Task-Robust Pre-Training for Worst-Case Downstream Adaptation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Pre-training has achieved remarkable success when transferred to downstream tasks. In machine learning, we care about not only the good performance of a model but also its behavior under reasonable shifts of condition. The same philosophy holds when pre-training a foundation model. However, the foundation model may not uniformly behave well for a series of related downstream tasks. This happens, for example, when conducting mask recovery regression where the recovery ability or the training instances diverge like pattern features are extracted dominantly on pre-training, but semantic features are also required on a downstream task. This paper considers pre-training a model that guarantees a uniformly good performance over the downstream tasks. We call this goal as *downstream-task robustness*. Our method first separates the upstream task into several representative ones and applies a simple minimax loss for pre-training. We then design an efficient algorithm to solve the minimax loss and prove its convergence in the convex setting. In the experiments, we show both on large-scale natural language processing and computer vision datasets our method increases the metrics on worse-case downstream tasks. Additionally, some theoretical explanations for why our loss is beneficial are provided. Specifically, we show fewer samples are inherently required for the most challenging downstream task in some cases.

1 Introduction

The rapid development of machine learning is promoting a shift in the learning paradigm, where one first trains a very large model, often called a foundation model, with massive data, and then adapts it to desired tasks using much less data. The hope is to obtain a model that serves as an infrastructure and is transferable to a wide range of tasks. Pre-training plays the role of an engine to acquire the foundation model. Typical pre-training methods are to minimize the average expected risks of the upstream tasks. Such pre-trained models can achieve good performance for a lot of downstream tasks but may fail in some hard cases. For example, a vision pre-trained model for animal and plant recognition may work well for typical characteristics species but fail when identifying mimicry animals and plants; a common green mantis can be correctly recognized as a mantis, while an orchid mantis might be falsely classified as an orchid.

In machine learning, one cares about not only the good performance of a model but also its behavior under reasonable shifts of conditions. The same philosophy holds for pre-training a foundation model. To guarantee a uniformly good performance over a series of tasks, one has to consider the robustness of pre-training. We call this *downstream-task robustness*. The aim is to develop a pre-training method to train the foundation model that admits a good adaptation performance over a series of downstream tasks. It is crucial to achieving the downstream-task robustness for pre-training: (i) safety is critical

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

for some applications, such as deep learning systems in medicine and finance; (ii) our goal of the foundation model requires reliably good performance on all downstream tasks.

In recent times, popular large-scale models like ChatGPT [29] have also faced safety issues, sparking discussions among various stakeholders [5]. The concerns largely stem from the potential misuse or unintended behavior of the model in real-world applications. Some argue that the vast and diverse knowledge base of these models, coupled with their ability to generate human-like text, could be leveraged for malicious purposes. Others worry about the potential of the model to generate inappropriate or harmful content [30].

Addressing these safety concerns is crucial, particularly in the context of pre-training foundation models that are intended for downstream tasks. An initial way to mitigate these safety issues is to consider downstream-task robustness. By focusing on downstream-task robustness, we expect that the models are resilient to perturbations in the input data, thereby reducing their susceptibility to adversarial attacks or misuse. This approach can help in maintaining consistent performance across a range of tasks and so enhance the safety and reliability of the model.

In recent years, the concept of Distributionally Robust Optimization (DRO) [3, 27, 21] has attracted wide attention among theorists and practitioners. Most DRO frameworks [33, 14, 13] consider training a parameterized model that minimizes the worst-case expectation loss over the data from a family of probability distributions. Downstream-task robustness can be considered a generalization of DRO. The destination of downstream-task robustness is to guarantee good worst-case performance for a series of downstream adaptations.

This paper proposes a pre-training method as a starting point for downstream-task robustness. To take a step forward, our method considers learning several upstream tasks. The choice of how to design the upstream tasks allows us to incorporate prior knowledge of the domain and problems. For example, in language models, we can design upstream tasks by masking different types of words; how we generate such upstream tasks by grouping samples reflects our prior knowledge of the natural language. Then instead of minimizing the average expected risk of the upstream tasks, we minimize the worst-case expected risk. We also introduce a simple but practical algorithm called softmax weighted gradient descent to pre-train the model. We prove the algorithm's convergence in the ideal convex setting and show its effectiveness in our empirical study.

We consider the application of the framework in two experiments — Part-of-Speech masked language models in section 4.1 and multi-modal masked image models in section 4.2. We first pre-train a foundation model with the proposed task-robust pre-training method on multiple upstream tasks generated by different masks and adapt the foundation model for downstream tasks. Compared with the average expected risk minimization, our method achieves better worst-case performance and comparable average performance on all downstream tasks.

We also explain why our framework can benefit downstream-task robustness. Specifically, by simplifying the model-task relationship, we show fewer samples are needed for the hardest downstream task. The key intuition is that proper worse-case training for upstream tasks leads to an initiation close to the solution for the worst-case downstream tasks, thus reducing the downstream burden.

The contributions of our study can be primarily encapsulated within three key areas: (i) The introduction of the concept of task-robust pretraining, a novel theoretical framework that holds significant potential for future research. (ii) The provision of a simple yet efficacious method for task-robust pretraining, accompanied by a comprehensive exposition of its theoretical feasibility. (iii) A series of empirical validations across multiple domains, substantiating the effectiveness of our methodology.

