Initialization Matters for Adversarial Transfer Learning

Andong Hua[1] **Jindong Gu**[2] **Zhiyu Xue**[1] **Nicholas Carlini**[3] **Eric Wong**[4] **Yao Qin**[1,3]
[1]University of California, Santa Barbara [2]University of Oxford
[3]Google [4]University of Pennsylvania

Abstract

With the prevalence of the Pretraining-Finetuning paradigm in transfer learning, the robustness of downstream tasks has become a critical concern. In this work, we delve into adversarial robustness in transfer learning and reveal the critical role of initialization, including both the pretrained model and the linear head. First, we discover the necessity of an adversarially robust pretrained model. Specifically, we reveal that with a standard pretrained model, Parameter-Efficient Finetuning (PEFT) methods either fail to be adversarially robust or continue to exhibit significantly degraded adversarial robustness on downstream tasks, even with adversarial training during finetuning. Leveraging a robust pretrained model, surprisingly, we observe that a simple linear probing can outperform full finetuning and other PEFT methods with random initialization on certain datasets. We further identify that linear probing excels in preserving robustness from the robust pretraining. Based on this, we propose Robust Linear Initialization (RoLI) for adversarial finetuning, which initializes the linear head with the weights obtained by adversarial linear probing to maximally inherit the robustness from pretraining. Across five different image classification datasets, we demonstrate the effectiveness of RoLI and achieve new state-of-the-art results. Our code is available at https://github.com/DongXzz/RoLI.

1. Introduction

With the advancement of large-scale deep learning models, the Pretraining-Finetuning paradigm takes a more dominant role compared to training from scratch across various tasks [2], including computer vision [12, 15, 22, 28], natural language processing [20, 36, 41], and speech recognition [35]. Under the paradigm of Pretraining-Finetuing, advanced parameter-efficient finetuning (PEFT) methods [4, 16–18, 27, 48] have emerged. Compared to full finetuning, PEFT methods either introduce small modules into a fixed pretrained model or optimize only part of the pretrained model, demonstrating exceptional performance while keeping storage usage low.

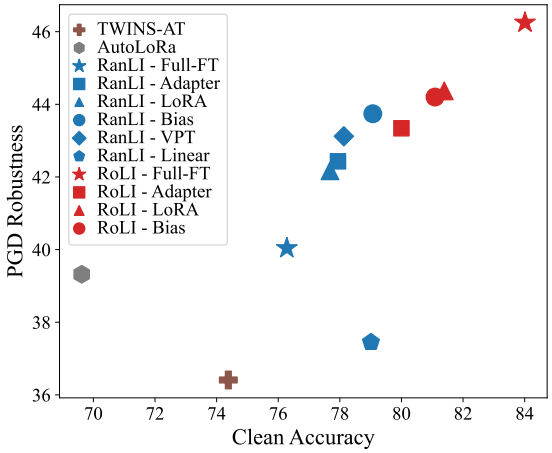

Figure 1. **Robust Linear Initialization (RoLI), significantly improves adversarial robustness.** RoLI, denoted in red, achieves an average 3.88% increase in clean accuracy and a 2.44% increase in robust accuracy compared to Random Linear Initialization (RanLI) of the linear head during adversarial finetuning, averaged across five downstream datasets. Our best-performing RoLI - Full-FT, which represents adversarial full finetuning with robust linear initialization, achieves a new state-of-the-art performance. We include six popular finetuning methods with Swin Transformer [28] and two existing state-of-the-art techniques for adversarial transfer learning: TWINS-AT [29] and AutoLoRA [43].

While the success of transfer learning is evident, the robustness of downstream models has become a critical concern in real-world applications. For example, as suggested by previous studies, full finetuning will distort the pretrained features, leading to compromised robustness in terms of image corruptions and out-of-distribution (OOD) performance [26, 44]. In addition, adversarial robustness poses another significant challenge to real-world deployment, as a non-robust model demonstrates nearly zero accuracy under adversarial attacks [40]. Many approaches have been proposed to enhance adversarial robustness in transfer learning, including improving the robustness of pretraining through self-supervised techniques [7, 19] and preserving robustness from pretrained models during finetuning [29, 43]. However, most previous works [29, 43] apply

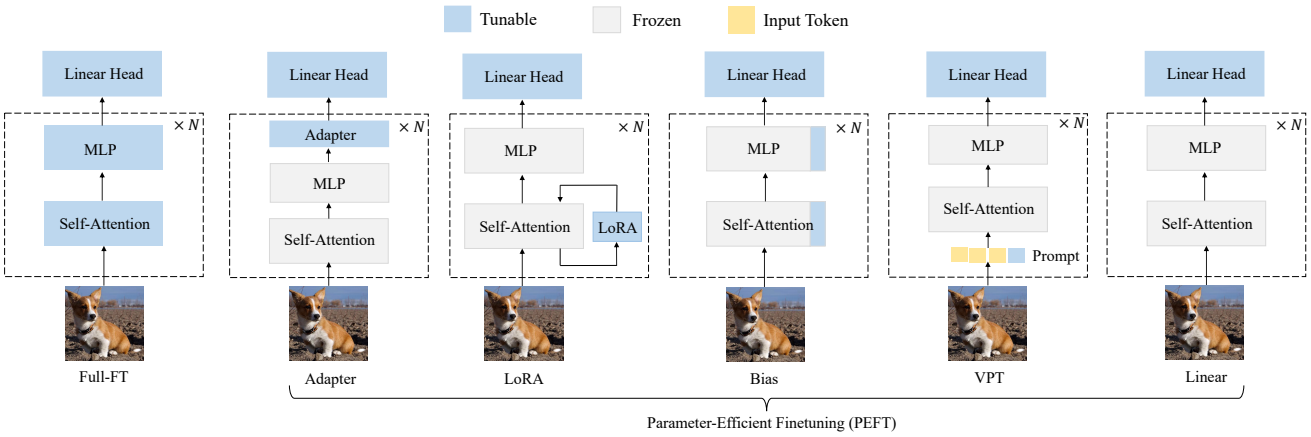

Figure 2. Illustration of six different finetuning techniques, arranged in descending order according to the number of tunable parameters.

full finetuning for adversarial transfer learning, overlooking the significance of different finetuning methods.

