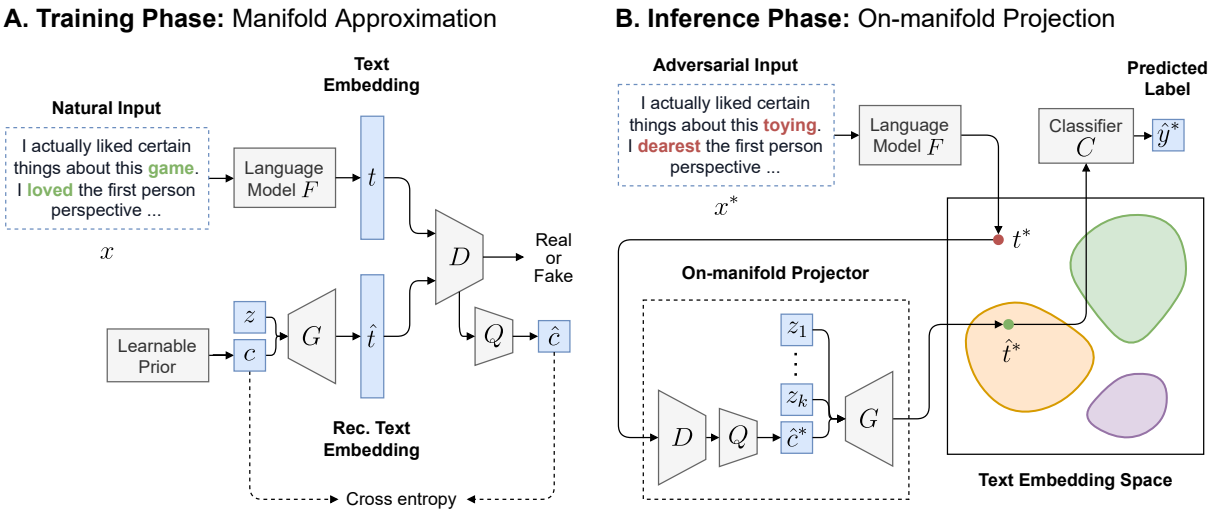


Figure 1: An overview of the Textual Manifold-based Defense. (A) Training Phase: All textual inputs x are transformed into continuous representations $t = F(x)$. An InfoGAN (Chen et al., 2016) with learnable prior is trained to distinguish between real embeddings t versus fake embeddings $\hat{t}$ to implicitly learn the natural text embedding manifold. (B) On-manifold projection: Once the generative model is trained, novel input embedding $t^* = F(x^*)$ is projected onto the approximated manifold using a sampling-based reconstruction strategy. The reconstructed embedding $\hat{t}^*$ is fed to the classifier $C(.)$ to produce the final predicted label $\hat{y}^*$. The colored blobs represent the approximated disjoint submanifolds in the contextualized embedding space.

later shows that on-manifold robustness is essentially generalization, i.e., they are simply generalization errors. Recent work from Shamir et al. (2021) independently finds similar observations to Tanay and Griffin (2016) and Stutz et al. (2019). They propose a conceptual framework called Dimpled Manifold Model and use it to explain various unanswered phenomena of adversarial examples (Ilyas et al., 2019; Tsipras et al., 2019).

Based on this property, many defenses have been developed. Samangouei et al. (2018) proposes Defense-GAN, which uses a Wasserstein Generative Adversarial Network (WGAN) (Arjovsky et al., 2017) to project adversarial examples onto the approximated manifold. A similar idea is used in Schott et al. (2019), in which the authors propose to use Variational Auto-Encoder (VAE) (Kingma and Welling, 2014) to learn the data manifold.

3 Textual Manifold-based Defense

In this section, we introduce our method in greater detail. We first introduce how we approximate the contextualized embedding manifold using deep generative models. Then, we describe how to perform on-manifold projection with the trained model in the inference phase. The general overview of our method is visualized in Figure 1.

3.1 Textual Manifold Approximation

Approximating the manifold of textual data is not straightforward due to its discrete nature. Instead of modeling at the sentence level, we relax the problem by mapping texts into their continuous representations.

More formally, let $F : \mathcal{X} \rightarrow \mathbb{R}^n$ be a language model that maps an input text $x \in \mathcal{X}$ to its corresponding n-dimensional embedding t, where $\mathcal{X}$ denotes the set of all texts, we first compute the continuous representations t for all x in the dataset. We assume that all t lie along a low-dimensional manifold $\mathcal{T} \subset \mathbb{R}^n$. Several studies have shown that the contextualized embedding space consists of disjoint clusters (Mamou et al., 2020; Cai et al., 2021). Based on these findings, we assume that $\mathcal{T}$ is a union of disconnected submanifolds, i.e., $\mathcal{T} = \bigcup \mathcal{T}_i,\ \forall i \neq j : \mathcal{T}_i \cap \mathcal{T}_j = \emptyset$. The problem is now simplified to approximating the contextualized embedding manifold $\mathcal{T}$. However, choosing the appropriate generative model is crucial as learning complex manifolds in high dimensional spaces heavily depends on the underlying geometry (Fefferman et al., 2016). While previous manifold-based defenses for images are able to achieve impressive results with simple models such as WGAN (Samangouei et al., 2018) or VAE (Schott et al.,

2019), we will show in Section 4.4 that this is not the case for contextualized embeddings.