2 Setup and Methodology

Consider a traditional machine learning task where the data $z \in \mathbb{Z}$ follows an underlying distribution P. Given a model, a parameter space $\Theta \subset \mathbb{R}^d$, a loss function $\ell : \Theta \times \mathbb{Z} \mapsto \mathbb{R}_+$, the goal is to find the optimal parameter θ^* for the model such that $\theta^* = \arg\min_{\theta\in\Theta} \mathbb{E}_{z\sim P}\left[\ell\left(\theta, z\right)\right]$. The classic empirical risk minimization (ERM) tackles the problem by first collecting i.i.d. training data $\{z_i\}_{i=1}^N$ from P and then finding a parameter $\hat{\theta}_{\text{ERM}}$ via minimizing the empirical risk:

$$\hat{\theta}_{\text{ERM}} := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{N} \ell\left(\theta, z_i\right). \tag{1}$$

In statistical learning theory, it is well-known that under mild conditions (such as the VC dimension of the model is bounded above), $\hat{\theta}_{\text{ERM}}$ is a good approximation of θ^* in the sense that with high probability at least $1-\delta$ ($0 < \delta \ll 1$),

$$\mathbb{E}_{z\sim P}\left[\ell\left(\hat{\theta}_{\text{ERM}}, z\right)\right] - \min_{\theta\in\Theta}\mathbb{E}_{z\sim P}\left[\ell\left(\theta, z\right)\right] \leq \epsilon, \tag{2}$$

when the number of training samples N is sufficiently large. We simply denote the training data requirement by $N(\epsilon, \delta)$.

In machine learning, a foundation model is trained and then adapted for each downstream task. The adaptation process involves initializing a downstream model with the pre-trained foundation model's parameters and then training on the downstream task. Denote the parameter of the foundation model by $\theta_{\text{foundation}}$. Let Λ be the downstream task space. The goal is to find an initial parameter $\theta_{\text{foundation}}$ that enables fast adaptations for each downstream task $\lambda \in \Lambda$. Different downstream tasks are characterized by different loss functions with a shared data distribution [1]. We use the foundation model's parameters $\theta_{\text{foundation}}$ as initialization to find an approximately optimal parameter $\hat{\theta}_{\lambda,\text{ERM}}$ by ERM. We study the sample complexity required to guarantee a good approximate solution with high probability while considering the effect of initialization. For the task λ and the model initialized by $\theta_{\text{foundation}}$, denote the number of samples required to find an ϵ-approximately optimal parameter by ERM with probability at least $1-\delta$ as $N_\lambda\left(\theta_{\text{foundation}}, \epsilon, \delta\right)$. The aim is to find the optimal initialization that minimizes the worst-case sample complexity required to find an approximately optimal parameter for all tasks, i.e.,

$$\theta^*_{\text{foundation}} := \underset{\theta_{\text{foundation}}\in\Theta}{\arg\min}\ \max_{\lambda\in\Lambda} N_\lambda\left(\theta_{\text{foundation}}, \epsilon, \delta\right). \tag{3}$$

Directly training for the optimal worst-case initialization is generally infeasible or computationally expensive. Pre-training provides a feasible alternative $\theta^*_{\text{pre-train}}$ for $\theta^*_{\text{foundation}}$ by training for available surrogate upstream tasks. When the upstream tasks are related to the downstream tasks, the pre-trained parameter can lead to lower initial expected risks and accelerate downstream training. For example, if we pre-train a model on generated upstream tasks of reconstructing images corrupted by different masks, the pre-trained model can learn some prior knowledge for general vision tasks; with the pre-trained parameter as the initialization, we can accelerate the training process of downstream vision tasks such as image classification and object detection [8]. Consider there are T representative upstream tasks. Denote the loss function of the task t as ℓ_t. A typical choice of the pre-trained parameter is the minimizer of the average expected risk over the T upstream tasks [26, 10, 17], i.e.,

$$\theta^*_{\text{average}} := \arg\min_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{z\sim P}\left[\ell_t\left(\theta, z\right)\right]. \tag{4}$$

However, minimizing the average expected risk over upstream tasks may neglect extreme cases and lead to limited benefit for some downstream tasks.

To alleviate the aforementioned issue of the average expected risk minimization, we propose to use the minimizer of the worst-case expected risks over the upstream tasks, i.e.,

$$\theta^*_{\text{max}} := \arg\min_{\theta\in\Theta}\max_{t\in[T]}\mathbb{E}_{z\sim P}\left[\ell_t\left(\theta, z\right)\right], \tag{5}$$

as the initial parameter, where $[m]$ denotes the set $\{1, \ldots, m\}$. We show that θ^*_{max} is a better choice than $\theta^*_{\text{average}}$ in terms of downstream-task robustness.