In this work, we comprehensively study six popular finetuning methods in adversarial transfer learning, as shown in Fig. 2. We discover that initialization, which includes 1) the backbone initialization (i.e., a pretrained model), and 2) the initialization of the linear head that adapts features to the target domain, plays a critical role in adversarial transfer learning. We start by comparing the performance of adversarial finetuning when initialized with either a robust or a standard pretrained model. Surprisingly, all PEFT methods fail or exhibit significantly inferior performance when initialized with a standard pretrained model compared to being initialized with a robust pretrained model. Furthermore, we observe that a large-scale pretrained model, such as CLIP [34], cannot alleviate this phenomenon, highlighting the critical role of robust pretraining in adversarial transfer learning.

Given a robust pretrained model, surprisingly, we discover that adversarial linear probing outperforms other methods on certain datasets, e.g., Caltech256 [13] and Stanford Dogs [21]. We further demonstrate that this is mainly because linear probing excels in inheriting robustness from pretraining. Moreover, we identify the ability to adapt features to the specific target domain as a key factor influencing linear probing's success, supported by a strong correlation between transferred accuracy and robustness. Based on these insights, we introduce Robust Linear Initialization (RoLI), which initializes the model's linear head with the weights obtained from adversarial linear probing. With RoLI followed by adversarial finetuning methods, we maximize the inherited robustness along with a strong feature adaptation ability, resulting in improved performance. In summary, our contributions are as follows:

- We comprehensively study six popular finetuning techniques for adversarial transfer learning. We discover that PEFT methods fail or exhibit significantly inferior performance when initialized with a standard pretrained model, even with adversarial finetuning on downstream data.
- We demonstrate that adversarial linear probing excels in preserving robustness from a robust pretrained model. Building upon this insight, we propose Robust Linear Initialization (RoLI) for adversarial finetuning to maximally inherit robustness from pretraining and effectively adapt features through adversarial finetuning.
- We demonstrate the effectiveness of RoLI across five downstream datasets. On average, RoLI improves the clean and robust accuracy by 3.88% and 2.44% compared with random initialization. This establishes a new state-of-the-art benchmark for adversarial transfer learning.

2. Related Works

Finetuning in Transfer Learning. Transfer learning aims to finetune a pretrained model on downstream tasks to gain better performance [23, 45, 47]. Linear probing and full finetuning are often applied in the finetuning process. Given the increasing size of pretrained models, various Parameter Efficient Finetuning (PEFT) methods have been proposed. PEFT methods can effectively reduce the finetuning cost and alleviate overfitting since only a small number of parameters are updated [16, 17]. Specifically, Bias [3, 46] only updates the bias term, while Scaling & Shifting [27] performs the linear transformation to adapt the features. Different from both of them, Adapter [16, 32, 33], LoRA [17], and VPT [18] introduce extra learnable modules into the pretrained models in the finetuning stage. Although there are various finetuning methods proposed to improve the accuracy, it is still unclear what matters to the robustness on the downstream tasks. [25] points out that feature distortion in finetuning hurts the out-of-distribution robustness on downstream tasks, while we discover initialization plays a critical role in adversarial robustness on downstream tasks.

Adversarial Robustness in Transfer Learning. The research on adversarial robustness in transfer learning has been studied from two perspectives, namely, robust pretraining and robust finetuning. Concretely, robust pertaining aims to achieve adversarial robustness in the pertaining stage and finetune pretrained models on the downstream tasks without adversarial training [5, 7, 19, 39]. Besides, [38] demonstrates that the adversarially pertaining also benefits standard performance on downstream tasks. In contrast, robust finetuning [29, 43] focuses on preserving the robustness of the pretrained model during the finetuning stage. For example, TWINS [29] incorporates a dual batch normalization in the model to keep the statistics of pretrained and finetuned datasets separately. And AutoLoRA [43] introduces a low-rank (LoRA) branch to disentangle clean and adversarial objectives. Existing works design various strategies to preserve the robustness from the pretrained model, but they simply apply full finetuning on downstream tasks. In this work, we study six different finetuning methods and introduce Robust Linear Initialization (RoLI) to enhance the adversarial robustness for different finetuning methods.

3. Background

In this section, we provide an overview of existing finetuning techniques, including full finetuning as well as five different PEFT methods, as illustrated in Fig. 2. Additionally, we introduce adversarial finetuning which integrates adversarial training during the finetuning stage.

3.1. Existing Finetuning Methods

In transfer learning, finetuning adapts the features from the pretrained domain to the target domain. In this section, we introduce six finetuning methods in descending order according to the number of tunable parameters.

Full Finetune (Full-FT): We initialize the model from a pretrained model and finetune all its parameters on the downstream tasks.

