Improved OOD Generalization via Adversarial Training and Pre-training

Mingyang Yi [1 2 †] **Lu Hou** [3] **Jiacheng Sun** [3] **Lifeng Shang** [3] **Xin Jiang** [3] **Qun Liu** [3] **Zhi-Ming Ma** [1 2]

Abstract

Recently, learning a model that generalizes well on out-of-distribution (OOD) data has attracted great attention in the machine learning community. In this paper, after defining OOD generalization via Wasserstein distance, we theoretically show that a model robust to input perturbation generalizes well on OOD data. Inspired by previous findings that adversarial training helps improve input-robustness, we theoretically show that adversarially trained models have converged excess risk on OOD data, and empirically verify it on both image classification and natural language understanding tasks. Besides, in the paradigm of first pre-training and then fine-tuning, we theoretically show that a pre-trained model that is more robust to input perturbation provides a better initialization for generalization on downstream OOD data. Empirically, after fine-tuning, this better-initialized model from adversarial pre-training also has better OOD generalization.

1. Introduction

In the machine learning community, the training and test distributions are often not identically distributed. Due to this mismatching, it is desired to learn a model that generalizes well on out-of-distribution (OOD) data though only trained on data from one certain distribution. OOD generalization is empirically studied in (Hendrycks et al., 2019; 2020a;b) by evaluating the performance of the model on the test set that is close to the original training samples. However, the theoretical understanding of these empirical OOD generalization behaviors remains unclear.

Intuitively, the OOD generalization measures the performance of the model on the data from a shifted distribution around the original training distribution (Hendrycks & Dietterich, 2018). This is equivalent to the distributional robustness (Namkoong, 2019; Shapiro, 2017) which measures the model's robustness to perturbations the distribution of training data. Inspired by this, we study the OOD generalization by utilizing the Wasserstein distance to measure the shift between distributions (Definition 1). We theoretically find that if a model is robust to input perturbation on training samples (namely, input-robust model), it also generalizes well on OOD data.

The connection of input-robustness and OOD generalization inspires us to find an input-robust model since it generalizes well on OOD data. Thus we consider adversarial training (AT) (Madry et al., 2018) as Athalye et al. (2018) show that a model is input-robust if it defends adversarial perturbations (Szegedy et al., 2013). Mathematically, AT can be formulated as a minimax optimization problem and solved by the multi-step SGD algorithm (Nouiehed et al., 2019). Under mild assumptions, we prove that the convergence rate of this multi-step SGD for AT is $\tilde{\mathcal{O}}(1/T)$ both in expectation and in high probability, where T is the number of training steps and $\tilde{\mathcal{O}}(\cdot)$ is defined in the paragraph of notations. Then, combining the convergence result with the relationship between input-robustness and OOD generalization, we theoretically show that for the model adversarially trained with n training samples for T steps, its excess risk on the OOD data is upper bounded by $\tilde{\mathcal{O}}(1/\sqrt{n} + 1/T)$, which guarantees its performance on the OOD data.

Besides models trained from scratch, we also study the OOD generalization on downstream tasks of pre-trained models, as the paradigm of first pre-training on a large-scale dataset and then fine-tuning on downstream tasks has achieved remarkable performance in both computer vision (CV) (Hendrycks et al., 2019; Kornblith et al., 2019) and natural language processing (NLP) domains (Devlin et al., 2019) recently. Given the aforementioned relationship of input-robustness and OOD generalization, we theoretically show that a pre-trained model more robust to input perturbation also provides a better initialization for generalization on downstream OOD data. Thus, we suggest conducting adversarial pre-training like (Salman et al., 2020a; Hendrycks et al., 2019; Utrera et al., 2021), to improve the OOD generalization in downstream tasks.

---

[*] Equal contribution [1] University of Chinese Academy of Sciences, Beijing, China [2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China [3] Huawei Noah's Ark Lab, Shenzhen, China. Correspondence to: Mingyang Yi <yimingyang17@mails.ucas.edu.cn>, Lu Hou <houlu3@huawei.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

We conduct various experiments on both image classification (IC) and natural language understanding (NLU) tasks to verify our theoretical findings.

For IC task, we conduct AT on `CIFAR10` (Krizhevsky & Hinton, 2009) and `ImageNet` (Deng et al., 2009), and then evaluate the OOD generalization of these models on corrupted OOD data `CIFAR10-C` and `ImageNet-C` (Hendrycks & Dietterich, 2018). For NLU tasks, we similarly conduct AT as in (Zhu et al., 2019) on datasets `SST-2`, `IMBD`, `MNLI` and `STS-B`. Then we follow the strategy in (Hendrycks et al., 2020b) to evaluate the OOD generalization. Empirical results on both IC and NLU tasks verify that AT improves OOD generalization.

To see the effect of the initialization provided by an input-robust pre-trained model, we adversarially pre-train a model on `ImageNet` to improve the input-robustness, and then fine-tune the pre-trained model on `CIFAR10`. Empirical results show that this initialization enhances the OOD generalization on downstream tasks after fine-tuning. Another interesting observation is that for language models, standard pre-training by masked language modeling (Devlin et al., 2019; Liu et al., 2019) improves the input-robustness of the model. Besides, models pre-trained with more training samples and updating steps are more input-robust. This may also explain the better OOD generalization on downstream tasks (Hendrycks et al., 2020b) of these models.

Notations. For vector $\boldsymbol{x} \in \mathbb{R}^{d_0}$, $\|\boldsymbol{x}\|_p$ is its ℓ_p-norm, and its ℓ_2-norm is simplified as $\|\boldsymbol{x}\|$. $\mathcal{P}(\mathcal{X})$ is the set of probability measures on metric space $(\mathcal{X}, \|\cdot\|_p)$ with $\mathcal{X} \subseteq \mathbb{R}^{d_0}$. $\mathcal{O}(\cdot)$ is the order of a number, and $\tilde{\mathcal{O}}(\cdot)$ hides a poly-logarithmic factor in problem parameters e.g., $\mathcal{O}(M_1 \log d_0) = \tilde{\mathcal{O}}(M_1)$. For $P, Q \in \mathcal{P}(\mathcal{X})$, let (P, Q) be their couplings (measures on $\mathcal{X} \times \mathcal{X}$). The p-th ($p < \infty$) Wasserstein distance (Villani, 2008) between P and Q is

$$\mathsf{W}_p(P, Q) = \left(\inf_{\pi \in (P,Q)} \mathbb{E}_{(\boldsymbol{u},\boldsymbol{v})\sim\pi}\left[\|\boldsymbol{u} - \boldsymbol{v}\|_p^p\right]\right)^{\frac{1}{p}}. \tag{1}$$

When $p = \infty$, the ∞-Wasserstein distance is $\mathsf{W}_\infty(P, Q) = \lim_{p\to\infty} \mathsf{W}_p(P, Q)$. In the sequel, the p-Wasserstein distance is abbreviated as W_p-distance. The total variation distance (Villani, 2008) is a kind of distributional distance and is defined as

$$\mathsf{TV}(P, Q) = \frac{1}{2}\int_{\mathcal{X}} |dP(\boldsymbol{x}) - dQ(\boldsymbol{x})|. \tag{2}$$

2. Related Work

OOD Generalization. OOD generalization measures a model's ability to extrapolate beyond the training distribution (Hendrycks & Dietterich, 2018), and has been widely explored in both CV (Recht et al., 2019; Schneider et al., 2020; Salman et al., 2020b) and NLP domains (Tu et al., 2020; Lohn, 2020). Hendrycks & Dietterich (2018) observe that the naturally trained models are sensitive to artificially constructed OOD data. They also find that adversarial logit pairing (Kannan et al., 2018) can improve a model's performance on noisy corrupted OOD data. Hendrycks et al. (2020b) also empirically find that pre-trained language models generalize on downstream OOD data. But the theoretical understanding behind these observations remains unclear.