Proposition 3.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be topological spaces and let $f : \mathcal{X} \to \mathcal{Y}$ be a continuous function. If $\mathcal{X}$ is connected then the image $f(\mathcal{X})$ is connected.*

The original GAN proposed by Goodfellow et al. (2014) trains a continuous generator $G : \mathcal{Z} \to \mathcal{X}'$ that maps a latent variable $z \in \mathcal{Z}$ sampled from some prior $P(z)$ to the target space $\mathcal{X}'$. The standard multivariate normal $\mathcal{N}(0, I)$ is often chosen as the distribution for $P(z)$. This implies that $P(z)$ is supported on the connected space $\mathbb{R}^d$. Therefore, the target space $\mathcal{X}'$ is also connected according to Proposition 3.1. This explains why such simple model fails to learn disconnected manifolds (Gurumurthy et al., 2017; Khayatkhoei et al., 2018; Tanielian et al., 2020).

To approximate $\mathcal{T}$, we need to introduce disconnectedness in the latent space $\mathcal{Z}$. We follow the InfoGAN (Chen et al., 2016) approach by adding an additional discrete latent code c to $c \sim Cat(K, r)$, where c is a K-dimensional one-hot vector, K is a hyperparameter for the number of submanifolds in $\mathcal{T}$ and $r \in \mathbb{R}^K$ denotes the probabilities for each value of c ($r_i \geq 0, \sum r_i = 1$). The generator now becomes $G : \mathcal{Z} \times \mathcal{C} \to \mathcal{T}$.

Naturally, we expect to have different values of c targeting different submanifold $\mathcal{T}_i$. This can be achieved by maximizing the mutual information $I(c; G(z, c))$ between the latent code c and the generated embeddings $G(z, c)$. However, directly optimize $I(c, G(z, c))$ can be difficult, so we instead maximize the lower bound of $I(c, G(z, c))$:

$$\begin{aligned}
&I(c; G(z, c)) \\
&= H(c) - H(c \mid G(z, c)) \\
&= \mathbb{E}_{t\sim P_g(t)} \left[\mathbb{E}_{c'\sim P(c|t)} \left[\log P(c' \mid t)\right]\right] + H(c) \\
&= \mathbb{E}_{t\sim P_g(t)} \left[D_{\mathrm{KL}} \left(P(\cdot \mid t) \parallel Q(\cdot \mid t)\right)\right. \\
&\quad \left. + \mathbb{E}_{c'\sim P(c|t)} \left[\log Q(c' \mid t)\right]\right] + H(c) \\
&\geq \mathbb{E}_{t\sim P_g(t)} \left[\mathbb{E}_{c'\sim P(c|t)} \left[\log Q(c' \mid t)\right]\right] + H(c) \\
&= \mathbb{E}_{c\sim P(c), t\sim P_g(t)} \left[\log(Q(c \mid t))\right] + H(c)
\end{aligned} \tag{1}$$

where $Q(c \mid t)$ is an auxiliary distribution parameterized by a neural network to approximate $P(c \mid t)$, $P_g(t)$ denotes the distribution of generated embeddings from $G(z, c)$. The problem of maximizing $I(c, G(z, c))$ becomes maximizing the following information-theoretic regularization:

$$L_I(G, Q) = \mathbb{E}_{c\sim P(c), t\sim P_g(t)} \left[\log(Q(c \mid t))\right] \tag{2}$$

Combine with the objective function of GAN,

$$\begin{aligned}
V(D, G) = &\mathbb{E}_{t\sim P_r(t)} \left[\log \left(D \left(t\right)\right)\right] \\
&+ \mathbb{E}_{\hat{t}\sim P_g(t)} \left[\log \left(1 - D \left(\hat{t}\right)\right)\right]
\end{aligned} \tag{3}$$

where $P_r(t)$ denotes the distribution of natural text embeddings, we obtain the following initial objective function to implicitly learn the manifold $\mathcal{T}$:

$$\min_{G,Q} \max_{D} V(D, G) - \lambda L_I(G, Q) \tag{4}$$

3.2 Learnable Prior

One limitation with the original InfoGAN design is the fixed uniform latent code distribution. Firstly, the true number of submanifolds in the target space is generally unknown. Secondly, data samples are not likely to be uniformly distributed over all submanifolds. To address these issues, we use expectation maximization to learn the optimal prior. Since $Q(c \mid t)$ approximates $P(c \mid t)$, $\mathbb{E}_{t\sim P_r} Q(c \mid t)$ gives us an approximation for the true prior distribution $P(c)$ which will then be used to train the InfoGAN in the M step. Doing this, the probabilities r_i are adaptively readjusted and redundant latent code values will have their weights zeroed out. Instead of optimizing r directly, we model it as $r = \mathrm{softmax}(\hat{r})$, $\hat{r} \in \mathbb{R}^K$ then optimize on the unconstrained $\hat{r}$. We train $\hat{r}$ by minimizing the following cross entropy:

$$\begin{aligned}
H(P(c), r) &= -\mathbb{E}_{c\sim P(c)}[\log r] \\
&= -\mathbb{E}_{t\sim P_r(t), c\sim P(c|t)}[\log r] \\
&= \mathbb{E}_{t\sim P_r(t)}[H(P(c \mid t), r)] \\
&\approx \mathbb{E}_{t\sim P_r(t)}[H(Q(c \mid t), r)]
\end{aligned} \tag{5}$$