3 Algorithm

Recall that our minimax pre-training method is to minimize the worst-case expected risks over the upstream tasks, i.e.,

$$\min_{\theta\in\Theta}\max_{t\in[T]}\mathbb{E}_{z\sim P}\left[\ell_t\left(\theta, z\right)\right] \tag{6}$$

[1] Data distributions vary for different tasks can be modelled by incorporating weighting functions from the data distributions to the loss functions. Consider a task where the data distribution is P_t and $\frac{dP_t}{dP}(z) > 0$ for all $z \in \text{supp}(P_t)$, the expected loss is $\mathbb{E}_{z\sim P_t}\left[\ell(\theta, z)\right]$. The expected loss can still be rewritten as $\mathbb{E}_{z\sim P}\left[\ell_t(\theta, z)\right]$, where $\ell_t(\theta, z) = \frac{dP_t}{dP}(z)\ell(\theta, z)$.

There is extensive literature on minimax optimization. The minimax optimization algorithms can be generally classified into two types: (i) minimization for the maximum function $\max_{t\in[T]} \mathbb{E}_{z\sim P}\left[\ell_t\left(\theta, z\right)\right]$ [7, 2, 36, 16], and (ii) direct minimax optimization for the objective [20, 35, 28, 22, 23]. We introduce a new simple optimization algorithm called softmax weighted gradient descent (Algorithm 1) that we find is very practical to pre-train the model. The algorithm can be roughly seen as the first type. It is an adaptation of the classic subgradient descent for minimizing the maximum function to enable its practical use in pre-training. Concretely, in one update, we take a descent step at the current point θ along the direction of the gradient weighted by softmax-type weights, i.e., $\sum_{t=1}^{T} w_{\alpha,t}\left(\theta\right) \nabla_\theta \mathbb{E}_{z\sim P}\left[\ell_t\left(\theta, z\right)\right]$, where

$$w_{\alpha,t}\left(\theta\right) := \frac{\exp\left(\alpha \mathbb{E}_{z\sim P}\left[\ell_t\left(\theta, z\right)\right]\right)}{\sum_{t'=1}^{T} \exp\left(\alpha \mathbb{E}_{z\sim P}\left[\ell_{t'}\left(\theta, z\right)\right]\right)}, \tag{7}$$

and $\alpha > 0$ is a hyperparameter. (In practice, we use estimations for the expected risks and the gradients on minibatch samples.)

The motivation behind softmax weighted gradient descent is to use the softmax weighted gradient to approximate the subgradient in the classic subgradient descent for the minimax optimization of (6). As the softmax weighted gradient descent algorithm optimizes for the minimax loss directly, it can achieve better worst-case loss than other pre-training methods. One advantage of the softmax approximation is that it avoids the non-differentiability caused by the maximum operator via the softmax approximation, making the algorithm easily implementable for pre-training applications. Also, as the softmax weighted gradient descent includes only single weighted gradient step in each update, it has computational efficiency comparable to gradient descent in deep learning. In contrast, standard minimax algorithms often cost several times gradient oracles in single step. Moreover, our algorithm can be directly combined with commonly-used optimization tricks in deep learning, such as momentum and adaptive learning rates. In our experiments, we observe that the algorithm with a simple implementation achieves better worst-case errors in various real-world tasks than a number of benchmark pre-training algorithms.

Algorithm 1 Softmax Weighted Gradient Descent

Input: Step sizes $\{\eta_k\}_{k=1}^{K-1}$, softmax hyperparameters $\{\alpha_k\}_{k=0}^{K-1}$ and an initial parameter θ_0;
for $k = 1, \ldots, K-1$ **do**
  Compute the softmax weights $\{w_{\alpha_k,t}\left(\theta_{k-1}\right)\}_{t=1}^{T}$ as in (7);
  Update the parameter as $\theta_k \leftarrow \theta_{k-1} - \eta_k \sum_{t=1}^{T} w_{\alpha_k,t}\left(\theta_{k-1}\right) \nabla_\theta \mathbb{E}_{z\sim P}\left[\ell_t\left(\theta_{k-1}, z\right)\right]$;
end for

For completeness, we also provide some convergence analysis for our proposed algorithm. We consider a relatively basic setting where for all $t \in [T]$, the loss function $\ell(\cdot, z)$ is convex and L'-Lipschitz continuous for any fixed $z \in \mathbb{Z}$ Intuitively, when the hyperparameter α_k is sufficiently large, the function $\sum_{t=1}^{T} w_{\alpha_k,t}\left(\theta_k\right) \mathbb{E}_{z\sim P}\left[\ell_t\left(\theta, z\right)\right]$ is a good differentiable approximation for the objective $\mathbb{E}_{z\sim P}\left[\ell_t\left(\theta, z\right)\right]$ Softmax weighted gradient descent can be roughly seen as a remedy for non-differentiability in subgradient descent, at the expense of controllable approximation errors. We show in Theorem 3.1 that Algorithm 1 can achieve a convergence rate $O\left(\frac{1}{\sqrt{K}}\right)$ if the hyperparameter α_k is as large as $\tilde{O}(\sqrt{k})$. This result is comparable to the standard convergence rate $O\left(\frac{1}{\sqrt{K}}\right)$ of subgradient descent [6, Chapter 3].