Adversarial Training. Adversarial training (Madry et al., 2018) is proposed to improve input-robustness by dynamically constructing the augmented adversarial samples (Szegedy et al., 2013; Goodfellow et al., 2015) using projected gradient descent across training. In this paper, we first show the close relationship between OOD generalization and distributional robustness (Ben-Tal et al., 2013; Shapiro, 2017), and then explore the OOD generalization by connecting input-robustness and distributional robustness.

The most related works to ours are (Sinha et al., 2018; Lee & Raginsky, 2018; Volpi et al., 2018). They also use AT to train distributionally robust models under the Wasserstein distance, but their results are restricted to a specialized AT objective with an additional regularizer. The regularizer can be impractical due to its large penalty parameter. Moreover, their bounds are built upon the entropy integral and increase with model capacity, which can be meaningless for high-dimensional models. On the other hand, our bound is (i) based on the input-robustness, regardless of how it is obtained; and (ii) irrelevant to model capacity.

Pre-Training. Pre-trained models transfer the knowledge in the pre-training stage to downstream tasks, and are widely used in both CV (Kornblith et al., 2019) and NLP (Devlin et al., 2019) domains. For instance, Dosovitskiy et al. (2021); Brown et al. (2020); Radford et al. (2021) pre-train the transformer-based models on large-scale datasets, and obtain remarkable results on downstream tasks. Standard pre-training is empirically found to help reduce the uncertainty of the model for both image data (Hendrycks et al., 2019; 2020a) and textual data (Hendrycks et al., 2020b). Adversarial pre-training is explored in (Hendrycks et al., 2019) and (Salman et al., 2020a), and is shown to improve the robustness and generalization on downstream tasks , respectively. In this work, we theoretically analyze the OOD generalization on downstream tasks from the perspective of the input-robustness of the pre-trained model.

3. Adversarial Training Improves OOD Generalization

In this section, we first show that the input-robust model can generalize well on OOD data after specifying the definition of OOD generalization. Then, to learn a robust model, we suggest adversarial training (AT) (Madry et al., 2018). Under mild conditions, we prove a $\tilde{\mathcal{O}}(1/T)$ convergence rate for AT both in expectation and in high probability. With this, we show that the excess risk of an adversarially trained model on OOD data is upper bounded by $\tilde{\mathcal{O}}(1/\sqrt{n} + 1/T)$ where n is the number of training samples.

3.1. Input-Robust Model Generalizes on OOD Data

Suppose $\{(\boldsymbol{x}_i, y_i)\}$ is the training set with n i.i.d. training samples $\{\boldsymbol{x}_i\}$ and their labels $\{y_i\}$. We assume the training sample distribution P has compact support $\mathcal{X} \subseteq \mathbb{R}^{d_0}$, thus there exists $D > 0$, such that $\forall \boldsymbol{u}, \boldsymbol{v} \in \mathcal{X}$, $\|\boldsymbol{u} - \boldsymbol{v}\|_1 \leq D$. For training sample $\boldsymbol{x}$ and its label y, the loss on $(\boldsymbol{x}, y)$ with model parameter $\boldsymbol{w}$ is $\mathcal{L}(\boldsymbol{w}, (\boldsymbol{x}, y))$, where $\mathcal{L}(\boldsymbol{w}, (\boldsymbol{x}, y))$ is continuous and differentiable for both $\boldsymbol{w}$ and $(\boldsymbol{x}, y)$. Besides, we assume $0 \leq \mathcal{L}(\boldsymbol{w}, (\boldsymbol{x}, y)) \leq M$ for constant M without loss of generality. We represent the expected risk under training distribution P and label distribution $P_{y|\boldsymbol{x}}$ [1] as $R_P(\boldsymbol{w}) = \mathbb{E}_P[\mathbb{E}_{P_{y|\boldsymbol{x}}}[\mathcal{L}(\boldsymbol{w}, (\boldsymbol{x}, y))]]$. For simplicity of notation, let $\mathbb{E}_{P_{y|\boldsymbol{x}}}[\mathcal{L}(\boldsymbol{w}, (\boldsymbol{x}, y))] = f(\boldsymbol{w}, \boldsymbol{x})$ in the sequel, e.g., $f(\boldsymbol{w}, \boldsymbol{x}_i) = \mathcal{L}(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$.

Intuitively, the OOD generalization is decided by the performance of the model on a shifted distribution close to the training data-generating distribution P_0 (Hendrycks & Dietterich, 2018; Hendrycks et al., 2020b). Thus defining OOD generalization should involve the distributional distance which measures the distance between distributions. We use the Wasserstein distance as in (Sinha et al., 2018).

Let $P_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\cdot = \boldsymbol{x}_i\}}$ be the empirical distribution, and $B_{\mathsf{W}_p}(P_0, r) = \{P : \mathsf{W}_p(P_0, P) \leq r\}$. Then we define the OOD generalization error as

$$\mathcal{E}_{\text{gen}}^{\text{ood}}(p, r) = \left| \sup_{P \in B_{\mathsf{W}_p}(P_0, r)} R_P(\boldsymbol{w}) - R_{P_n}(\boldsymbol{w}) \right|, \qquad (3)$$

under the W_p-distance with $p \in \{2, \infty\}$. Extension to the other OOD generalization with $p < \infty$ is straightforward by generalizing the analysis for $p = 2$. Note that (3) reduces to the generalization error on in-distribution data when $r = 0$.

Definition 1. *A model is (r, ϵ, P, p)-input-robust, if*

$$\mathbb{E}_P\left[\sup_{\|\boldsymbol{\delta}\|_p \leq r} |f(\boldsymbol{w}, \boldsymbol{x} + \boldsymbol{\delta}) - f(\boldsymbol{w}, \boldsymbol{x})|\right] \leq \epsilon. \qquad (4)$$

With the input-robustness in Definition 1, the following Theorems 1 and 2 give the generalization bounds on the OOD data drawn from $Q \in B_{\mathsf{W}_p}(P_0, r_0)$ with $p \in \{2, \infty\}$.

Theorem 1. *If a model is $(2r, \epsilon, P_n, \infty)$-input-robust, then with probability at least $1 - \theta$,*

$$\mathcal{E}_{\text{gen}}^{\text{ood}}(\infty, r_0) \leq \epsilon + M\sqrt{\frac{(2d_0)^{\frac{2D}{r^2}+1}\log 2 + 2\log\left(\frac{1}{\theta}\right)}{n}}, \qquad (5)$$

for any $r_0 \leq r$. Here D is the ℓ_1-diameter of data support $\mathcal{X}$ with dimension d_0, and M is an upper bound of $f(\boldsymbol{w}, \boldsymbol{x})$.