We obtain the objective function for prior learning as follow:

$$L_P(\hat{r}) = \mathbb{E}_{t\sim P_r(t)} \left[H(Q(c \mid t), r)\right] \tag{6}$$

Combine Equation 6 with Equation 4, we arrive at the final objective function for manifold approximation:

$$\min_{G,Q,\hat{r}} \max_{D} V(D, G) - \lambda L_I(G, Q) + L_P(\hat{r}) \tag{7}$$

3.3 On-manifold Projection

Once G is trained, the next step is to develop an on-manifold projection method. Given an input embedding t, projecting it onto the approximated manifold is essentially finding an embedding $\hat{t}$ on $G(z, c)$ that is the closest to t, i.e., solving $\min_{z,c} \|G(z, c) - t\|_2$. Conveniently, we can utilize the auxiliary network $Q(c \mid t)$ to determine the optimal value for c, which is the submanifold id that t belongs to. The problem simplified to $\min_z \|G(z, c_t) - t\|_2$, where $c_t = Q(c \mid t)$. For the latent variable z, previous works in the GAN inversion literature often use optimization-based approaches to find the optimal z (Samangouei et al., 2018; Creswell and Bharath, 2019; Abdal et al., 2019). However, we will show in Section 4.6 that simply sampling k candidate $z_i \sim P(z)$ and selecting the one that produces minimal reconstruction loss results in better robustness and faster inference speed. In summary, given an embedding t, we compute the reconstructed on-manifold embedding $\hat{t}$ as follow:

$$\begin{aligned} & \hat{t} = G(z^*, c_t) \\ \text{where} \quad & c_t = Q(c \mid t) \\ & z^* = \arg\min_{z \sim P(z)} \|G(z, c_t) - t\|_2 \end{aligned} \tag{8}$$

The reconstructed embedding $\hat{t}$ is then fed into a classifier $C : \mathcal{T} \to \mathcal{Y}$ to produce the final predicted label $\hat{y} \in \mathcal{Y}$.

4 Experiments

4.1 Experimental Setting

Datasets We evaluate our method on three datasets: AG-News Corpus (AGNEWS) (Zhang et al., 2015a), Internet Movie Database (IMDB) (Maas et al., 2011), and Yelp Review Polarity (YELP) (Zhang et al., 2015b). The AGNEWS dataset contains over 120000 samples, each belonging to one of the four labels: World, Sports, Business, Sci/Tech. The IMDB dataset contains 50000 data samples of movie reviews with binary labels for negative and positive sentiments. The YELP dataset contains nearly 600000 samples of highly polar Yelp reviews with binary labels. However, due to limitations in computing resources, we only use a subset of 63000 samples of the YELP dataset. In addition, we randomly sample 10% of the training set for validation in all datasets.

Model Architectures To test if our method is able to generalize to different architectures, we apply TMD on three state-of-the-art pretrained language models: $\text{BERT}_{\text{base}}$ (Devlin et al., 2019), $\text{RoBERTa}_{\text{base}}$ (Liu et al., 2019), and $\text{XLNet}_{\text{base}}$ (Yang et al., 2019). BERT is a Transformer-based language model that has brought the NLP research by storm by breaking records on multiple benchmarks. Countless variants of BERT have been proposed (Xia et al., 2020) in which RoBERTa and XLNet are two of the most well-known. In addition to the above models, we also experiment with their larger versions in Appendix B.

Adversarial Attacks We choose the following state-of-the-art attacks to measure the robustness of our method: (1) PWWS (Ren et al., 2019) is a word synonym substitution attack where words in a sentence are greedily replaced based on their saliency and maximum word-swap effectiveness. (2) TextFooler (Jin et al., 2020) utilize nearest neighbor search in the counter-fitting embeddings (Mrkšić et al., 2016) to construct the dictionaries. Subsequently, words are swapped greedily based on their importance scores. They also introduce constraints such as part-of-speech and semantic similarity checking to ensure the generated adversarial example looks natural and does not alter its original label. (3) BERT-Attack (Li et al., 2020) is similar to TextFooler, but instead of using synonyms for substitution, they use BERT masked language model to produce candidate words that fit a given context. They also introduce subtle modifications to the word importance scoring function. All attacks above are implemented using the TextAttack framework (Morris et al., 2020).