Theorem 3.1. *Suppose that for all $t \in [T]$ the loss function $\ell_t(\cdot, z)$ is convex, L'-Lipschitz continuous and bounded by B for all $\theta \in \Theta$ and any fixed $z \in \mathbb{Z}$. Denote the optimal solution of* (6) *as θ^* and the distance $\|\theta_0 - \theta^*\|$ as R_0. If the step size $\eta_k = \eta = \frac{R_0}{L'\sqrt{K}}$ and the softmax hyperparameter $\alpha_k \geq \frac{4\sqrt{k+1}}{R_0 L'} \log \frac{4TB\sqrt{k+1}}{R_0 L'}$ for all $k = 0, \ldots, K-1$, the average $\bar{\theta}_K$ of the iteration points in Algorithm 1, i.e., $\bar{\theta}_K = \frac{1}{K}\sum_{k=0}^{K-1} \theta_k$, satisfies*

$$\max_{t\in[T]} \mathbb{E}_{z\sim P}\left[\ell_1\right] - \min_{\theta\in\Theta} \max_{t\in[T]} \mathbb{E}_{z\sim P}\left[\ell_2\right] \leq \frac{2R_0 L'}{\sqrt{K}}, \tag{8}$$

where $\ell_1 = \ell_t\left(\bar{\theta}_K, z\right)$ and $\ell_2 = \ell_t\left(\theta, z\right)$.

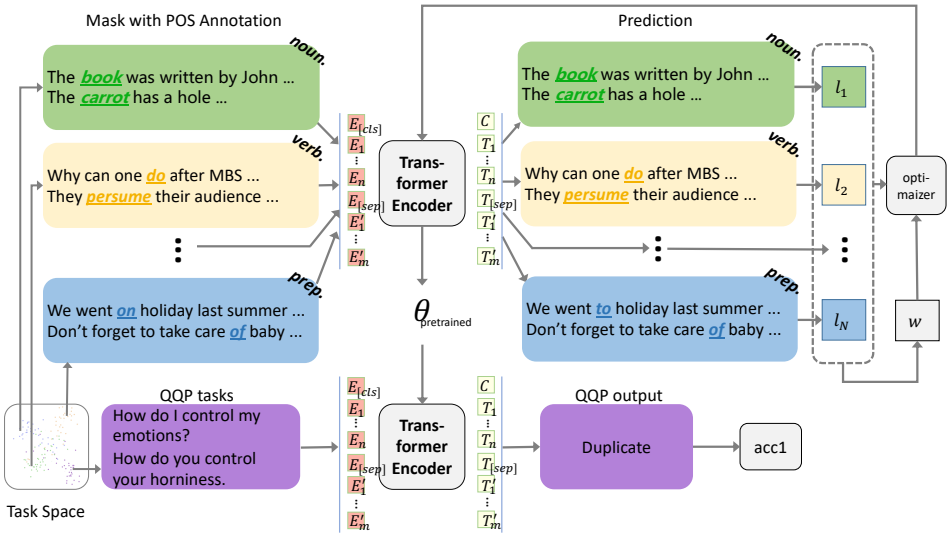


Figure 1: **Part-of-Speech Mask BERT.** We first sample the datasets from the task space and group them according to different Part-of-Speech types, and then recover the predicted sentence by a BERT encoder. The optimizer selects the most challenging task and then updates the model's weight through a minimax layer.

Table 1: Results on GLUE. The "Averages" are obtained from GLUE leaderboard. F1 scores are reported for QQP and MRPC. , spearman correlations are reported for STS-B, Matthews correlations are reported for CoLA, and accuracy scores are reported for the other tasks.

Task-Balancing	**MNLI**	**QQP**	**QNLI**	**SST-2**	**CoLA**	**STS-B**	**MRPC**	**RTE**	**Avg.**
None	84.6	71.2	**90.5**	**93.5**	52.2	**85.8**	**88.9**	66.4	79.6
Minimax (Ours)	**85.6**	**76.9**	88.6	91.3	**61.4**	84.2	88.2	**70.7**	**81.4 (+1.8)**

Remark 3.2. The convexity assumption on the loss functions is an oversimplification in deep learning. However, for some cases such as the neural tangent kernel [18], the deep neural networks exhibit properties similar to convexity. In our experiments with non-convex models, we also observe that the algorithm behaves well.

Remark 3.3. The above analysis requires increasing softmax hyperparameters $\{\alpha_k\}_{k=0}^{K-1}$, i.e., $\alpha_k = \tilde{O}\left(\sqrt{k}\right)$ to guarantee the convergence rate. In our experiments, however, we find that constant softmax hyperparameters, or more concretely $\alpha_k = 1$ for all $k = 0, \ldots, K-1$, work well for most problems. We attribute these phenomena to some properties of deep neural networks, which are left for future exploration.

4 Experiments

In this section, we subject our methods to rigorous testing through two experiments, each encompassing tasks germane to the fields of Natural Language Processing (NLP) and Computer Vision (CV). A minimalist design approach was adopted for both the models and the tasks to demonstrate the universality of our design across a broad spectrum of model tasks. We extended the functionalities of BERT and MAE, thereby constructing Part-of-Speech Mask BERT (PoS-BERT) and Multi-Modal Mask MAE (MM-MAE), respectively.