Theorem 2. *If a model is $(2r/\epsilon, \epsilon, P_n, 2)$-input-robust, then with probability at least $1 - \theta$,*

$$\mathcal{E}_{\text{gen}}^{\text{ood}}(2, r_0) \leq (M+1)\epsilon + M\sqrt{\frac{(2d_0)^{\frac{2\epsilon^2 D}{r^2}+1}\log 2 + 2\log\left(\frac{1}{\theta}\right)}{n}}, \qquad (6)$$

for any $r_0 \leq r$, where the notations follow Theorem 1.

Remark 1. *When $r_0 = 0$, the bounds in Theorems 1 and 2 become the generalization bounds on in-distribution data.*

Remark 2. *The ϵ in Theorem 2 can not be infinitely small, as the model is required to be robust in $B(\boldsymbol{x}_i, 2r/\epsilon)$ for each $\boldsymbol{x}_i$. Specifically, when $\epsilon \to 0$, the robust region $B(\boldsymbol{x}_i, 2r/\epsilon)$ can cover the data support $\mathcal{X}$, then the model has almost constant output in $\mathcal{X}$.*

Remark 3. *The bounds (5) and (6) become vacuous when r is large. Thus, our results can not be applied to those OOD data from distributions far away from the original training distribution. For example, ImageNet-R (Hendrycks et al., 2020a) consists of data from different renditions e.g., photo vs. cartoon, where most pixels vary, leading to large $\|\boldsymbol{u} - \boldsymbol{v}\|_p^p$ in (1), and thus large distributional distance.*

The proofs of Theorems 1 and 2 are in Appendix A.1. Lemmas 1 and 2 in Appendix A show that the OOD data concentrates around the in-distribution data with high probability. Thus, the robustness of model on training samples guarantees the generalization on OOD data. The observations from Theorems 1 and 2 are summarized as follows.

1. The right-hand sides of bounds (5) and (6) imply that a more input-robust model (i.e., a larger r and a smaller ϵ in Definition 1) has smaller OOD generalization bound, and thus better performance on OOD data.

2. For both (5) and (6), a larger number of training samples n results in smaller upper bounds. This indicates that in a high-dimensional data regime with a large feature dimension d_0 of data and diameter D of data support, more training samples can compensate for generalization degradation caused by large d_0 and D.

3. The bounds (5) and (6) are independent of the model capacity. Compared with other uniform convergence generalization bounds which increase with the model capacity (e.g., Rademacher complexity (Yin et al., 2019) or entropy integral (Sinha et al., 2018)), our bounds are superior for models with high capacity.

---

[1] $P_{y|\boldsymbol{x}_i}(\cdot) = \mathbf{1}_{\{\cdot = y_i\}}$ where $\mathbf{1}_{\{\cdot = y_i\}}$ is the indicator function.

Algorithm 1 Multi-Step SGD.

```
Input: Number of training steps $T$, learning rate for model parameters $\eta_{\boldsymbol{w}_t}$ and adversarial input $\eta_{\boldsymbol{x}}$, two initialization points $\boldsymbol{w}_1, \boldsymbol{\delta}_1$, constant $p \in \{2, \infty\}$ and perturbation size $r$.
Return $\boldsymbol{w}_{T+1}$.
1: for $t = 1, \cdots, T$ do
2:     Uniformly sample $i_t$ from $\{1, \cdots, n\}$.
3:     for $k = 1, \cdots, K$ do
4:         $\boldsymbol{\delta}_{k+1} = \mathrm{Proj}_{B_p(\mathbf{0}, r)}\left(\boldsymbol{\delta}_k + \eta_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{w}_t, \boldsymbol{x}_{i_t} + \boldsymbol{\delta}_k)\right)$.
5:     end for
6:     $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_{\boldsymbol{w}_t} \nabla_{\boldsymbol{w}} f(\boldsymbol{w}_t, \boldsymbol{x}_{i_t} + \boldsymbol{\delta}_{K+1})$.
7: end for
```

3.2. Adversarial Training Improves Input-Robustness

As is justified in Theorems 1 and 2, the input-robust model can generalize on OOD data. Thus we consider training an input-robust model with the following objective

$$\begin{aligned}\min_{\boldsymbol{w}} \tilde{R}_{P_n}(\boldsymbol{w}, p) &= \min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \sup_{\|\boldsymbol{\delta}\|_p \le r(p)} f(\boldsymbol{w}, \boldsymbol{x}_i + \boldsymbol{\delta}) \\ &= \min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} [\underbrace{f(\boldsymbol{w}, \boldsymbol{x}_i)}_{\text{clean acc}} + \underbrace{\sup_{\|\boldsymbol{\delta}\|_p \le r(p)} (f(\boldsymbol{w}, \boldsymbol{x}_i + \boldsymbol{\delta}) - f(\boldsymbol{w}, \boldsymbol{x}_i))}_{\text{input-robustness}}],\end{aligned} \tag{7}$$

which is from AT (Madry et al., 2018), and can be decomposed into the clean accuracy term and the input-robustness term. We consider $p \in \{2, \infty\}$ as in Section 3.1, with $r(2) = 2r/\epsilon_0$, $r(\infty) = 2r$ for any given small constant ϵ_0.

Besides the general assumptions in Section 3.1, we also use the following mild assumptions in this subsection.

Assumption 1. *The loss $f(\boldsymbol{w}, \boldsymbol{x})$ satisfies the following Lipschitz smoothness conditions*

$$\begin{aligned}\|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}_1, \boldsymbol{x}) - \nabla_{\boldsymbol{w}} f(\boldsymbol{w}_2, \boldsymbol{x})\| &\le L_{11} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|, \\ \|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x}_1) - \nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x}_2)\| &\le L_{12} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \\ \|\nabla_{\boldsymbol{x}} f(\boldsymbol{w}_1, \boldsymbol{x}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{w}_2, \boldsymbol{x})\| &\le L_{21} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|, \\ \|\nabla_{\boldsymbol{x}} f(\boldsymbol{w}, \boldsymbol{x}_1) - \nabla_{\boldsymbol{x}} f(\boldsymbol{w}, \boldsymbol{x}_2)\| &\le L_{22} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|.\end{aligned} \tag{8}$$

Assumption 2. *$\|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x})\|$ is upper bounded by G.*

Assumption 3. *For $p \in \{2, \infty\}$, $\tilde{R}_{P_n}(\boldsymbol{w}, p)$ in (7) satisfies the PL-inequality:*

$$\frac{1}{2} \|\nabla_{\boldsymbol{w}} \tilde{R}_{P_n}(\boldsymbol{w}, p)\|^2 \ge \mu_{\boldsymbol{w}} \left(\tilde{R}_{P_n}(\boldsymbol{w}, p) - \inf_{\boldsymbol{w}} \tilde{R}_{P_n}(\boldsymbol{w}, p)\right). \tag{9}$$

For any $\boldsymbol{w}$ and training sample $\boldsymbol{x}_i$, $f(\boldsymbol{w}, \boldsymbol{x}_i + \boldsymbol{\delta})$ is $\mu_{\boldsymbol{x}_i}$-strongly concave in $\boldsymbol{\delta}$ for $\|\boldsymbol{\delta}\|_p \le r(p)$:

$$f(\boldsymbol{w}, \boldsymbol{x}_i + \boldsymbol{\delta}) - f(\boldsymbol{w}, \boldsymbol{x}_i) \le \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{w}, \boldsymbol{x}_i), \boldsymbol{\delta} \rangle - \frac{\mu_{\boldsymbol{x}_i}}{2} \|\boldsymbol{\delta}\|^2, \tag{10}$$

where $\mu_{\boldsymbol{w}}$ and $\mu_{\boldsymbol{x}_i}$ are constants.