For a fair comparison, we follow the same settings as Li et al. (2021), in which all attacks must follow the following constraints: (1) the maximum percentage of modified words ρ_{max} for AGNEWS, IMDB, and YELP must be 0.3, 0.1, and 0.1 respectively, (2) the maximum number of candidate replacement words K_{max} is set to 50, (3) the minimum semantic similarity[1] ϵ_{min} between original input x and adversarial example x' must be 0.84, and (4) the maximum number of queries to the victim model is $Q_{max} = K_{max} \times L$, where L is the length of the original input x.

[1]The semantic similarity between x and x' is approximated by measuring the cosine similarity between their text embeddings produced from Universal Sentence Encoder (Cer et al., 2018)

Model	Defense	AGNEWS				IMDB				YELP			
		CA	PW	TF	BA	CA	PW	TF	BA	CA	PW	TF	BA
BERT	Vanilla	94.39	39.2	27.9	39.0	92.15	6.5	2.3	1.0	95.28	13.7	10.5	2.8
	ASCC	91.57	32.8	31.4	32.1	88.48	15.1	12.4	11.2	91.46	19.4	15.7	12.2
	DNE	94.09	34.0	33.6	52.3	89.97	25.7	23.0	20.6	93.97	33.3	31.2	**43.8**
	SAFER	**94.42**	39.3	35.5	42.3	**92.26**	**41.4**	39.1	30.7	**95.39**	29.8	25.8	23.7
	TMD	94.29	**70.0**	**50.0**	**55.2**	92.17	38.7	**44.2**	**33.7**	95.24	**36.8**	**40.9**	28.6
RoBERTa	Vanilla	**95.04**	44.1	34.5	44.5	93.24	4.4	1.0	0.1	96.64	37.0	20.1	9.0
	ASCC	92.62	48.1	41.0	49.1	92.62	23.1	13.5	11.8	95.42	15.0	8.6	4.5
	DNE	94.93	58.0	46.5	54.5	**94.20**	48.8	26.9	16.0	**96.76**	64.4	64.0	45.2
	SAFER	94.58	51.3	41.9	46.1	93.92	52.8	47.1	40.6	96.59	65.6	67.9	48.3
	TMD	95.03	**68.3**	**54.0**	**56.7**	93.26	**60.5**	**66.8**	**51.6**	96.62	**68.9**	**70.9**	**51.0**
XLNet	Vanilla	94.80	34.4	28.0	37.9	93.59	7.1	2.3	1.4	96.23	28.0	14.0	7.2
	ASCC	92.64	38.6	33.4	41.6	92.57	15.8	11.1	10.5	95.52	39.3	24.2	12.8
	DNE	**94.99**	48.8	38.4	44.1	93.53	**42.9**	33.2	**26.6**	**96.64**	55.4	53.0	41.3
	SAFER	93.87	40.4	32.1	38.1	93.48	22.7	16.7	6.7	96.17	48.9	36.7	23.7
	TMD	94.57	**66.9**	**54.2**	**56.5**	**93.62**	25.3	**37.9**	21.3	96.17	**61.4**	**64.8**	**51.2**

Table 1: The robustness of different defenses on AGNEWS, IMDB, and YELP. We denote the clean accuracy, accuracy under PWWS, TextFooler, Bert-Attack as CA, PW, TF, BA, respectively. The best performance for each model is **bolded**, and the second-best performance is underlined.

Baseline Defenses We compare our method with other families of defenses. For adversarial training defenses, we choose the Adversarial Sparse Convex Combination (ASCC) (Dong et al., 2021a) and Dirichlet Neighborhood Ensemble (DNE) (Zhou et al., 2021). Both methods model the perturbation space as the convex hull of word synonyms. The former introduces an entropy-based sparsity regularizer to better capture the geometry of real word substitutions. The latter expands the convex hull for a word x to cover all synonyms of x's synonyms and combines Dirichlet sampling in this perturbation space with adversarial training to improve the model robustness. For certified defenses, we choose SAFER (Ye et al., 2020), which is a method based on the randomized smoothing technique. Given an input sentence, SAFER constructs a set of randomized inputs by applying random synonym substitutions and leveraging statistical properties of the predicted labels to certify the robustness.

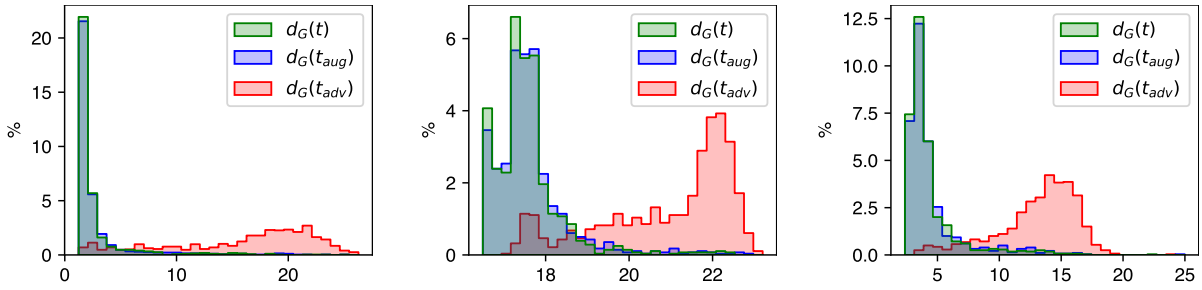
Implementation Details In practice, we precompute the text embeddings for all inputs to reduce the computational cost. For BERT and RoBERTa, we use the [CLS] representation as the text embedding. For XLNet, we use the last token's representation. More implementation details such as training InfoGAN, choosing hyperparameters can be found in Appendix A.