4.1 NLP Scenario: Part-of-Speech Mask BERT

4.1.1 Model and Settings

Architectures An overview of the Part-of-Speech Mask BERT model is shown in Figure 1. PoS-BERT model first samples the datasets from the task space and groups them according to different Part-of-Speech types. The loss function term is calculated separately for each data group entering

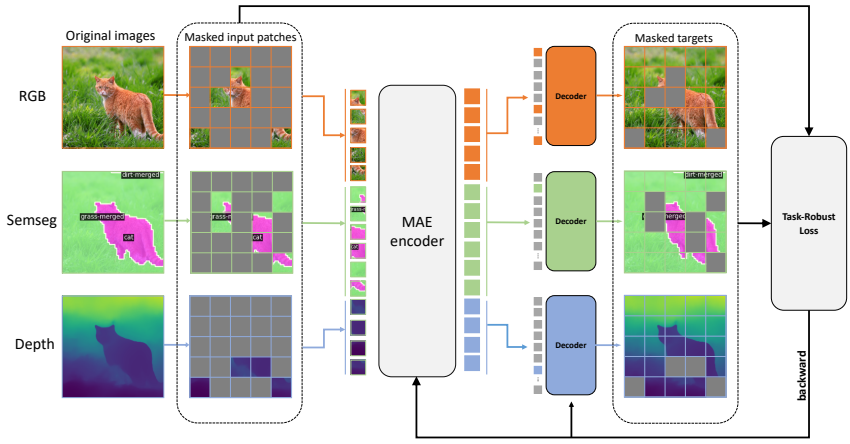


Figure 2: **Multi-modal Mask MAE:** Randomly sampled patches from multiple modalities are projected to tokens. Task-specific decoders reconstruct the masked-out patches by first performing a cross-attention step from queries to the encoded tokens.

the BERT encoder. The loss term with the highest weight is selected to enter the optimizer through a minimax layer. Finally, we run experiments on downstream tasks to compare our minimax task balancing learning algorithm with other methods to verify our theory. We follow the common practice to design the feature representations for masked language modeling and next-sentence prediction.

Tasks and Datasets During pre-training, Part-of-Speech Mask BERT has two objectives: masked language modeling and next-sentence prediction. We define 9 task categories that recover different parts of speech-type words on masked language modeling: (1)verb, (2) noun, (3) adjective, (4) determiner, (5) adverb, (6) pronoun, (7) preposition, (8) conjunction, (9) interjection. Following previous work [10, 31, 25] , we evaluate our pre-trained models on downstream tasks using the GLUE [38] benchmarks. Downstream tasks we fine-tuned include MNLI, QQP, QNLI, SST-2, CoLA, STS-B, MRPC, and RTE. By swapping out the appropriate inputs and outputs, Part-of-Speech Mask BERT can model many downstream tasks and has a unified way to handle the tasks that involve single text and text pairs. After setting the masks, we utilize Natural Language Toolkitannotate (NLTK) [4] to pseudo-label the words with POS annotations.

4.1.2 Quantitative Result

Quantitative results are presented in Table 1. Our model obtains comparable results on GLUE tasks. PoS-BERT outperforms on half tasks by a substantial margin and obtains a 1.8% average score improvement over BERT. As for the most challenging training task, CoLA, which has the lowest accuracy on BERT, our method gets a 9.2% improvement which is a significant boost among downstream tasks. Benefiting from task-robust grouping, on QQP and RTE tasks, our method outperforms the original BERT by 5.7 F1-score and 4.3% accuracy, respectively. Our method also shows superiority on the challenging downstream task, MNLI, by achieving a 1.1% higher matched accuracy. As compensation for working better on the more challenging tasks, our method loses little correctness on some downstream tasks that already transfer well. On QNLI, SST-2, STS-B, and MRPC, our results are lower than that of the original BERT model by a margin of 1% accuracy, 1.9% accuracy, 1.6 spearson correlation, and 0.7 F1-score. The empirical result shows that, with our task-robust pre-training strategy, the downstream tasks perform on the whole, especially those tricky tasks. We expect future work to further improve these results by incorporating more sophisticated multi-task and grouping procedures.

4.2 CV Scenario: Multi-Modal Mask MAE

4.2.1 Model and Settings

Architectures An overview of MM-MAE is shown in Figure 4.1.2. Multi-Modal Mask MAE contains three encoders, each of which processes different modalities of one image. During pre-

Table 2: Comparison between task-robust method and other task-balancing methods on ImageNet1K and ImageNetS50 pre-training. Our methodology demonstrates superior performance across the majority of downstream tasks, particularly excelling in the most challenging among them.

Pre-training			**Downstream**		
Data	Epoch	Task-Balancing	Cls. (Top-1 Acc. %)	Semseg. (mIoU)	Depth. (δ_1 Acc. %)
ImageNetS50	800	None	92.2	51.9	52.1
		Uncertainty [19]	92.6	54.5	70.2
		GradNorm [9]	93.0	56.5	65.8
		DWA [24]	**93.4**	52.7	65.7
		Minimax(Ours)	91.8	**61.5**	**74.1**
ImageNet1K	400	Uncertainty	**82.6**	48.9	85.2
		Minimax(Ours)	82.3	**50.1**	**85.3**
	1600	Uncertainty	**83.3**	52.0	86.4
		Minimax(Ours)	83.0	**53.2**	**86.8**

Table 3: A qualitative comparison between task balancing techniques. T representing the computation cost when no additional task-balancing techniques are employed.