Assumptions 1 and 2 are widely used in minimax optimization problems (Nouiehed et al., 2019; Sinha et al., 2018). PL-inequality in Assumption 3 means that although $f(\boldsymbol{w}, \boldsymbol{x})$ may be non-convex on $\boldsymbol{w}$, all the stationary points are global minima. This is observed or proved recently for over-parameterized neural networks (Xie et al., 2017; Du et al., 2019; Allen-Zhu et al., 2019; Liu et al., 2020). The local strongly-concavity in Assumption 3 is reasonable when the perturbation size $\|\boldsymbol{\delta}\|_p$ is small.

To solve the minimax optimization problem (7), we consider the multi-step stochastic gradient descent (SGD) in Algorithm 1 (Nouiehed et al., 2019). $\mathrm{Proj}_A(\cdot)$ in Algorithm 1 is the ℓ_2-projection operator onto A. Note that the update rule of $\boldsymbol{\delta}_k$ in Algorithm 1 is different from that in PGD adversarial training (Madry et al., 2018), where $\nabla_{\boldsymbol{x}} f(\boldsymbol{w}_t, \boldsymbol{x}_{i_t} + \boldsymbol{\delta}_k)$ in Line 4 is replaced with the sign of it.

The following theorem gives the convergence rate of Algorithm 1 both in expectation and in high probability.

Theorem 3. *Let $\boldsymbol{w}_t$ be updated by Algorithm 1, $p \in \{2, \infty\}$, $\eta_{\boldsymbol{w}_t} = \frac{1}{\mu_{\boldsymbol{w}} t}$, $\eta_{\boldsymbol{x}} = \frac{1}{L_{22}}$, $K \ge \frac{L_{22}}{\mu_{\boldsymbol{x}}} \log\left(\frac{8T\mu_{\boldsymbol{w}} d_0 r^2(p)}{GL}\right)$, where $\mu_{\boldsymbol{x}} = \min_{1 \le i \le n} \mu_{\boldsymbol{x}_i}$ and $L = L_{11} + \frac{L_{12} L_{21}}{\mu_{\boldsymbol{x}}}$. Under Assumptions 1, 2, and 3, we have*

$$\mathbb{E}[\tilde{R}_{P_n}(\boldsymbol{w}_{T+1}, p)] - \tilde{R}_{P_n}(\boldsymbol{w}^*, p) \le \frac{G^2 L}{T \mu_{\boldsymbol{w}}^2}, \tag{11}$$

and with probability at least $1 - \theta$,

$$\begin{aligned}&\tilde{R}_{P_n}(\boldsymbol{w}_{T+1}, p) - \tilde{R}_{P_n}(\boldsymbol{w}^*, p) \\ &\quad \le \frac{G^2 \log(\log(T/\theta))(64L + 16\mu_{\boldsymbol{w}}) + G^2 L}{T \mu_{\boldsymbol{w}}^2},\end{aligned} \tag{12}$$

for $0 < \theta < 1/e$, $T \ge 4$, with $\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w}} \tilde{R}_{P_n}(\boldsymbol{w}, p)$.

This theorem shows that Algorithm 1 is able to find the global minimum of the adversarial objective (7) both in expectation and in high probability. Specifically, the convergence rate of Algorithm 1 is $\mathcal{O}(1/\lceil T/K \rceil) = \mathcal{O}(K/T) = \tilde{\mathcal{O}}(1/T)$, since the number of inner loop steps K is $\mathcal{O}(\log(T d_0 r(p)^2))$, which increases with the feature dimension of input data d_0 and the size of perturbation r. The proof of Theorem 3 is in Appendix A.2.

The following Proposition 1 (proof is in Appendix A.2.2) shows that the model trained by Algorithm 1 has a small error on clean training samples, and satisfies the condition of input-robustness in Theorems 1 and 2.

Proposition 1. *If $\tilde{R}_{P_n}(\boldsymbol{w}) \le \epsilon$ for $\boldsymbol{w}$ and a constant ϵ, then $R_{P_n}(\boldsymbol{w}) \le \epsilon$, and $f(\boldsymbol{w}, \boldsymbol{x})$ is $(r(p), 2\epsilon, P_n, p)$-input-robust.*

According to Theorem 3 and Proposition 1, after T training steps in Algorithm 1, we can obtain a $(r(p), \tilde{\mathcal{O}}(1/T), P_n, p)$-input-robust model when $\tilde{R}_{P_n}(\boldsymbol{w}^*)$ is close to zero. Thus, combining Theorems 1 and 2, we get the following corollary which shows that the adversarially trained model generalizes on OOD data.

Corollary 1. *For* $p \in \{2, \infty\}$*, with the same notations as Theorem 1 and 3, if* $\tilde{R}_{P_n}(\boldsymbol{w}^*, p) \leq \epsilon_0$*, then with probability at least* $1 - \theta$*,*

$$\begin{aligned}\sup_{P \in B_{\mathsf{W}_2}(P_0, r/\epsilon_0)} & R_P(\boldsymbol{w}_{T+1}, 2) \leq (2M+3)\epsilon_0 \\ & + (2M+3)\left(\frac{G^2 \log(\log(2T/\theta))(64L + 16\mu_{\boldsymbol{w}}) + G^2 L}{T\mu_{\boldsymbol{w}}^2}\right) \\ & + M\sqrt{\frac{(2d_0)^{\frac{2\epsilon_0^2 D}{r^2}+1} \log 2 + 2\log(2/\theta)}{n}},\end{aligned}$$

and

$$\begin{aligned}\sup_{P \in B_{\mathsf{W}_\infty}(P_0, r)} & R_P(\boldsymbol{w}_{T+1}, \infty) \leq 3\epsilon_0 \\ & + \frac{G^2 \log(\log(2T/\theta))(192L + 48\mu_{\boldsymbol{w}}) + 3G^2 L}{T\mu_{\boldsymbol{w}}^2} \\ & + M\sqrt{\frac{2d_0^{\frac{2D}{r^2}+1} \log 2 + 2\log(2/\theta)}{n}},\end{aligned}$$

for any $0 \leq \theta \leq 1/e$ *and* $T \geq 4$*.*

This corollary is directly obtained by combining Theorem 1, 2, 3, and Proposition 1. It shows that the excess risk (i.e., the terms in the left-hand side of the above two inequalities) of the adversarially trained model on OOD data is upper bounded by $\tilde{\mathcal{O}}(1/\sqrt{n} + 1/T)$ after T steps. The dependence of the bounds on hyperparameters like input data dimension d_0, ℓ_1-diameter D of data support $\mathcal{X}$ are from the OOD generalization bounds (5), (6), and convergence rate (12).