Adapter [16, 32, 33]: Adapter introduces a module after the MLP block in every layer. The adapter module consists of a down-sampling layer and an up-sampling layer, with a non-linear activation. During finetuning, we will update the adapter modules as well as the linear head.

Low-Rank Adaptation (LoRA) [17]: LoRA proposes to learn a residual from pretrained weight represented by two low-rank metrics. Specifically, we apply the LoRA branch on the query and value projection in the self-attention block. During finetuning, the pretrained model remains frozen while updates are applied to the LoRA branch and head.

Bias [3, 46]: We update the bias term and keep other parameters unchanged. Additionally, We also update the classification head (including weights and bias).

Visual Prompt Tuning (VPT) [18]: VPT appends additional learnable tokens (embeddings), called prompts, into the input space of each attention layer in vision transformers. Specifically, we use the structure of VPT-Deep, where every layer introduces a fixed number of trainable prompts independently. During the finetuning stage, we will tune both the prompts and the linear head on downstream tasks while freezing the entire pretrained backbone. We refer readers to [18] for more details.

Linear Probing (Linear): We only finetune the classification head on the downstream tasks.

3.2. Adversarial Finetuning

Adversarial training corresponds to a min-max optimization problem [31]. In the context of adversarial finetuning, we can formulate the optimization problem as follows:

$$\min_{\hat{\theta}} \mathbb{E}_{(x,y)\sim D}\left[\max_{\|\delta\|_\infty \leq \varepsilon} \mathcal{L}(x+\delta, y; \theta \cup \hat{\theta})\right] \tag{1}$$

where θ and $\hat{\theta}$ represent the frozen parameters and tunable parameters in model parameters respectively. For example, in adversarial full finetuning, $\hat{\theta}$ stands for all the model parameters as we update all of them during the finetuning stage. In contrast, in adversarial VPT, $\hat{\theta}$ denotes the extra prompts and linear head while θ represents the frozen parameters in the backbone architecture. In addition, δ represents adversarial perturbation, whose $\|\cdot\|_\infty$ is bounded by ε. During adversarial finetuning, we use the pretrained model together with additional modules introduced by PEFT methods to generate adversarial examples, while only updating the tunable parameters $\hat{\theta}$ for each method. If not otherwise specified, we use PGD-7 with $\varepsilon = 8/255$ and step size $\alpha = 2/255$ to generate adversarial attacks during adversarial finetuning.

4. Robust Pretrained Model Matters

Previous research [29, 39] suggests that although standard pretraining results in performance inferior to that of adversarially robust pretraining, it can still yield an adversarially robust model with adversarial finetuning. In this section, our goal is to investigate the significance of pretraining in PEFT methods. Specifically, we seek to address the question: *Is an adversarially robust model necessary for PEFT methods?*

To answer this question, we initialize model parameters using two different pretrained models. One is a standard pretrained model on ImageNet-1k [11] while the other is adversarially pretrained with PGD [31] attacks. These pretrained models are off-the-shelf and their specifics are outlined in the supplementary material. Following initialization, we perform adversarial finetuning on downstream tasks. As shown in Fig. 3, we evaluate six adversarial

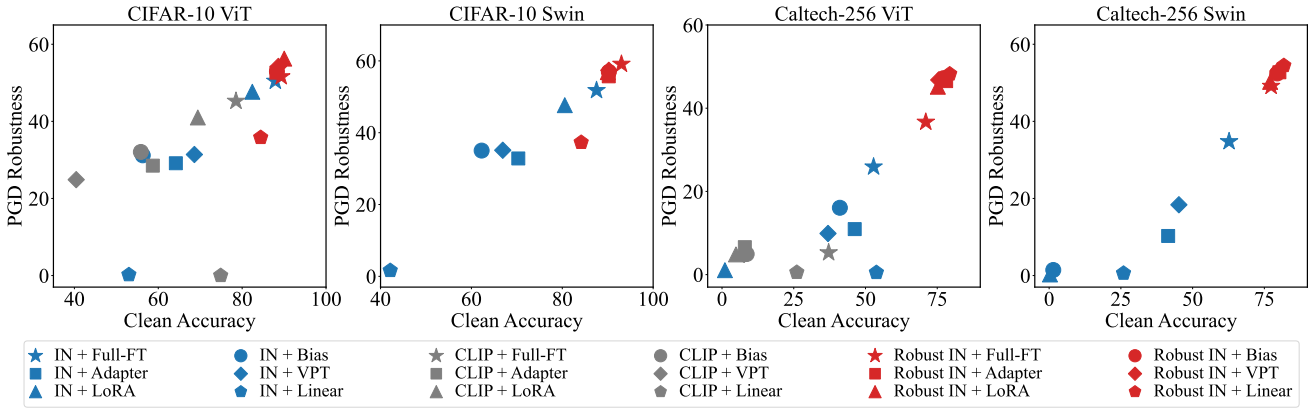

Figure 3. **PEFT methods fail or exhibit significantly degraded performance with a standard pretrained model (in blue and gray).** Models finetuned from a robust pretrained model (in red) exhibit high accuracy and robustness, as they are consistently positioned in the top-right corner. Among six finetuning techniques, Full-FT outperforms others when starting with a standard pretrained model; however, it falls short of other methods when starting with a robust pretrained model. The specific numerical results are provided in the supplementary.

finetuning techniques across two model architectures (ViT-B [12] and Swin-B [28]) on the CIFAR10 [24] and Caltech256 [13]. The adversarial robustness is measured with PGD-10 bounded with $\varepsilon = 8/255$.