Method	**Balance Magnitude**	**Balance Learning**	**Grads Required**	**No Extra Tuning**	**FLOPs**	**Motivation**
None	✓			✓	T	/
Uncertainty	✓			✓	$2T$	Homoscedastic uncertainty
Gradnorm	✓	✓	✓	✓	$4T$	Balance learning and magnitudes
DWA		✓			$3T$	Balance learning
Minimax (Ours)	✓			✓	$2T$	Task robust

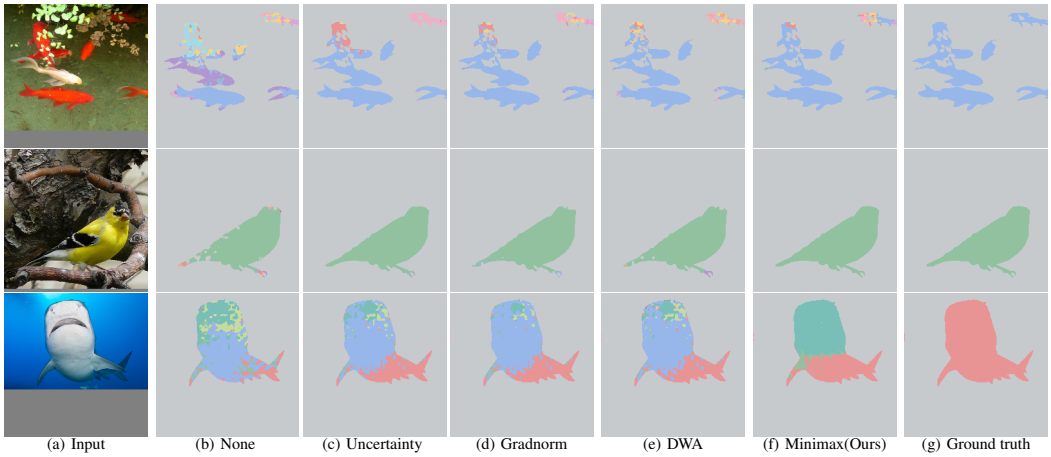
training, we try to recover each modality from its masked tokens. Each modality is divided into 16×16 patches and then tokenize the patches with modality-dependent linear projections. Projected patches are concatenated into a sequence of tokens and given as input to the same transformer encoder. We also add a global token with 2D sine-cosine positional embeddings. Each task owns a specialized decoder, and the computational cost of decoders scales linearly with the number of tasks.

Tasks and Datasets We select two datasets with different scales, ImageNet1K [32] and ImageNetS50 [15], to conduct unsupervised training upstream to see whether the minimax pre-training method can help the downstream tasks with poor performance. The classification task is evaluated on the validation part of the original dataset, while the semantic segmentation and depth estimation tasks are validated on the NYUv2 dataset [34] by fine-tuning. Due to the absence of a sizeable multi-task dataset with aligned task images [11, 1] we generate pseudo-labels on ImageNet and ImageNetS50 with GPT-3 and Mask2Former.

4.2.2 Quantitative Result

Classification tasks The quantitative results are presented in Table 2. We evaluate our models and baseline by fine-tuning them on the supervised ImageNetS50 and ImageNet1K. We fine-tune our models for 100 epochs and report the top-1 validation accuracy. The result shows a tiny gap between our method and the average method in the classification task. Classification tasks are regarded as the least challenging task category of the three. Cause different downstream tasks have different optimal parameter requirements, this gap is unavoidable. After the pre-training of the model reaches a specific step, the training weight of the classification tasks will continue to decrease.

Semantic segmentation tasks We further evaluate our models on semantic segmentation tasks on the NYUv2 dataset. We report the mean intersection over the union (mIoU) metric. Notice that semantic segmentation is the most challenging task of these downstream transfers. Our method benefits more than the average loss model from pseudo-labeled modalities as input. In particular, the correctness is improved by 9.6% on ImageNetS50 pre-training. With the progress of model training, our task-robust loss forces the model to improve poorly trained semantic segmentation tasks by increasing the training weight. The following section 5 will explain why a simple strategy can significantly help worst-case downstream tasks.



(a) Input (b) None (c) Uncertainty (d) Gradnorm (e) DWA (f) Minimax(Ours) (g) Ground truth

Figure 3: Comparative intermediate results. The first and final columns represent the image input and ground truth, respectively, while the intermediary images depict the intermediate results yielded by various task-balancing methodologies. Distinct colors correspond to the prediction of different objects. Our approach ensures robustness in downstream tasks.

Depth evaluate tasks For depth estimation, we use NYUv2 . We report δ_1 on the NUYv2 test set, showing the percentage of pixels p with error max $\{\frac{y_p}{\hat{y_p}}, \frac{\hat{y_p}}{y_p}\}$ less than 1.25 [12]. According to Table 2, the accuracy is improved by 3.9% on ImageNetS50 pre-training with the help of the downstream-task robustness loss function. The depth estimation task in the same data volume has a higher tolerance for prediction errors per pixel than semantic segmentation. However, it is still more complicated than the classification task, which only predicts the image once. After the classification task is well-trained, the depth estimation task will benefit from our strategy in the subsequent training.