4. Robust Pre-Trained Model has Better Initialization on Downstream Tasks

The paradigm of "first pre-train and then fine-tune" has been widely explored recently (Radford et al., 2021; Hendrycks et al., 2020b). In this section, we theoretically show that the input-robust pre-trained model provides an initialization that generalizes on downstream OOD data.

Assume the m i.i.d. samples $\{\boldsymbol{z}_i\}$ in the pre-training stage are from distribution Q_0. For a small constant ϵ_{pre} and given $r(2) = r/\epsilon_{\text{pre}}, r(\infty) = r$, the following Theorems 4 and 5 show that the pre-trained model with a small excess risk on OOD data in the pre-training stage also generalizes on downstream OOD data. The proofs are in Appendix B.1.

Theorem 4. *If* $\sup_{Q \in B_{\mathsf{W}_\infty}(Q_0, r(\infty))} R_Q(\boldsymbol{w}_{\text{pre}}) \leq \epsilon_{\text{pre}}$*, then*

$$\sup_{P \in B_{\mathsf{W}_\infty}(P_0, r(\infty))} R_P(\boldsymbol{w}_{\text{pre}}) \leq \epsilon_{\text{pre}} + 2M\mathsf{TV}(P_0, Q_0), \quad (13)$$

and with probability at least $1 - \theta$*,*

$$\tilde{R}_{P_n}(\boldsymbol{w}_{\text{pre}}, \infty) \leq \epsilon_{\text{pre}} + 2M\mathsf{TV}(P_0, Q_0) + M\sqrt{\frac{\log(1/\theta)}{2n}}. \quad (14)$$

Theorem 5. *If* $\sup_{Q \in B_{\mathsf{W}_2}(Q_0, r_0)} R_Q(\boldsymbol{w}_{\text{pre}}) \leq \epsilon_{\text{pre}}$ *with* $r_0 = \sqrt{2D^2\mathsf{TV}(P_0, Q_0) + r(2)^2}$*, then*

$$\sup_{P \in B_{\mathsf{W}_2}(P_0, r(2))} R_P(\boldsymbol{w}_{\text{pre}}) \leq \epsilon_{\text{pre}} + 2M\mathsf{TV}(P_0, Q_0). \quad (15)$$

Remark 4. *The self-supervised pre-training (e.g., masked language modeling in BERT (Devlin et al., 2019)) can also be included into the* $f(\boldsymbol{w}, \boldsymbol{x})$ *in Section 3.1, if we take label* $y \sim P_{y|\boldsymbol{x}}$ *as the distribution of the artificially constructed labels (e.g., masked tokens in BERT).*

When we implement fine-tuning on downstream tasks, the model is initialized by $\boldsymbol{w}_{\text{pre}}$. Combining the results in Theorems 1 and 2 (an input-robust model has small OOD generalization error) with Theorems 4 and 5, we conclude that the input-robust model has small excess risk on the OOD data in the pre-training stage, and thus generalizes on the OOD data of downstream tasks. Specifically, (13) and (15) show that the initial OOD excess risk in the fine-tuning stage $\sup_{P \in B_{\mathsf{W}_p}(P_0, r(p))} R_P(\boldsymbol{w}_{\text{pre}})$ is decided by terminal OOD excess risk in pre-training stage $\sup_{Q \in B_{\mathsf{W}_p}(Q_0, r(p))} R_Q(\boldsymbol{w}_{\text{pre}})$ and the total variation distance $\mathsf{TV}(P_0, Q_0)$. The intuition is that if $\boldsymbol{w}_{\text{pre}}$ generalizes well on distributions around Q_0, and P_0 is close to Q_0 under the total variation distance, then $\boldsymbol{w}_{\text{pre}}$ generalizes on downstream OOD data.

To satisfy the condition $\sup_{Q \in B_{\mathsf{W}_p}(Q_0, r(p))} R_Q(\boldsymbol{w}_{\text{pre}}) \leq \epsilon_{\text{pre}}$ in Theorems 4 and 5, we can use adversarial pre-training. Corollary 1 implies $\epsilon_{\text{pre}} = \mathcal{O}(1/\sqrt{m})$ by implementing sufficient adversarial pre-training. Thus, massive training samples m in the adversarial pre-training stage improves the OOD generalization on downstream tasks as $\epsilon_{\text{pre}} = \mathcal{O}(1/\sqrt{m})$ appears in the bounds (13) and (15).

Radford et al. (2021); Hendrycks et al. (2020b) empirically verify that the standardly pre-trained model also generalizes well on downstream OOD data. It was shown that sufficient standard training by gradient-based algorithm can also find the most input-robust model under some mild conditions (Soudry et al., 2018; Lyu & Li, 2019). Thus, $\sup_{Q \in B_{\mathsf{W}_\infty}(Q_0, r(p))} R_Q(\boldsymbol{w}_{\text{pre}}) \leq \epsilon_{\text{pre}}$ can hold even for standardly pre-trained model. However, the convergence to the most input-robust model of standard training is much slower compared with AT, e.g., for linear model (Soudry et al., 2018; Li et al., 2020). Hence, to efficiently learn an input-robust model in the pre-training stage, we suggest adversarial pre-training.

5. Experiments

5.1. Adversarial Training Improves OOD Generalization

In this section, we verify our conclusion in Section 3 that OOD generalization can be improved by AT (Corollary 1).

Table 1: Clean and corruption accuracy (%) of ResNet34 on `CIFAR10-C` and `ImageNet-C` using standard training and adversarial training under both ℓ_2-norm and ℓ_∞-norm.

Dataset	Method	Clean	Noise			Blur				Weather				Digital				Avg.
			Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	
`CIFAR10-C`	Std	94.82	34.75	40.43	25.45	59.85	48.95	67.58	63.85	73.31	62.87	**67.03**	**90.69**	**36.83**	76.00	42.89	75.84	57.75
	Adv-ℓ_2	94.93	70.39	74.24	45.17	72.77	71.34	73.51	80.26	83.28	81.36	51.08	89.37	19.49	83.39	79.78	**89.52**	71.00
	Adv-ℓ_∞	93.48	**80.18**	**80.80**	**62.73**	**77.71**	**77.10**	**75.46**	**82.47**	**83.45**	**82.32**	41.00	88.15	16.10	**83.82**	**85.98**	89.36	**73.78**
`ImageNet-C`	Std	74.01	18.97	18.39	12.98	6.32	9.76	11.49	9.37	8.78	12.98	6.21	33.74	4.31	18.29	23.91	29.08	14.97
	Adv-ℓ_2	73.66	**30.13**	**28.93**	**25.05**	**32.91**	25.61	**34.50**	32.84	**27.39**	**33.82**	**36.52**	**62.18**	**31.73**	42.91	47.86	51.55	**36.26**
	Adv-ℓ_∞	68.36	25.94	25.61	21.17	24.56	**32.81**	32.20	**34.57**	26.70	33.47	11.22	56.07	12.34	**47.67**	**57.32**	**59.10**	33.38

5.1.1. EXPERIMENTS ON IMAGE CLASSIFICATION

Data. We use the following benchmark datasets.