First, we find that initializing the model with an adversarially robust pretrained model can significantly improve adversarial robustness on downstream tasks. This consistently holds across all the experimental settings we have evaluated, e.g., 6 adversarial finetuning techniques across 2 model architectures on 2 different datasets. More surprisingly, we discover that a robust pretrained model is especially important for PEFT methods, as they struggle to achieve satisfactory adversarial robustness when initialized with a standard pretrained model. For example, all the PEFT methods fail to achieve adversarial robustness higher than 20% on the Caltech256 dataset. In addition, adversarial linear probing can only achieve $< 5\%$ adversarial robustness on both datasets when initiated with a standard pretrained model. This strongly suggests that adversarial PEFT can not sufficiently infuse adversarial robustness in the finetuning stage if the model starts with a non-robust pretrained model. Finally, we observe that Full-FT consistently outperforms other finetuning methods with a non-robust pretraining, indicating the preference for using full finetuning in the absence of robust pretraining.

Does Pretraining on a Larger Dataset Help Adversarial Finetuning? Due to the high computational cost of adversarial pretraining, we aim to investigate whether a model pretrained on a larger dataset could substitute for the necessity of an adversarially robust pretrained model. To this end, we initialize the model with an off-the-shelf large-scale pretrained model, CLIP [34], and perform the same adversarial finetuning techniques on downstream tasks. Fig. 3 demonstrates that a standard pretrained model on a larger dataset, like CLIP, does not yield higher robustness compared to the non-robust ImageNet-1k pretrained model. It notably lags far behind the robustness achieved by a smaller-scale robust pretrained model, such as one trained on ImageNet-1k. This observation implies that opting for a smaller yet robust pretrained model proves to be a more effective strategy for attaining adversarial robustness in downstream tasks compared to using a larger, non-robust pretrained model.

Methods	ViT		Swin	
	Clean	PGD	Clean	PGD
RanLI - Full-FT	80.17	44.15	85.19	54.13
RanLI - Adapter	83.14	49.57	85.24	54.22
RanLI - LoRa	82.63	50.72	83.52	53.57
RanLI - Bias	82.69	50.07	84.88	54.69
RanLI - VPT	82.09	50.52	84.99	55.23
RanLI - Linear	81.89	42.02	82.98	45.89

Table 1. **PEFT methods, excluding Linear, demonstrate strong performance on average across CIFAR10 and Caltech256 datasets.** With adversarial finetuning, PEFT methods outperform Full-FT and Linear with RanLI. Full results are available in **??**

5. Initialization Matters For Finetuning

In the previous section, we understand the necessity of initializing a model with an adversarially robust pretrained model. We now ask: *given a robust pretrained model, what matters for adversarial finetuning?*

5.1. Random Linear Initialization

To answer this question, we first evaluate the adversarial robustness of six adversarial finetuning techniques. To adapt the robust pretrained model to the downstream tasks, we initialize the linear head with Random Linear Initialization (RanLI) following [39]. We perform adversarial finetuning

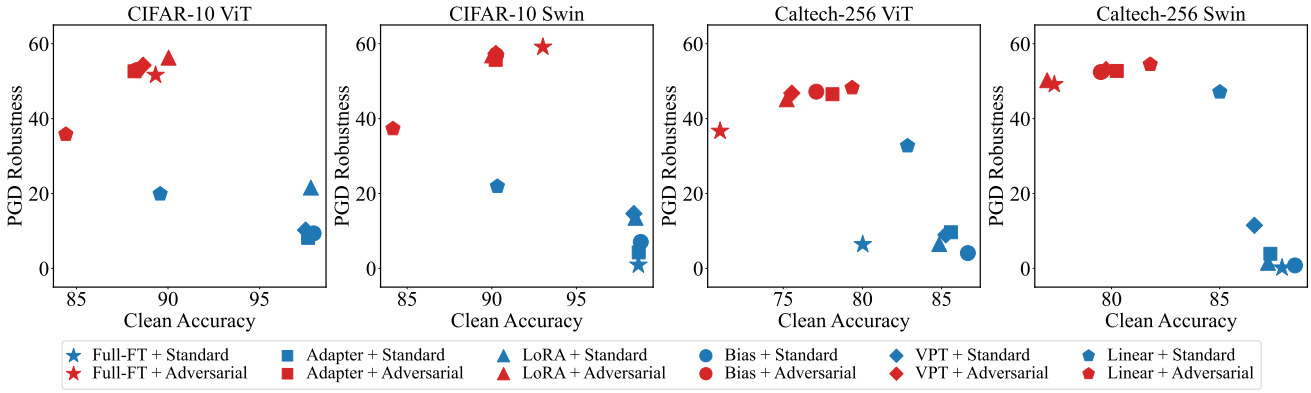

Figure 4. **Linear probing with adversarial finetuning surpasses other methods on Caltech256.** With standard finetuning, the PEFT methods exhibit inherited robustness (Robustness > 0) from the pretraining. Notably, linear probing surpasses the other methods by a substantial margin in robustness while maintaining similar clean data accuracy on Caltech256. Additionally, adversarial training during finetuning is effective in enhancing all methods' robustness.

across two model architectures, e.g., ViT and Swin, on the CIFAR10 and Caltech256 datasets. Additional details regarding the training process can be found in Sec. 5.4 and in supplementary **??**

As shown in Tab. 1, PEFT methods except Linear consistently demonstrate strong average results with adversarial finetuning. In particular, LoRA achieves the highest average robustness at 50.72% for ViT, while VPT performs exceptionally well with 55.23% on Swin.