4.2 Main Results

To evaluate the robustness of different defenses, we randomly select 1000 samples from the test set and evaluate their accuracy under attacks. For the clean accuracy, we evaluate it on the entire test set.

The results for the AGNEWS dataset are shown in Table 1. We denote the "Vanilla" method as the original model without any defense mechanism. As we can see from the results, TMD outperforms previous methods under various settings by a large margin. Despite a slight decline in the clean accuracy, TMD achieves state-of-the-art robustness for BERT, RoBERTa, and XLNet with 23.03%, 18.63%, and 25.77% average performance gain over all attacks, respectively.

Interestingly, slightly different trends are found in the IMDB and YELP datasets. First of all, all models are generally more vulnerable to adversarial examples. This could be explained by the long average sentence length in IMDB (313.87 words) and YELP (179.18 words). This value is much larger than the AGNEWS, about 53.17 words. Longer sentences result in less restricted perturbation space for the attacker to perform word-substitution attacks, hence increasing the attack success rate. Regarding robustness, our method outperforms other methods in the majority of cases.



(a) BERT (b) RoBERTa (c) XLNet

Figure 2: The distributions of distance-to-manifold of clean, augmented, and adversarial embeddings in the IMDB dataset. The adversarial examples are generated using BERT-Attack (Li et al., 2020). Augmented examples are generated using random synonym substitution.

4.3 Testing the Off-Manifold Conjecture in NLP

Equipped with a way to approximate the contextualized embedding manifold, we aim to validate the off-manifold conjecture in the NLP domain. To achieve this, we first define the distance of an embedding t to the approximated manifold $G(z, c)$ as

$$d_G(t) = \left\| t - \hat{t} \right\|_2 \tag{9}$$

where $\hat{t}$ is the on-manifold projection computed from Equation 8. The off-manifold conjecture states that adversarial examples tend to leave the underlying natural data manifold. Therefore, if the off-manifold conjecture holds, we should observe small $d_G(t)$ for clean examples, while adversarial examples t_{adv} have large values of $d_G(t_{adv})$.

For each sentence x in the test set, we find its corresponding adversarial example x_{adv}. Additionally, to ensure that large $d_G(t_{adv})$ is not simply caused by word substitution, we randomly substitute x with synonyms to make an augmented sentence x_{aug} and see if the resulted distribution $d_G(t_{aug})$ diverges away from $d_G(t)$. We set the modifying ratio equal to the average percentage of perturbed words in x_{adv} for a fair comparison. The distributions of $d_G(t)$, $d_G(t_{aug})$, and $d_G(t_{adv})$ are visualized in Figure 2.

From the figure, we can see a clear divergence between the distribution of $d_G(t)$ and $d_G(t_{adv})$ on all models. Furthermore, the distribution $d_G(t_{aug})$ remains nearly identical to $d_G(t)$. This shows that simple word substitution does not cause the embedding to diverge off the natural manifold. Additionally, it is important to emphasize that since all x are unseen examples from the test set, low values of $d_G(t)$ are not simply due to overfitting in the generative model. These results provide empirical evidence to support the off-manifold conjecture in NLP.

Method		RL	CLN	AUA
BERT	TMD-InfoGAN	**3.121**	**92.15**	**33.70**
	TMD-DCGAN	3.707	92.04	4.30
RoBERTa	TMD-InfoGAN	**17.678**	**93.24**	**51.60**
	TMD-DCGAN	19.644	93.07	2.90
XLNet	TMD-InfoGAN	**4.679**	**93.59**	**19.20**
	TMD-DCGAN	7.056	93.47	4.00

Table 2: The accuracy under attack comparison between TMD-InfoGAN with disconnected support and TMD-DCGAN with connected support under BERT-Attack. We denote the reconstruction loss, clean accuracy, and accuracy under attack as RL, CLN, and AUA, respectively. All results are measured on the IMDB dataset.

4.4 The Importance of Disconnectedness in Contextualized Manifold Approximation

In this experiment, we study the significance of using disconnected generative models in TMD to improve robustness. We replace InfoGAN with DCGAN (Radford et al., 2016), which has a Gaussian prior distribution $z \sim \mathcal{N}(0, I)$ supported on the connected space $\mathbb{R}^d$ and has the same backbone as InfoGAN for a fair comparison. We then measure its reconstruction capability, the resulted generalization, robustness and compare the differences between the two versions of TMD. The reconstruction loss is defined similarly as Equation 9.