- `CIFAR10` (Krizhevsky & Hinton, 2009) has 50000 colorful images as training samples from 10 object classes. `CIFAR10-C` simulates OOD colorful images with 15 types of common visual corruptions, which serves as a benchmark to verify the OOD generalization of model trained on `CIFAR10`. Each type of corruption has five levels of severity, and each severity has 10000 validation samples. The 15 types of corruptions are divided into 4 groups: Noise, Blur, Weather and Digital.

- `ImageNet` (Deng et al., 2009) contains colorful images with over 1 million training samples from 1,000 categories. Similar to `CIFAR10-C`, `ImageNet-C` serves as a benchmark of OOD data with 15 types of corruptions. Each type of corruption has five levels of severity with 50000 validation samples in it. A visualization of `ImageNet-C` is in Figure 2 in Appendix.

Setup. The model used in this subsection is ResNet34 (He et al., 2016). To verify that adversarial training helps improve OOD performance, we conduct Algorithm 1 on `CIFAR10`, `ImageNet` and evaluate the model on `CIFAR10-C` and `ImageNet-C`, respectively. The number of inner loop steps K is 8 for `CIFAR10`, and 3 for `ImageNet`. The models are trained by SGD with momentum. The number of training epochs is 200 for `CIFAR10`, and 100 for `ImageNet`. The learning rate starts from 0.1 and decays by a factor 0.2 at epochs 60, 120, 160 (resp. 30, 60, 90) for `CIFAR10` (resp. `ImageNet`). Detailed hyperparameters are in Appendix C.

We compare adversarial training under ℓ_2- and ℓ_∞-norm (respectively abbreviated as "Adv-ℓ_2" and "Adv-ℓ_∞") against standard training (abbreviated as "Std"). For Adv-ℓ_∞, we replace $\nabla_{\boldsymbol{x}} f(\boldsymbol{w}_t, \boldsymbol{x}_{i_t} + \boldsymbol{\delta}_k)$ in Line 4 of Algorithm 1 with the sign of it as in (Madry et al., 2018), in order to find stronger adversarial perturbation (Goodfellow et al., 2015).

Main Results. In Table 1, for each type of corruption, we report the test accuracy on `CIFAR10-C` under the strongest corruption severity level 5[2]. For `ImageNet-C`, we report the average test accuracy of five severity levels as in (Hendrycks & Dietterich, 2018). We also report the test accuracy on `CIFAR10` and `ImageNet` in the column of "Clean" for comparison.

As can be seen, Adv-ℓ_2 and Adv-ℓ_∞ improve the average accuracy on OOD data, especially under corruption types Noise and Blur. This supports our finding in Section 3 that AT makes the model generalize on OOD data. Though AT improves the OOD generalization on all corruption types for `ImageNet-C`, it degenerates the performance for data corrupted under types Fog, Bright and Contrast in `CIFAR10-C`. We speculate this is because these three corruptions intrinsically rescale the adversarial perturbation, and refer readers to Appendix D.1 for a detailed discussion.

Ablation Study. We study the effect of perturbation size r and the number of training samples n for adversarial training in bounds (5) and (6). Due to the space limit, we put the implementation details and results in Appendix D.

The results for the effect of perturbation size r are in Figures 3-4 in Appendix D.1. As can be seen, the accuracy on OOD data `CIFAR10-C` first increases and then decreases with an increasing r. This is because the upper bounds of excess risk in (5) and (6) are decided by both the clean accuracy and input-robustness. However, an increasing perturbation size r improves the input-robustness, but harms the clean accuracy (Raghunathan et al., 2019). Specifically, when the perturbation size r is small, the clean accuracy is relatively stable and the robustness dominates. Thus the overall OOD performance increases as r increases. However, when r is relatively large, a larger r leads to worse clean accuracy though better robustness, and can lead to worse overall OOD performance. Thus, to achieve the optimal performance on OOD data, we should properly choose the perturbation size r rather than continually increasing it.

The results for the effect of the number of training samples n are in Figures 5-6 in Appendix D.2. The accuracy on OOD data increases with the number of training samples, which is consistent with our findings in Theorems 1 and 2.

---

[2] Lighter severity levels exhibit similar trends but with smaller performance gaps between adversarial and standard training

Table 2: Performance of BERT base model on NLU tasks using standard training and adversarial training under both ℓ_2-norm and ℓ_∞-norm.

Dataset	Train	Test	Std	Adv-ℓ_2	Adv-ℓ_∞
`STS-B`	Images	Images	98.38	97.81	96.39
		MSRvid	89.52(-8.86)	**90.61**(-7.20)	90.09(-6.30)
	MSRvid	MSRvid	98.55	97.45	96.65
		Images	**84.12**(-14.43)	83.63(-13.82)	83.11(-13.54)
	Headlines	Headlines	97.59	96.73	95.75
		MSRpar	62.07(-35.52)	64.48(-32.25)	**67.67**(-28.08)
	MSRpar	MSRpar	97.55	97.33	97.55
		Headlines	75.58(-21.97)	75.27(-22.06)	**76.12**(-21.43)
`SST-2; IMDb`	SST-2	SST-2	93.57	93.57	93.92
		IMDb	90.06(-3.51)	**91.50**(-2.07)	91.32(-2.60)
	IMDb	IMDb	94.36	94.88	94.68
		SST-2	87.00(-7.36)	**88.53**(-6.35)	88.07(-6.61)
`MNLI`	Telephone	Telephone	83.01	83.16	82.90
		Letters	82.45(-0.56)	83.76(+0.60)	**84.07**(+1.17)
		Face-to-face	81.56(-1.45)	**83.59**(+0.43)	**83.59**(+0.69)

5.1.2. EXPERIMENTS ON NATURAL LANGUAGE UNDERSTANDING

Data. As in (Hendrycks et al., 2020b), we use three pairs of datasets as the original and OOD datasets for NLU tasks.

- `SST-2` (Socher et al., 2013) and `IMDb` (Maas et al., 2011) are sentiment analysis datasets, with pithy expert and full-length lay movie reviews, respectively. As in (Hendrycks et al., 2020b), we train on one dataset and evaluate on the other. Then we report the accuracy of a review's binary sentiment predicted by the model.
- `STS-B` consists of texts from different genres and sources. It requires the model to predict the textual similarity between pairs of sentences (Cer et al., 2017). As in (Hendrycks et al., 2020b), we use four sources from two genres: MSRpar(news), Headlines (news); MSRvid(captions), Images(captions). The evaluation metric is Pearson's correlation coefficient.
- `MNLI` is a textual entailment dataset which contains sentence pairs from different genres of text (Williams et al., 2018). We select training samples from two genres of transcribed text (Telephone and Face-to-Face) and the other of written text (Letters) as in (Hendrycks et al., 2020b), and report the classification accuracy.

Setup. For a pre-trained language model e.g., BERT, each input token is encoded as a one-hot vector and then mapped into a continuous embedding space. Instead of adding perturbations to the one-hot vectors, we construct adversarial samples in the word embedding space as in (Zhu et al., 2019).