Looking more closely at the performance on each individual dataset in Fig. 4, we observe that adversarial linear probing achieves the strongest robustness on Caltech256. This is rather surprising, given that linear probing has the lowest computational cost compared to others and never outperforms other finetuning methods on any single dataset in standard transfer learning [4, 6, 18].

5.2. Why does Adversarial Linear Probing Work?

To understand the remarkable robustness achieved by adversarial linear probing, we then ask: *what contributes to adversarial robustness on the downstream tasks given a robust pretrained model?* We hypothesize that there are two confounding factors: 1) the robustness inherited from pretrained models, and 2) the robustness achieved by adversarial finetuning. To validate our hypothesis, we conduct a comparative analysis of adversarial robustness in downstream tasks while using standard finetuning versus adversarial finetuning, given the same robust pretrained model.

Robustness Inherited from Pretrained Models. In this section, we focus on *standard finetuning* as it does not bring in additional adversarial robustness to downstream tasks. Therefore, the adversarial robustness achieved during standard finetuning primarily arises from the robust pretrained model, representing the capability of each method to preserve robustness from the pretrained model. As shown in Fig. 4, all standard finetuning methods (colored in blue) achieve non-zero robustness under PGD-10 attacks. This supports our hypothesis that robustness inherited from the pretrained model contributes to the robustness in downstream tasks. In addition, linear probing significantly outperforms all other standard finetuning methods in preserving robustness from the pretrained model. For example, linear probing exhibits significantly stronger robustness (at least 20% higher) compared to other methods while maintaining similar accuracy on Caltech256. We explain this as linear probing avoids distorting the robust features inherited from the pretrained model by only updating the linear head. However, on CIFAR10, linear probing does not exhibit a significant advantage over robustness and displays lower clean accuracy compared to others. This is likely due to its limited capability to adapt features from the pretrained source domain to the downstream target domain.

Robustness Achieved by Adversarial Finetuning. By comparing adversarial robustness achieved by standard finetuning (colored in blue) and adversarial finetuning (colored in red) in Fig. 4, we can see the extra robustness gain achieved by adversarial training during the finetuning stage. Except for Linear, adversarial finetuning contributes a significant portion of their robustness to the final robustness. For example, Full-FT on Swin, achieving a 99% final robustness on Caltech256 through adversarial finetuning. Even VPT, which has the fewest tunable parameters after Linear, demonstrates a 78% final robustness from adversarial finetuning. On the contrary, Linear only achieved a final robustness of 13% from this stage, with the primal robustness originating from the pretraining model.

Taken together, we validate our hypothesis that both robustness inherited from the pretrained model and achieved through adversarial finetuning contribute to adversarial robustness on downstream tasks. In addition, we conclude

that the exceptional robustness exhibited by adversarial linear probing is mainly because linear probing excels in preserving robustness from pretrained models compared to others, which primarily obtain their robustness from adversarial training during finetuning stage.

5.3. Transferred Robustness Correlates with Transferred Accuracy.

As we have seen in Fig. 4, adversarial linear probing achieves the highest adversarial robustness on Caltech256 but falls far behind other methods on CIFAR10. This performance discrepancy across different datasets raises a question: *how can we estimate if adversarial linear probing achieves the best robustness?* Based on our analysis in Sec. 5.2, where the robustness of adversarial linear probing primarily arises from preserving the robustness in the pretrained model, we hypothesize that the transferred robustness of adversarial linear probing is highly correlated with the transferred accuracy of standard linear probing. This suggests that if standard linear probing successfully adapts the features from the pretrained source domain to the downstream target domain, adversarial linear probing will similarly preserve the robustness in the pretrained model.

To validate this, we define **transferred accuracy** of standard linear probing (acc^T_{std}) and **transferred robustness** of adversarial linear probing (rob^T_{adv}) as follows, using full finetune as a baseline to normalize the performance across different datasets:

$$\text{acc}^T_{std} = \frac{\text{acc}(LP_{std}) - \text{acc}(FT_{std})}{\text{acc}(FT_{std})} \tag{2}$$

$$\text{rob}^T_{adv} = \frac{\text{rob}(LP_{adv}) - \text{rob}(FT_{adv})}{\text{rob}(FT_{adv})} \tag{3}$$

where $\text{acc}(\cdot)$ and $\text{rob}(\cdot)$ denote the accuracy and robustness of each method, respectively. FT_{std} and LP_{std} represent standard full finetune and standard linear probing whereas FT_{adv} and LP_{adv} stand for their respective adversarial counterparts. We trained 51 models with different hyper-parameter settings across five datasets and present transferred accuracy and robustness in Fig. 5.

In Fig. 5, the transferred robustness of adversarial linear probing is strongly correlated with the transferred accuracy of standard linear probing, with a high Pearson correlation coefficient (0.896). When the transferred accuracy approaches or surpasses 0, the transferred robustness becomes positive, indicating the improved performance of adversarial linear probing over adversarial fully finetuning. This strong correlation validates our hypothesis that if standard linear probing successfully adapts features from the pretrained to the target domain, its adversarial counterpart is capable of effectively achieving adversarial robustness on downstream tasks.