As shown in Table 2, DCGAN performs worse than InfoGAN on approximating the contextualized embedding manifold of all language models, which leads to a degradation in clean accuracy and a significant drop in robustness against adversarial attacks. The robustness of BERT, RoBERTa, and XLNET under TMD-DCGAN drop by 87.24%,

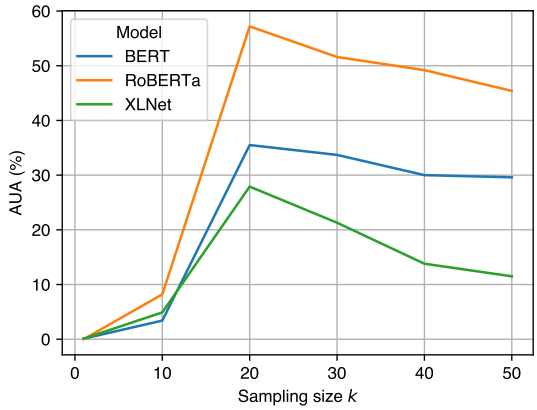


Figure 3: The relationship between the sampling size k and the accuracy under BERT-Attack for different language models.

94.38%, and 79.17%, respectively. This result is consistent with previous works on approximating disconnected manifold (Khayatkhoei et al., 2018; Tanielian et al., 2020) and shows that using disconnected generative models is crucial for robustness improvement in TMD.

4.5 The Effect of Sampling Size k on Robustness

We now study the effect of different k values on TMD performance. From Figure 3, we can see an optimal value for k across all models. A larger sampling size does not correspond with better robustness but slightly degrades it. We hypothesize that when k is set too large, z may be sampled from low-probability regions in the latent space, producing smaller reconstruction loss but not necessarily lying on the true embedding manifold. On the other hand, when k is too small, it affects the reconstruction quality of InfoGAN, and the projected embedding may not well reassemble the original one.

4.6 Comparison with Optimization-based Reconstruction Strategy

Method	BERT-Attack
TMD	33.70
TMD-GD ($\alpha = 1, N = 10$)	26.50
TMD-GD ($\alpha = 0.1, N = 10$)	33.10
TMD-GD ($\alpha = 0.01, N = 10$)	30.60
TMD-GD ($\alpha = 0.001, N = 10$)	31.20

Table 3: Comparison between sampling-based and optimization-based reconstruction strategies

In addition to the sampling-based reconstruction, we also experiment with the optimization-based approach, one of the most common methods in the GAN inversion literature (Xia et al., 2021; Creswell and Bharath, 2019; Abdal et al., 2019). Particularly, we evaluate the effectiveness of a reconstruction method similar to DefenseGAN Samangouei et al. (2018). Given an embedding t, we sample k initial values for z from the prior distribution. We then perform N steps of gradient descent (GD) with step size α separately on each k value of z to minimize the reconstruction loss. The optimal latent variable z is chosen to compute the final on-manifold projection $\hat{t}$. To put it shortly, this reconstruction method extends our method in Equation 8 with an additional step of GD optimization. We refer to this version of TMD as TMD-GD and compare several hyperparameter settings of this optimization-based approach with our sampling-based approach.

As can be seen in Table 3, the TMD-GD reconstruction is outperformed by our sampling-based approach. We hypothesize that since TMD-GD does not consider the probability over the latent space, it can move the latent variables to low-probability regions, producing embeddings that may not lie on the true natural manifold. This problem is similar to sampling-based reconstruction with too large sampling size. Another drawback of TMD-GD is the additional computational cost introduced by GD, which can negatively affect the inference phase.

5 Conclusion

In this paper, we show empirical evidence that the manifold assumption indeed holds for the textual domain. Based on this observation, we propose a novel method that projects input embeddings onto the approximated natural embedding manifold before classification to defend against adversarial examples. Extensive experiments show that our method consistently outperforms previous defenses by a large margin under various adversarial settings. Future research on developing better manifold approximation and on-manifold projection methods are interesting directions to further improve the robustness of this type of defense. We hope that the findings in this work can enable broader impacts on improving NLP robustness.

Limitations

Despite the many advantages of TMD, it still has some limitations that can be improved. One prob-

lem, in particular, is to reduce the computational overhead of on-manifold projection. Since we are adding an additional reconstruction step, it adds latency to the inference phase. Other interesting questions that have not been fully addressed in this work due to time constraints include the effect of TMD when applying reconstruction to intermediate layers, alternative methods to construct text embeddings (e.g., by averaging all token embeddings instead of using [CLS] token), more sophisticated choices of manifold approximation models and reconstruction methods. These are interesting research directions that can extend the understanding and effectiveness of TMD.

Acknowledgements

This work is supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 (RS21/20).

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space? In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4431–4440. IEEE.

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

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

Anish Athalye, Nicholas Carlini, and David Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283. PMLR.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2172–2180.

Antonia Creswell and Anil Anthony Bharath. 2019. Inverting the generator of a generative adversarial network. *IEEE Trans. Neural Networks Learn. Syst.*, 30(7):1967–1974.

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

Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. 2021a. Towards robustness against natural language word substitutions. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Xinshuai Dong, Anh Tuan Luu, Min Lin, Shuicheng Yan, and Hanwang Zhang. 2021b. How should pre-trained language models be fine-tuned towards adversarial robustness? *Advances in Neural Information Processing Systems*, 34:4356–4369.