The backbone model is the base version of BERT (Devlin et al., 2019) which has been widely used in the NLP community. We conduct AT in the fine-tuning stage to see its effectiveness on OOD generalization. The models are trained by AdamW (Loshchilov & Hutter, 2018) for 10 epochs. Detailed hyperparameters are in Appendix C. As in Section 5.1.1, we compare Adv-ℓ_2 and Adv-ℓ_∞ with Std.

Main Results. In Table 2, we report the results on in-distribution data and OOD data, and the gap between them (in the brackets) as in (Hendrycks et al., 2020b). The gaps in brackets are used to alleviate the interference by the general benefits from AT itself, since it was shown in (Zhu et al., 2019) that AT can improve the generalization ability of model on in-distribution textual data.

As can be seen, adversarially trained models perform similarly or even better than standardly trained models on in-distribution data, while significantly better on OOD data especially for `MNLI`. The smaller gaps between in-distribution and OOD data support our finding that AT can be used to improve OOD generalization.

5.2. Robust Pre-Trained Model Improves OOD Generalization

Previously in Section 4, we theoretically show that an input-robust pre-trained model gives a better initialization for fine-tuning on downstream task, in terms of OOD generalization. In this section, we empirically show that this better initialization also leads to better OOD generalization after finetuning on image classification tasks.

Setup. Following (Salman et al., 2020a), we pre-train the model on `ImageNet` and then fine-tune it on `CIFAR10`. To get an input-robust model in the pre-training stage, we consider adversarially pre-train the model. We compare adversarial pre-training (Adv-ℓ_2 and Adv-ℓ_∞) against standard pre-training and no pre-training as in Section 5.1.1. In the fine-tuning stage, the data from `CIFAR10` are resized to 224×224 as in (Salman et al., 2020a). We also compare stadard fine-tuning and adversarial fine-tuning under both ℓ_2- and ℓ_∞-norm. After fine-tuning, we verify the OOD generalization on `CIFAR10-C`. The other settings are the same as Section 5.1.1.

Main Results. The results are shown in Table 3. As can be seen, for all fine-tuning methods, adversarially pre-trained models consistently achieve better performance on OOD data than standardly pre-trained models or models without pre-training. Thus, the initialization from the adversarially pre-trained input-robust model leads to better OOD generalization on downstream tasks after fine-tuning. In addition, standard pre-training slightly improves the OOD generalization compared with no pre-training when we conduct Adv-ℓ_∞ fine-tuning or standard fine-tuning. We also observe that for all four kinds of pre-training, adversarial fine-tuning under ℓ_∞-norm has better performance than ℓ_2-norm. This agrees with the observations in Section 5.1.1. Note that the results of models without pre-training are different from

Table 3: Clean and corruption accuracy (%) of ResNet34 on `CIFAR10-C` with no pre-training, standard pre-training, and adversarial pre-training under ℓ_2-norm and ℓ_∞-norm.

Fine-Tuning	Pre-Training	Clean	Noise			Blur				Weather				Digital				Avg.
			Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	
Std	No	95.21	40.55	40.64	19.91	83.21	67.77	77.86	90.31	80.71	77.91	67.27	**90.88**	48.14	80.80	81.99	80.84	68.59
	Std	94.65	41.25	42.91	22.58	85.19	71.03	78.49	**90.82**	82.78	**80.04**	67.66	89.97	45.70	**83.89**	82.03	**80.99**	69.69
	Adv-ℓ_2	95.06	**45.10**	**50.58**	27.57	87.27	**72.95**	79.08	90.57	**83.29**	77.25	65.41	90.15	**50.41**	82.81	78.01	78.95	70.63
	Adv-ℓ_∞	94.30	40.94	46.42	**29.39**	**87.60**	70.79	**81.44**	90.69	82.77	79.28	**68.84**	89.19	45.29	83.59	**83.13**	80.86	**70.68**
Adv-l_2	No	94.43	56.82	60.58	29.34	85.44	71.67	81.80	90.08	83.68	80.37	61.68	89.96	34.76	83.76	85.16	83.24	71.89
	Std	94.09	57.64	60.96	26.35	86.78	**73.52**	82.16	90.46	82.12	80.64	62.58	88.98	34.68	84.29	83.42	83.42	71.87
	Adv-ℓ_2	94.45	**58.98**	**62.99**	**35.08**	87.07	72.29	81.66	91.07	83.53	81.38	62.82	89.52	**39.53**	84.35	**86.60**	88.55	73.69
	Adv-ℓ_∞	95.25	58.64	62.18	29.86	**88.15**	73.00	**82.95**	**91.98**	**84.76**	**83.86**	**64.76**	**91.00**	37.35	**84.65**	86.57	**88.59**	**73.89**
Adv-ℓ_∞	No	92.46	80.91	81.69	52.00	79.58	80.94	77.42	80.21	80.57	79.35	35.41	83.15	18.06	83.51	87.79	87.44	72.54
	Std	92.05	80.21	81.06	**63.02**	77.94	77.80	75.60	80.04	**83.77**	81.22	41.57	**89.94**	**19.04**	82.39	85.49	**88.76**	73.86
	Adv-ℓ_2	92.55	**81.96**	**82.86**	58.95	**80.51**	**82.66**	**78.21**	**86.56**	81.49	81.10	42.07	89.76	18.56	**84.58**	**88.53**	88.05	**75.06**
	Adv-ℓ_∞	92.28	81.74	82.37	56.96	80.34	81.90	77.94	85.76	81.48	**81.70**	**42.99**	89.00	18.45	84.50	88.07	87.50	74.71

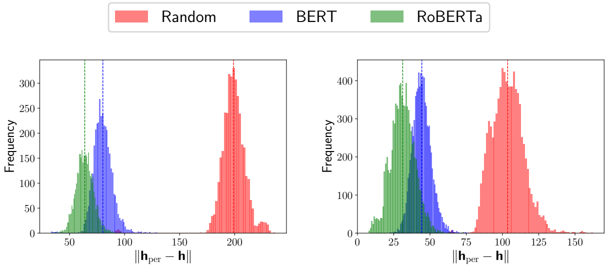
those in Table 1 due to the resized input data.

5.3. Discussion

It is shown in (Hendrycks et al., 2020b) that the language model BERT (Devlin et al., 2019) pre-trained on large corpus generalizes well on downstream OOD data, and RoBERTa (Liu et al., 2019) pre-trained with more training data and updates generalizes even better than BERT. We speculate this is because (i) sufficient pre-training obtains an input-robust model as discussed in Section 4, and this better-initialization leads to better OOD generalization after finetuning as observed in Section 5.2; and (ii) the objective of masked language modeling predicts the masked (perturbed) input tokens and enables a certain amount of input-robustness.