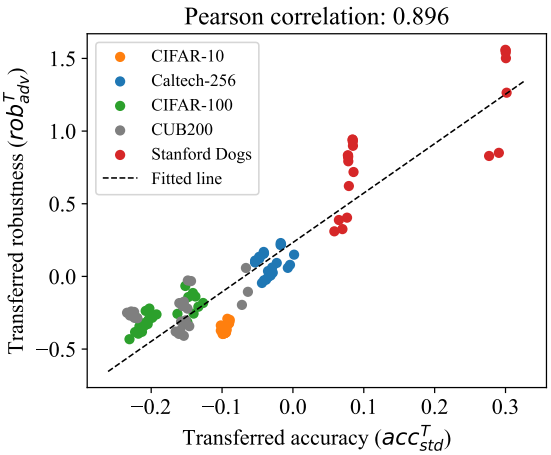

Figure 5. Transferred robustness of adversarial linear probing strongly correlates with transferred accuracy of standard linear probing.

5.4. Robust Linear Initialization (RoLI)

Based on our observation that linear probing can outperform other methods when it effectively adapts features to the target domain, we propose Robust Linear Initialization (denoted as RoLI) for adversarial finetuning. Specifically, we initialize the linear head of a robust pretrained model with weights obtained through adversarial linear probing instead of random initialization. Then, we perform adversarial finetuning, using either full finetune or PEFT methods. Since adversarial linear probing excels in preserving robustness from a pretrained model, RoLI provides a more robust initialization for subsequent adversarial finetuning, maximizing the robustness inherited from a robust pretrained model. In addition, the following adversarial finetuning can further enhance robustness by introducing additional robustness through adversarial training. In the following sections, we will evaluate the effectiveness of RoLI applied to four different adversarial finetuning methods across five different datasets and demonstrate it achieves new state-of-the-art robustness in adversarial transfer learning.

Datasets. In our experiments, we use five datasets: CIFAR10 [24], CIFAR100 [24], Caltech256 [13], Caltech-UCSD Birds-200-2011 (CUB200) [42], and Stanford Dogs (Dogs) [21]. For low-resolution image datasets like CIFAR10 and CIFAR100, we resize the image to match the model input shape. For high-resolution image datasets, we apply random resizing and cropping in training and center crop in testing. It's important to note that we integrate the resizing and normalization process directly into the model since the attacker does not have access to pre-processed images. Additional details are provided in the supplementary **??**.

Methods. We provide the results from TWINS-AT [29] AutoLoRa [43] as baselines. We proceed to assess six fine-

Methods	CIFAR10		CIFAR100		Caltech256		CUB200		Dogs		Avg	
	Clean	PGD	Clean	PGD	Clean	PGD	Clean	PGD	Clean	PGD	Clean	PGD
TWINS-AT [29]	91.24	52.73	70.72	31.08	76.86	48.40	68.09	29.24	64.98	20.58	74.38	36.41
AutoLoRa [43]	86.93	57.16	66.20	35.25	69.85	47.82	62.78	31.07	62.33	25.32	69.62	39.32
RanLI - Full-FT	93.02	59.12	74.36	35.62	77.35	49.13	70.83	31.55	65.85	24.79	76.28	40.04
RanLI - Adapter	90.23	55.70	71.57	34.74	80.24	52.73	59.67	26.73	87.98	42.24	77.94	42.43
RanLI - LoRA	90.02	56.88	71.98	34.83	77.02	50.25	64.74	29.65	84.63	39.24	77.68	42.17
RanLI - Bias	90.25	56.94	71.50	36.72	79.50	52.43	66.69	29.88	87.41	42.75	79.07	43.74
RanLI - VPT	90.24	57.38	72.37	35.60	79.74	53.07	61.94	26.46	86.34	43.10	78.13	43.12
RanLI - Linear	84.16	37.30	67.65	25.46	81.79	54.47	69.88	24.56	91.55	45.45	79.01	37.45
RoLI - Full-FT	**94.18**	**60.85**	**76.25**	**37.38**	**83.78**	**55.42**	**74.28**	**32.03**	91.55	45.57	**84.01**	**46.25**
RoLI - Adapter	90.31	55.69	71.75	34.33	81.74	54.32	64.20	25.22	**92.02**	47.12	80.00	43.34
RoLI - LoRA	92.23	58.62	72.87	35.18	81.79	54.50	68.88	26.77	91.20	46.76	81.39	44.37
RoLI - Bias	91.83	58.39	72.17	35.17	81.78	54.49	68.14	25.70	91.52	**47.27**	81.09	44.20

Table 2. **RoLI significantly improves the performance and achieves SOTA across five image classification tasks.** Robust Linear Initialization (RoLI) enhances robustness by 3.88% and clean accuracy by 2.44% on average compared to Random Initialization (RanLI). Notably, RoLI - Full-FT achieves the best overall performance, with an accuracy of 84.01% and 46.25% robustness.

Methods	CIFAR10	CIFAR100	Caltech256	CUB200	Dogs	Avg
RanLI - Full-FT	54.9	31.8	62.6	49.7	42.5	48.3
RanLI - Adapter	50.8	29.2	66.7	40.3	68.5	51.1
RanLI - LoRa	53.0	29.0	62.8	44.7	63.3	50.6
RanLI - Bias	52.3	31.0	65.8	46.2	67.3	52.5
RanLI - VPT	52.3	29.7	66.3	40.3	67.3	51.2
RanLI - Linear	26.6	18.4	65.7	40.4	73.5	44.9
RoLI - Full-FT	**56.3**	**32.7**	**70.4**	**52.1**	73.5	**57.0**
RoLI - Adapter	51.1	29.1	65.8	40.6	**75.2**	52.4
RoLI - LoRa	53.9	30.0	65.7	44.6	74.3	53.7
RoLI - Bias	53.9	29.8	65.7	43.5	74.8	53.5

Table 3. **RoLI consistently outperforms RanLI under AutoAttack.** Note, we use $\varepsilon = 8/255$ attack for CIFAR10, CIFAR100, and $\varepsilon = 4/255$ for Caltech256, CUB200 and Stanford Dogs. RoLI shows an average 3.53 AA gain compared with RanLI, and notably, RoLI - Full-FT surpasses other methods by achieving a robustness of 57.0 under AutoAttack. The clean accuracy is illustrated in Tab. 2.