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

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

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049. Funding Information: The first author was supported by NSF grant DMS 1265524, AFOSR grant FA9550-12-1-0425 and U.S.-Israel Binational Science Foundation grant 2014055. The

second author was supported by NSF grant EECS-1135843. Publisher Copyright: © 2016 American Mathmatical Society.

J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian J. Goodfellow. 2018. Adversarial spheres. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

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*.

Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R. Venkatesh Babu. 2017. Deligan: Generative adversarial networks for diverse and limited data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4941–4949. IEEE Computer Society.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. 2019. Achieving verified robustness to symbol substitutions via interval bound propagation. 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)*, pages 4083–4093, Hong Kong, China. Association for Computational Linguistics.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. 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 125–136.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. 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)*, pages 4129–4142, Hong Kong, China. Association for Computational Linguistics.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025.

Mahyar Khayatkhoei, Maneesh Singh, and Ahmed Elgammal. 2018. Disconnected manifold learning for generative adversarial networks. 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 7354–7364.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. 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*. The Internet Society.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Zongyi Li, Jianhan Xu, Jiehang Zeng, Linyang Li, Xiaoqing Zheng, Qi Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021. Searching for an effective defender: Benchmarking defense against adversarial word substitution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3137–3147, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Jonathan Mamou, Hang Le, Miguel Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2020. Emergence of separable manifolds in deep language representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6713–6723. PMLR.

Dongyu Meng and Hao Chen. 2017. Magnet: A two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*, pages 135–147. ACM.

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*. OpenReview.net.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California. Association for Computational Linguistics.

Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

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 Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Marco Tulio 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 (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. 2018. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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*, volume 31. Curran Associates, Inc.

Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. 2019. Towards the first adversarially robust neural network model on MNIST. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Adi Shamir, Odelia Melamed, and Oriel BenShmuel. 2021. The dimpled manifold model of adversarial examples in machine learning. *CoRR*, abs/2106.10151.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. 2020. Robustness verification for transformers. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1569–1576, Online. Association for Computational Linguistics.

Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. 2017. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *CoRR*, abs/1710.10766.

David Stutz, Matthias Hein, and Bernt Schiele. 2019. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Thomas Tanay and Lewis Griffin. 2016. A boundary tilting persepective on the phenomenon of adversarial examples.

Ugo Tanielian, Thibaut Issenhuth, Elvis Dohmatob, and Jérémie Mary. 2020. Learning disconnected manifolds: a no gan's land. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9418–9427. PMLR.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Patrick Xia, Shijie Wu, and Benjamin Van Durme. 2020. Which *BERT? A survey organizing contextualized encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7516–7533, Online. Association for Computational Linguistics.

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2021. GAN inversion: A survey. *CoRR*, abs/2101.05278.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mao Ye, Chengyue Gong, and Qiang Liu. 2020. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3465–3475, Online. Association for Computational Linguistics.

Jiehang Zeng, Xiaoqing Zheng, Jianhan Xu, Linyang Li, Liping Yuan, and Xuanjing Huang. 2021. Certified robustness to text adversarial attacks by randomized [MASK]. *CoRR*, abs/2105.03743.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015b. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Haiteng Zhao, Chang Ma, Xinshuai Dong, Anh Tuan Luu, Zhi-Hong Deng, and Hanwang Zhang. 2022. Certified robustness against natural language attacks by causal intervention. *arXiv preprint arXiv:2205.12331*.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. 2021. Defense against synonym substitution-based adversarial attacks via Dirichlet neighborhood ensemble. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5482–5492, Online. Association for 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*. OpenReview.net.

A Implementation Details

A.1 Preparing Datasets

We use off-the-shelf datasets from HuggingFace Datasets (Lhoest et al., 2021). Based on the average text length of each dataset, we set the model's max length to 128, 256, and 256 for AGNEWS, IMDB, and YELP, respectively.

A.2 Finetuning Language Models to Downstream Tasks

For pretrained language models, we utilize the base models provided by HuggingFace Transformers (Wolf et al., 2020). We then finetune them on downstream tasks for ten epochs. The optimal learning rates for each pair of datasets and model are achieved using a simple grid search from 1e-6 to 9e-4. The optimally finetuned models are kept for robustness evaluation.

A.3 Training InfoGAN

Hyperparameter	Values
α_g	1e-5, 3e-5, 5e-5, 7e-5, 9e-5, 1e-4, 3e-4, 5e-4
α_d	7e-5, 9e-5, 1e-4, 3e-4, 5e-4, 7e-4, 9e-4, 1e-3
α_p	0, 1e-4, 1e-3, 1e-2
K	50, 100, 200
d	10, 20, 30, 40
d/g	1, 2, 3, 4, 5

Table 4: The value ranges for random search on hyperparameters, where α_g, α_d, α_p, K, d, d/g denote the G's learning rate, D's learning rate, Prior's learning rate, latent variable z dimension, latent code c dimension, and the number of D's iterations per G updates.