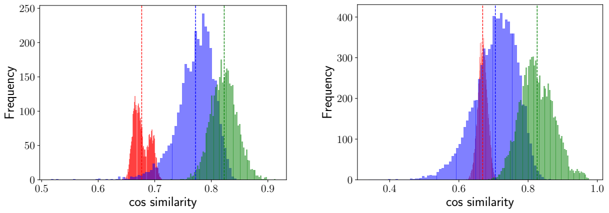
In this section, we empirically show that the model initialized by BERT has higher input-robustness than a randomly initialized model. Besides, compared with BERT, RoBERTa is pre-trained with more training samples and updating steps and the model initialized by it is more robust to input perturbations.

Setup. We compare the input-robustness of the base versions of pre-trained language model BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), against a randomly initialized model whose parameters are independently sampled from $\mathcal{N}(0, 0.02^2)$ (Wolf et al., 2020). The three models have exactly the same structure. Compared with BERT, RoBERTa is pre-trained on a larger corpus for more updating steps. Experiments are performed on `MRPC` and `CoLA` datasets from the GLUE benchmark (Wang et al., 2018), with 3.7k and 8.5k training samples, respectively. Similar as Section 5.1.2, we add adversarial perturbations in the embedding space. We use 3 steps of ℓ_∞-norm attack to construct perturbation. The perturbation size is 0.001 and the perturbation step size 0.0005. Since the the last classification layer of BERT or RoBERTa is randomly initialized during downstream task fine-tuning, we study the difference in the hidden states of the last Transformer layer before the classification layer. Denote $\mathbf{h}, \mathbf{h}_{\text{per}} \in \mathbb{R}^{128\times 768}$ as the hidden states from the original input and the adversarially perturbed input, respectively. We use the ℓ_2-norm $\|\mathbf{h}_{\text{per}} - \mathbf{h}\|$ and the cosine similarity $\langle \mathbf{h}, \mathbf{h}_{\text{per}}\rangle / (\|\mathbf{h}\|\|\mathbf{h}_{\text{per}}\|)$ to measure the difference. The cosine similarity is used to alleviate the potential interference caused by the scale of $\mathbf{h}$ over different pre-trained models. The results are in Figure 1.



(a) `MRPC`.

(b) `CoLA`.



(c) `MRPC`.

(d) `CoLA`.

Figure 1: Difference of hidden states in the last Transformer layer between the original input and adversarially perturbed input, measured by ℓ_2-norm and cosine similarity. The models compared are randomly initialized model, BERT, and RoBERTa. The datasets used are `MRPC` and `CoLA` from the GLUE benchamrk. The dashed lines in the upper and bottom figures are respectively the mean of $\|\mathbf{h}_{\text{per}} - \mathbf{h}\|$ and $\langle \mathbf{h}, \mathbf{h}_{\text{per}}\rangle / (\|\mathbf{h}\|\|\mathbf{h}_{\text{per}}\|)$ from all samples in a dataset.

Main Results. The histograms of $\|\mathbf{h}_{\text{per}} - \mathbf{h}\|$ and $\langle \mathbf{h}, \mathbf{h}_{\text{per}}\rangle / (\|\mathbf{h}\|\|\mathbf{h}_{\text{per}}\|)$ from all training samples in MRPC and CoLA are shown in Figure 1a, 1b and Figure 1c, 1d, respectively. We can observe that (i) BERT is more robust than the randomly initialized model, indicating that the masked language modeling objective and sufficient pretraining improves input-robustness, and leads to better ood performance after fine-tuning; (ii) RoBERTa is more input-robust compared with BERT, which implies that that more training samples and updating steps in the pre-training stage improve the input-robustness. Combining with that a more input-robust pre-trained model also leads to better OOD generalization on downstream tasks empirically (Section 4), the above observations (i) and (ii) may also explain the finding in (Hendrycks et al., 2020b) that BERT generalizes worse on downstream OOD data than RoBERTa, but much better than the model without pretraining.

6. Conclusion

In this paper, we explore the relationship between the robustness and OOD generalization of a model. We theoretically show that the input-robust model can generalize well on OOD data under the definition of OOD generalization via Wasserstein distance. Thus, for a model trained from scratch, we suggest using adversarial training to improve the input-robustness of the model which results in better OOD generalization. Under mild conditions, we show that the excess risk on OOD data of an adversarially trained model is upper bounded by $\tilde{\mathcal{O}}(1/\sqrt{n} + 1/T)$. For the framework of first pre-training and then fine-tuning, we show that a pre-trained input-robust model provides a theoretically good initialization which empirically improves OOD generalization after fine-tuning. Various experiments on CV and NLP verify our theoretical findings.

References

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, 2019.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. Preprint, 2020.

Bubeck, S. Convex optimization: Algorithms and complexity. Preprint arXiv:1405.4980, 2014.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. Semeval-2017 task 1: semantic textual similarity multilingual and crosslingual focused evaluation. In *International Workshop on Semantic Evaluation*, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, 2019.

Duchi, J. Lecture notes for statistics 311/electrical engineering 377, 2016. `http://web.stanford.edu/class/stats311/lecture-notes.pdf`.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. Preprint arXiv:2006.16241, 2020a.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. In *Annual Conference of the Association for Computational Linguistics*, 2020b.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial logit pairing. Preprint arXiv:1803.06373, 2018.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

Lee, J. and Raginsky, M. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems*, 2018.

Li, Y., Fang, E. X., Xu, H., and Zhao, T. Inductive bias of gradient descent based adversarial training on separable data. In *International Conference on Learning Representations*, 2020.

Liu, C., Zhu, L., and Belkin, M. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. Preprint arXiv:2003.00307, 2020.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. Preprint arXiv:1907.11692, 2019.

Lohn, A. J. Estimating the brittleness of ai: safety integrity levels and the need for testing out-of-distribution performance. Preprint arXiv:2009.00802, 2020.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2019.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Annual Conference of the Association for Computational Linguistics*, 2011.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Namkoong, H. *Reliable machine learning via distributional robustness*. PhD thesis, Stanford University, 2019.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision, 2021. `https://openai.com/blog/clip/`.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. Adversarial training can hurt generalization. Preprint arXiv:1906.06032, 2019.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *International Coference on Machine Learning*, 2012.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019.

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. Do adversarially robust imagenet models transfer better? Preprint arXiv:2007.08489, 2020a.

Salman, H., Ilyas, A., Engstrom, L., Vemprala, S., Madry, A., and Kapoor, A. Unadversarial examples: designing objects for robust vision. Preprint arXiv:2012.12235, 2020b.

Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., and Bethge, M. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, 2020.

Shapiro, A. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing*, 2013.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. Preprint arXiv:1312.6199, 2013.

Tu, L., Lalwani, G., Gella, S., and He, H. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 2020.

Utrera, F., Kravitz, E., Erichson, N. B., Khanna, R., and Mahoney, M. W. Adversarially-trained deep nets transfer better. In *International Conference on Learning Representations*, 2021.

van der Vaart, A. W. and Wellner, J. A. *Weak convergence and empirical processes*. Springer series in statistics. Springer, 2000.

Vershynin, R. *High-dimensional probability: an introduction with applications in data science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, 2018.

Wainwright, M. J. *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2018.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., et al. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing*, 2020.

Xie, B., Liang, Y., and Song, L. Diversity leads to generalization in neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A. L., and Le, Q. V. Adversarial examples improve image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Xu, H. and Mannor, S. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, 2019.

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*, 2019.