Methods	$\varepsilon = 8/255$	$\varepsilon = 4/255$
RanLI - Full-FT	20.10	42.50
RanLI - Adapter	36.30	68.50
RanLI - LoRa	32.40	63.30
RanLI - Bias	35.90	67.30
RanLI - VPT	34.40	67.30
RanLI - Linear	30.90	73.50
RoLI - Full-FT	30.90	73.50
RoLI - Adapter	37.00	75.20
RoLI - LoRa	38.40	74.30
RoLI - Bias	38.20	74.80

Table 4. **RoLI consistently improves adversarial robustness even against $\boldsymbol{\varepsilon = 8/255}$ AutoAttack.** RanLI - Linear performs well under $\varepsilon = 4/255$ AutoAttack, but poorly under $\varepsilon = 8/255$ AutoAttack on Stanford Dogs. This discrepancy demonstrates that adversarial linear probing obtains its robustness mainly from the pretrained model, whereas other methods obtain robustness mainly from adversarial finetuning.

tuning techniques introduced in Sec. 3.1 with random initialization. We use Swin Transformer [28] as a backbone architecture for our experiments, whereas TWINS and AutoLoRa use ResNet50 [14]. We apply RoLI to Full-FT, Bias, Adapter, and LoRa. We initialize the newly introduced modules in Adapter and LoRA with zero to ensure that they start with robust linear probing without distorting the robust features preserved in the pretrained model.

Training Details. We use the same optimizer, AdamW [30], and a cosine scheduler with warm-up for all finetuning methods. To determine the optimal learning rate and weight decay values, we conduct a grid search, and the ranges of hyper-parameters along with the optimal combinations can be found in the supplementary **??**. We keep method-specific hyper-parameters constant across all datasets. For instance, we set the prompt length to 10 for the VPT method. To avoid adversarial overfitting [37], we apply early stopping with the epoch that has the best performance on the validation set. During adversarial finetuning, we utilize the PGD-7 attack with $\varepsilon = 8/255$ and a step size of $\alpha = 2/255$. When reporting test accuracy and robustness, we evaluate the models under a PGD-10 attack with $\varepsilon = 8/255$. Additionally, we report the performance under AutoAttack (AA), which is an ensemble attack and provides a more reliable approach to evaluate the adversarial robustness. We follow the standard AutoAttack setting, i.e., untargeted APGDCE, targeted APGD-DLR, targeted FAB [8], and Square Attack [1].

PGD-10 Results. Looking at Random Initialization (RanLI) in Tab. 2, the same trend holds as in Sec. 5.1: 1) PEFT methods except linear probing exhibit strong average results across five different datasets, with Bias achiev-

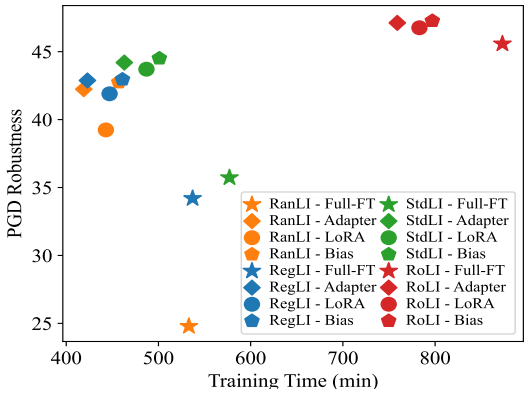

Figure 6. **Trade-off between speed and robustness.** We conduct the speed test on Stanford Dogs with random linear initialization (RanLI), logistic regression linear initialization (RegLI), standard linear initialization (StdLI), and robust linear initialization (RoLI). RoLI achieves the best performance at the cost of increased training time. Meanwhile, both RegLI and StdLI can improve their performance without incurring substantial time costs.

ing the best performance (79.25% accuracy and 43.58% robustness). 2) Linear probing achieves the highest robustness on Caltech256 and Stanford Dogs, although it falls behind others on the remaining three datasets.

Compared to RandLI, we can observe from Tab. 2 that RoLI significantly enhances the robustness on Caltech256 and Stanford Dogs for all finetuning methods. In particular, RoLI - Full-FT achieves a robustness improvement of 6.29% on Caltech256 and 20.78% on Stanford Dogs. On other datasets where adversarial linear probing with RanLI does not outperform others, RoLI performs comparable or even better than RanLI. This strongly supports that a robust initialization of the linear head is critical in adversarial transfer learning. Lastly, our RoLI - Full-FT achieves the highest performance across five datasets and sets a new state-of-the-art benchmark.

AutoAttack Results. We extend our evaluation to Stanford Dogs using $\varepsilon = 8/255$ and $\varepsilon = 4/255$, as illustrated in Tab. 4. Interestingly, we observe that adversarial linear probing outperforms other methods under $\varepsilon = 4/255$ attack but underperforms when tested against $\varepsilon = 8/255$ adversarial attacks. We conjecture that the pretrained model lacks robustness against $\varepsilon = 8/255$ attacks as it is trained using $\varepsilon = 4/255$ adversarial attacks.