4.3 Qualitative Comparison

Table 2 delineates several strategies designed to equilibrate the contribution of each task during the training of a multi-task network. For a qualitative comparison of these methods, refer to Table 3. We appraise these strategies based on several criteria [37]. An overview of our examination suggests that our approach achieves a synergistic blend of simplicity, efficiency, and effectiveness.

To facilitate a more intuitive comparison of the differences in results throughout the training process, we undertook downstream tasks in semantic segmentation, comparing the performance of various approaches midway through pre-training (400 epoch, ImageNetS50). As depicted in Figure 3, our methodology exhibits superior performance, even under conditions of insufficient training.

5 Explanation

We show why the proposed minimax pre-training method can be more effective than the average expected risk minimization in some cases. We consider a simplification of the model and the task relationship. Such a simplification makes our analysis convenient and intuitive. We assume that for all $t \in [T]$, the function $\ell_t(\cdot, z)$ is μ-strongly-convex, L-smooth, and L'-Lipschitz continuous for any fixed $z \in \mathbb{Z}$ and the function $\ell_t(\cdot, \cdot) \leq B$ for all $\theta \in \Theta$ and $z \in \mathbb{Z}^2$. Note that pre-training on irrelevant upstream tasks does little help to the downstream tasks in general. Here, we only discuss the case where the upstream tasks and the downstream tasks are closed related. We ideally assume that the loss functions of the downstream tasks are convex combinations of the loss functions of the upstream tasks, i.e.,

$$\ell_\lambda = \sum_{t=1}^{T} \lambda_t \ell_t, \text{ for all } \lambda \in \Lambda = \Delta_T, \tag{9}$$

[2]In fact, it is sufficient to assume that the expectation $\mathbb{E}_z[\ell_t(\cdot, z)]$ is μ-strongly-convex, L-smooth, and L'-Lipschitz continuous.

where Δ_T is the $(T-1)$-dimensional probability simplex. We further assume that for each task there exists a parameter such that the expected risk of the task is zero, i.e., $\min_{\theta\in\Theta}\mathbb{E}_{z\sim P}\left[\ell_t\left(\theta,z\right)\right]=0$ for all $t\in[T]$.

We first show that in the above setting, the proposed minimax optimization pre-training method can guarantee a better worst-case initial expected risk than the minimization method.

Proposition 5.1. *Let θ^*_{max} and $\theta^*_{average}$ be the pre-trained parameters obtained by minimizing the maximal expected risk and the average expected risk, respectively. Then for the worst-case expected risks of the downstream tasks, we have*

$$\max_{\lambda\in\Lambda}\mathbb{E}_{z\sim P}[\ell_\lambda(\theta^*_{max},z)]\leq\max_{\lambda\in\Lambda}\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta^*_{average},z\right)\right]. \tag{10}$$

Remark 5.2. The gap between $\max_{\lambda\in\Lambda}\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta^*_{\text{max}},z\right)\right]$ and $\max_{\lambda\in\Lambda}\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta^*_{\text{average}},z\right)\right]$ can be large. We provide an example in the appendix, where the ratio between $\max_{\lambda\in\Lambda}\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta^*_{\text{average}},z\right)\right]$ and $\max_{\lambda\in\Lambda}\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta^*_{\text{max}},z\right)\right]$ is as large as $O(T)$.

We then illustrate that a good initialization can serve as an implicit regularization. We suppose that the downstream tasks are trained with gradient descent. (For stochastic gradients with bounded variances, the analysis below also holds for sufficiently small step sizes, within neglectable approximation errors.) For the downstream task λ, with certain step sizes, the parameters will always be in a subset

$$\Theta_\lambda(\theta_0)=\left\{\theta\in\Theta\mid\|\theta-\theta^*_\lambda\|^2\leq\frac{2}{\mu}\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta_0,z\right)\right]\right\}, \tag{11}$$

where θ_0 is the initial parameter and $\theta^*_\lambda=\arg\min_{\theta\in\Theta}\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta,z\right)\right]$.

Proposition 5.3. *Suppose that a function $f:\mathbb{R}^d\mapsto\mathbb{R}$ is μ_f-strongly-convex and L_f-smooth for all $x\in\mathbb{R}^d$ and $x^*\in\arg\min_{x\in\mathbb{R}^d}f(x)$. Let $\{x_k\}_{k=0}^{K-1}$ be the sequence generated by gradient descent with a step size $\eta>0$, i.e., $x_k=x_{k-1}-\eta\nabla f(x_{k-1})$ for all $k\in[K-1]$. If the step size $\eta\leq\frac{1}{L_f}$, then we have*

$$\|x_k-x^*\|^2\leq\frac{2}{\mu_f}\left(f(x_0)-f(x^*)\right), \tag{12}$$

for all $k=0,1,\ldots,K-1$.

By Proposition 5.3, we can deem that the downstream task's parameter space is the subset $\Theta_\lambda(\theta_0)$.

Consider the worst sample complexity to find an ϵ-approximately optimal parameter by ERM within the parameter space $\Theta_\lambda(\theta_0)$ for a downstream task $\lambda\in\Lambda$.