Since InfoGAN uses the DCGAN backbone, we find it relatively stable during training. The only additional training trick we employ is tuning the ratio of discriminator iterations per generator update. We perform a random search on the hyperparameters, where their value ranges are shown in Table 4. Additionally, the number of training epochs is set to 100 for all experiments. In general, we find that the most significant hyperparameters are α_g, α_d, and d/g, while α_p, K, and d do not contribute too much to the final robustness. Any reasonably large value for d dimension is sufficient since redundant c values eventually vanished during prior learning. The optimal hyperparameters after random search are shown in Table 6. The detailed architectures for G and D are shown in Tables 7 and 8.

B Additional Experiments

B.1 Runtime Analysis

Defense	Runtime (s)
Vanilla	27.21
ASCC	460.76
SAFER	27.07
DNE	59.77
TMD	56.19

Table 5: Inference speed comparison with other defenses. Tested with BERT on 1000 samples from the IMDB dataset.

In this experiment, we want to measure how much overhead TMD introduces to the inference phase compared to other defenses. We sample 1000 inputs from the IMDB dataset and record the inference speed of BERT when equipped with different defenses. The results are shown in Table 5. Despite an additional latency in the inference phase, TMD still achieves competitive performance.

B.2 TMD on Varying Model Sizes

In this experiment, we want to support the significance of TMD by showing its generalization across different model architectural sizes. Since larger versions of language models often use a hidden size of 1024, we have to employ a larger version of InfoGAN to adopt the larger embeddings. The results are shown in Table 9. As we can see from the table, combining TMD with larger models results in impressive robustness against various attacks.

	AGNEWS						IMDB						YELP					
	α_g	α_d	α_p	d/g	K	d	α_g	α_d	α_p	d/g	K	d	α_g	α_d	α_p	d/g	K	d
BERT	1e-4	1e-4	0	1	50	20	1e-4	1e-4	0	1	50	40	1e-4	9e-5	0	1	100	20
RoBERTa	2e-4	2e-4	1e-2	1	100	20	3e-4	3e-4	0	1	200	20	3e-5	1e-3	1e-4	2	100	20
XLNet	1e-4	1e-4	1e-4	1	100	20	1e-4	1e-4	1e-4	1	100	20	5e-4	9e-5	1e-3	5	100	30

Table 6: Optimal hyperparameters after random search.

Operation	Kernel	Strides	Padding	Feature Maps	Batch Norm.	Activation	Shared?
$G(z, c) : z \sim P(z), c \sim P(c)$				$(K + d) \times 1$			
ConvTranspose1d	32×32	1×1	0	512×32	Y	ReLU	N
ConvTranspose1d	4×4	2×2	1	384×64	Y	ReLU	N
ConvTranspose1d	4×4	2×2	1	256×128	Y	ReLU	N
ConvTranspose1d	4×4	2×2	1	128×256	Y	ReLU	N
ConvTranspose1d	5×5	3×3	1	1×768	N	Tanh	N

Table 7: Model architecture of the Generator Network G

Operation	Kernel	Strides	Padding	Feature Maps	Batch Norm.	Activation	Shared?
$D(t), Q(t)$				1×768			
Conv1d	5×5	3×3	1	128×256	N	LeakyReLU (slope = 0.2)	Y
Conv1d	4×4	2×2	1	256×128	Y	LeakyReLU (slope = 0.2)	Y
Conv1d	4×4	2×2	1	384×64	Y	LeakyReLU (slope = 0.2)	Y
Conv1d	4×4	2×2	1	512×32	Y	LeakyReLU (slope = 0.2)	Y
Conv1d	4×4	2×2	1	768×16	Y	LeakyReLU (slope = 0.2)	Y
D Conv1d	16×16	1×1	0	1×1	N	Sigmoid	N
Q Fully Connected				$1 \times K$	N	Softmax	N

Table 8: Model architecture of the Discriminator D and Auxiliary Network Q

Model	Defense	CA	PW	TF	BA
BERT-base	Vanilla	92.15	6.50	2.30	1.00
	TMD	**92.17**	**38.70**	**44.20**	**33.70**
BERT-large	Vanilla	93.04	29.30	21.10	16.50
	TMD	**93.14**	**50.10**	**58.20**	**44.10**
RoBERTa-base	Vanilla	93.24	4.40	1.00	0.10
	TMD	**93.26**	**60.50**	**66.80**	**51.60**
RoBERTa-large	Vanilla	**95.05**	31.73	12.01	4.10
	TMD	94.95	**70.60**	**74.77**	**61.60**

(a) IMDB

Model	Defense	CA	PW	TF	BA
BERT-base	Vanilla	**94.39**	39.20	27.90	39.00
	TMD	94.29	**70.00**	**50.00**	**55.20**
BERT-large	Vanilla	**94.59**	30.30	20.80	26.20
	TMD	92.83	**56.40**	**30.50**	**30.00**
RoBERTa-base	Vanilla	**95.04**	44.10	34.50	44.50
	TMD	95.03	**68.30**	**54.00**	**56.70**
RoBERTa-large	Vanilla	**95.34**	52.40	35.40	40.70
	TMD	94.88	**71.80**	**50.80**	**51.20**

(b) AGNEWS

Table 9: Performance on different architectural sizes.