This performance disparity confirms that adversarial linear probing inherently derives robustness from pretraining, whereas other methods predominantly achieve robustness from adversarial finetuning. Notably, RoLI consistently improves adversarial robustness even against $\varepsilon = 8/255$ attacks, further validating the importance of a robust linear initialization for adversarial transfer learning.

To maintain consistency with the pre-training, we follow the conventional settings [9, 10] to set $\varepsilon = 8/255$ attack for low-resolution datasets (CIFAR10, CIFAR100) and

StdLI -	CIFAR-10		Caltech-256	
	Clean	PGD	Clean	PGD
Full-FT	93.34(-0.84)	59.34(-1.51)	83.23(-0.55)	52.85(-2.57)
Bias	93.34(-0.04)	58.12(-0.27)	82.93(+1.15)	52.93(-1.56)
Adapter	90.08(-0.23)	55.19(-0.50)	82.72(+0.98)	53.01(-1.31)
LoRa	90.71(-1.52)	58.17(-0.45)	82.87(+1.08)	51.47(-3.03)

Table 5. **StdLI is less robust than RoLI.** The values in parentheses denote the performance gap (StdLI - RoLI).

$\varepsilon = 4/255$ attack for high-resolution datasets (Caltech256, CUB200, Stanford Dogs). Tab. 3 demonstrates the robustness under AutoAttack and the clean accuracy is illustrated in Tab. 2. RoLI exhibits higher robustness over RanLI, demonstrating an average 3.53 AA gain. Notably, RoLI - Full-FT achieves a robustness of 57.0, surpassing other methods.

6. Discussion

Comparison between RoLI and StdLI. Standard Linear Initialization (StdLI) refers to the use of a simple linear probing to initialize the head, while RoLI uses adversarial linear probing. In Tab. 5, we compare the performance of RoLI and StdLI on CIFAR-10 and Caltech-256. We observe that StdLI exhibits lower performance compared to RoLI, with a robustness decrease of 2.48 across three datasets, highlighting the effectiveness of RoLI.

Robustness vs. Speed. Since robust linear initialization involves a two-step adversarial training process, it tends to be considerably slower than adversarial fine-tuning with random initialization. In this section, we show a trade-off between robustness and speed by obtaining the initialization from four different approaches: adversarial linear probing (RoLI), standard linear probing (StdLI), logistic regression (RegLI), and randomization (RanLI). From Fig. 6, it's evident that both RegLI and StdLI enhance robustness compared to random linear initialization within an acceptable time cost. In addition, RoLI achieves the highest performance but at the expense of slower speed.

7. Conclusion

This paper systematically investigates how to achieve adversarial robustness in downstream tasks. We highlight the necessity of an adversarially robust pretraining. Given a robust pretrained model, we propose to use robust linear initialization (RoLI) followed by adversarial full finetuning or PEFT methods to achieve the best performance. We demonstrate that RoLI outperforms random linear initialization across five image classification tasks. We hope the insights from this study greatly advance research aimed at enhancing adversarial robustness in downstream tasks.

Acknowledgment. We thank Chenhe Gu and Xuan Yang for the valuable discussion and insightful feedback.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pages 484–501. Springer, 2020. 7

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

[3] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. In *Advances in Neural Information Processing Systems*, pages 11285–11297, 2020. 2, 3

[4] Arnav Chavan, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen. One-for-all: Generalized lora for parameter-efficient fine-tuning. *arXiv preprint arXiv:2306.07967*, 2023. 1, 5

[5] Dian Chen, Hongxin Hu, Qian Wang, Li Yinli, Cong Wang, Chao Shen, and Qi Li. Cartl: Cooperative adversarially-robust transfer learning. In *International Conference on Machine Learning*, pages 1640–1650. PMLR, 2021. 3

[6] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 5

[7] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 699–708, 2020. 1, 3

[8] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pages 2196–2205. PMLR, 2020. 7

[9] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 8

[10] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 8

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 4

[13] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 2, 4, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[16] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 1, 2, 3

[17] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 2, 3

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 1, 2, 3, 5

[19] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Advances in neural information processing systems*, 33: 16199–16210, 2020. 1, 3

[20] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 1

[21] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*. Citeseer, 2011. 2, 6

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 1

[23] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. 2

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4, 6

[25] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2021. 2

[26] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. 1

[27] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35:109–123, 2022. 1, 2

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1, 4, 7

[29] Ziquan Liu, Yi Xu, Xiangyang Ji, and Antoni B Chan. Twins: A fine-tuning framework for improved transferability of adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16436–16446, 2023. 1, 3, 6, 7

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 7

[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 3

[32] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, 2020. 2, 3

[33] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, 2021. 2, 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[35] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 1

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1

[37] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020. 7

[38] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020. 3

[39] Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning. In *International Conference on Learning Representations*, 2019. 3, 4

[40] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-jui Hsieh. On the adversarial robustness of vision transformers. In *Annual Conference on Neural Information Processing Systems*, 2022. 1

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[42] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[43] Xilie Xu, Jingfeng Zhang, and Mohan Kankanhalli. Autolora: A parameter-free automated robust fine-tuning framework. *arXiv preprint arXiv:2310.01818*, 2023. 1, 3, 6, 7

[44] Yutaro Yamada and Mayu Otani. Does robustness on imagenet transfer to downstream tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9215–9224, 2022. 1

[45] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 2

[46] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, 2022. 2, 3

[47] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 2

[48] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022. 1