Theorem 5.4. *The worst-case sample complexity $\max_{\lambda\in\Lambda}N_\lambda\left(\theta_0,\epsilon,\delta\right)$ with initialization θ_0 satisfies*

$$\max_{\lambda\in\Lambda}N_\lambda\left(\theta_0,\epsilon,\delta\right)\leq\frac{8dB^2}{\epsilon^2}\log\left(1+\frac{16L'}{\epsilon}\sqrt{\frac{2}{\mu}\max_{\lambda\in\Lambda}\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta_0,z\right)\right]}\right)+\frac{8B^2}{\epsilon^2}\log\frac{2}{\delta}. \tag{13}$$

Theorem 5.4 characterizes the upper bound of the worst-case sample complexity of downstream tasks. If we regard ϵ and δ as constants, the worst-case sample complexity with respect to the initialization θ_0 is $O\left(\log\max_{\lambda\in\Lambda}\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta_0,z\right)\right]\right)$. Combined with (10), Theorem 5.4 demonstrates that the proposed minimax pre-training procedure implies tighter sample complexity than the average minimization pre-training procedure in the worst case. Though the dependency on the worst-case initial expected risk is logarithmic in the upper bound analysis, we find that the initialization can have an evident effect on the generalization of the downstream tasks in practice. We claim that the upper bound for general cases may not be tight for our deep learning applications. Special structures in applications might lead to tighter bounds for generalization errors, which remains for further study.

6 Conclusion

This paper introduces the concept of downstream-task robustness for pre-training, aiming to improve the performance of foundation models across various downstream tasks. As models such as ChatGPT become more prevalent, safety and consistent performance are increasingly important. Our proposed minimax loss for pre-training, validated through extensive experiments, offers a potential solution to enhance the robustness and safety of such models. In the future, we will explore grouping the upstream tasks adaptively. We would say though still in its early stages, the study of downstream-task robustness holds significant promise for the reliable and safe deployment of AI infrastructure.

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. arXiv preprint arXiv:2204.01678, 2022.

[2] Adil M Bagirov, Bülent Karasözen, and Meral Sezer. Discrete gradient method: derivative-free method for nonsmooth optimization. Journal of Optimization Theory and Applications, 137(2):317–334, 2008.

[3] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. Management Science, 59(2):341–357, 2013.

[4] Steven Bird, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. "O'Reilly Media, Inc.", 2009.

[5] Ali Borji. A categorical archive of chatgpt failures. arXiv preprint arXiv:2302.03494, 2023.

[6] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. Foundations and Trends® in Machine Learning, 8(3-4):231–357, 2015.

[7] Jacques Antonin Chatelon, Donald W Hearn, and Timothy J Lowe. A subgradient algorithm for certain minimax and minisum problems. Mathematical Programming, 15(1):130–145, 1978.

[8] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12299–12310, 2021.

[9] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In International conference on machine learning, pages 794–803. PMLR, 2018.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019.

[11] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In International Conference on Computer Vision, 2017.

[12] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In International Conference on Computer Vision, 2017.

[13] John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. Mathematics of Operations Research, 46(3):946–969, 2021.

[14] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. The Annals of Statistics, 49(3):1378–1406, 2021.

[15] Shang-Hua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. arXiv preprint arXiv:2106.03149, 2021.

[16] Warren Hare and Mason Macklem. Derivative-free optimization methods for finite minimax problems. Optimization Methods and Software, 28(2):300–312, 2013.

[17] 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.

[18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. Advances in neural information processing systems, 31, 2018.

[19] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7482–7491, 2018.

[20] Galina M Korpelevich. The extragradient method for finding saddle points and other problems. Matecon, 12:747–756, 1976.

[21] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. Advances in Neural Information Processing Systems, 33:8847–8860, 2020.

[22] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In International Conference on Machine Learning, pages 6083–6093. PMLR, 2020.

[23] Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In Conference on Learning Theory, pages 2738–2779. PMLR, 2020.

[24] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1871–1880, 2019.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. International Conference on Learning Representations, 2019.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26, 2013.

[27] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. Advances in neural information processing systems, 30, 2017.

[28] Yurii Nesterov and Laura Scrimali. Solving strongly monotone variational and quasi-variational inequalities. Discrete and Continuous Dynamical Systems, 31(4):1383–1396, 2011.

[29] OpenAI. Gpt-4 technical report, 2023.

[30] Oscar Oviedo-Trespalacios, Amy E Peden, Thomas Cole-Hunter, Arianna Costantini, Milad Haghani, Sage Kelly, Helma Torkamaan, Amina Tariq, James David Albert Newton, Timothy Gallagher, et al. The risks of using chatgpt to obtain common safety-related information and advice. Available at SSRN 4346827, 2023.

[31] Jason Phang, Thibault Févry, and Samuel R. Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. Computing Research Repository, 2018.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3):211–252, 2015.

[33] Soroosh Shafieezadeh Abadeh, Peyman M Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. Advances in Neural Information Processing Systems, 28, 2015.

[34] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In European Conference on Computer Vision, 2012.

[35] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. Journal of Computational and Applied Mathematics, 60(1-2):237–252, 1995.

[36] Paul Tseng. Accelerated proximal gradient methods for convex optimization. Technical report, Technical report, University of Washington, Seattle, 2008., 2008.

[37] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(7):3614–3633, 2021.

[38] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.