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

Anonymous Author(s)
Affiliation
Address
email

Contents

A Related Work

Minimax Optimization in Deep Learning. Minimax optimization has wide application in deep learning, e.g., adversarial robustness [6, 9], distributional robustness [12, 20, 27], adversarial generative models [15], imitation learning [35, 17]. In the field of pre-training, previous work has employed minimax optimization for adversarial robustness [8, 19]. Our work proposes a minimax optimization procedure for pre-training but with a different formulation. While the previous works aims at adversarial robustness, our work focus on the worst-case generalization of downstream tasks.

Masked language modeling. Masked language modeling (MLM) was first proposed by [28] as a Cloze task. Adapted as a novel pre-training task, MLM and its autoregressive counterpart ,*e.g.*, BERT [10], GPT [23, 24, 5], and T5 [25], are highly successful methods to deal with NLP problems. MLM first masks out some tokens from the input sentences and then trains the model to retrieve the missing context from the rest of the tokens. These methods have been demonstrated to scale excellently so that various downstream tasks can utilize the pre-trained representations. In particular,

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

BERT is constructed based on the transformer [30] model. After preparing the input samples, an embedding layer and a stack of Transformer layers are followed to conduct the bi-directional semantic modeling. We exploit BERT, the most typical masked language model, as the backbone model to process our ablation experiments.

Masked image modeling. Encouraged by transformers, which have gradually become a primary architecture for generic language understanding, ViT [11] later illusion the potential of adopting a pure transformer in image tasks. To generalize better for vision tasks and motivated by the success of BERT [10] in NLP, many recent works propose various masked image prediction methods for pre-training vision models in a self-supervised way. These methods reconstruct the target such as pixels [1, 7, 11, 13, 16, 33], discrete tokens [4, 34], and (deep) features [3, 32]. Notably, the masked autoencoder (MAE) [16] adopts an asymmetric design to allow the large encoder to operate only on unmasked patches and is followed by a lightweight decoder to reconstruct the complete signal from the latent representation along with mask tokens. MultiMAE [2] leverages the efficiency of the MAE approach and extends it to multi-modal and multitask settings. Based on MultiMAE, we apply our approach to increase the transfer capability for downstream tasks.

B Proofs

B.1 Convergence Rate of Algorithm 1

B.1.1 Proof of Theorem 3.1

For notation simplicity, we define that

$$\begin{aligned}
f_t(\theta) &:= \mathbb{E}_{z\sim P}\left[\ell_t\left(\theta, z\right)\right], \\
F_k(\theta) &:= \sum_{t=1}^{T} w_{\alpha,t}\left(\theta_k\right)\mathbb{E}_{z\sim P}\left[\ell_t\left(\theta, z\right)\right], \\
F(\theta) &:= \max_{t\in[T]} f_t(\theta),
\end{aligned}$$

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

Our proof consists of two parts. The first part is to show how well $F_k(\theta_k)$ approximates $F(\theta_k)$ with our choice of the softmax hyperparameter α. The second part is to analyze the dynamics of the algorithm and the total convergence rate.

We first show that $F_k(\theta_k) \geq F(\theta_k) + \frac{R_0 L'}{2\sqrt{k+1}}$ for $\alpha_k \geq \frac{4\sqrt{k+1}}{R_0 L'}\log\frac{4TB\sqrt{k+1}}{R_0 L'}$. Define $T_{k,\epsilon}(\theta) = \{t \in [T] \mid f_t\left(\theta_k\right) \geq F(\theta_k) - \epsilon\}$. If $\alpha_k \geq \frac{1}{\epsilon_k}\log\frac{TB}{\epsilon_k}$ for an $\epsilon_k > 0$, we have

$$\begin{aligned}
F_k(\theta_k) &= \sum_{t=1}^{T} w_{\alpha_k,t}\left(\theta_k\right) f_t(\theta_k) \\
&\geq \sum_{t\in T_{k,\epsilon_k}(\theta_k)} w_{\alpha_k,t}\left(\theta_k\right) f_t(\theta_k) \\
&\overset{(A)}{\geq} \frac{\sum_{t\in T_{k,\epsilon_k}(\theta_k)}\exp\left(\alpha_k f_t(\theta_k)\right)}{\sum_{t'=1}^{T}\exp\left(\alpha_k f_{t'}(\theta_k)\right)}\left(F(\theta_k) - \epsilon_k\right) \\
&= \left[1 + \frac{\sum_{t\notin T_{k,\epsilon_k}(\theta_k)}\exp\left(\alpha_k f_t(\theta_k)\right)}{\sum_{t'\in T_{k,\epsilon_k}(\theta_k)}\exp\left(\alpha_k f_{t'}(\theta_k)\right)}\right]^{-1}\left(F(\theta_k) - \epsilon_k\right) \\
&\overset{(B)}{\geq} \left[1 + \frac{T\exp\left(\alpha_k\left(F(\theta_k) - \epsilon_k\right)\right)}{\exp\left(\alpha_k F(\theta_k)\right)}\right]^{-1}\left(F(\theta_k) - \epsilon_k\right) \\
&= \left[1 + T\exp\left(-\alpha_k\epsilon_k\right)\right]^{-1}\left(F(\theta_k) - \epsilon_k\right) \\
&\overset{(C)}{\geq} F(\theta_k) - 2\epsilon_k.
\end{aligned}$$

The inequality A is due to the definition of $T_{k,\epsilon}(\theta)$. The inequality B is because $\sum_{t \notin T_{k,\epsilon_k}(\theta_k)} \exp\left(\alpha_k f_t(\theta_k)\right) \leq T \exp\left(\alpha_k \left(F(\theta_k) - \epsilon_k\right)\right)$ by the definition of $T_{k,\epsilon}(\theta)$ and there exists $t_k^* \in T_{k,\epsilon_k}(\theta)$ such that $f_{t_k^*} = F(\theta_k)$, which further implies $\sum_{t' \in T_{k,\epsilon_k}(\theta_k)} \exp\left(\alpha_k f_{t'}(\theta_k)\right) \geq \exp\left(\alpha_k F(\theta_k)\right)$. The inequality C is because the hyperparameter $\alpha_k \geq \frac{1}{\epsilon_k} \log \frac{TB}{\epsilon_k}$. Specifically, let $\epsilon_k = \frac{R_0 L'}{4\sqrt{k+1}}$ and $\alpha_k \geq \frac{4\sqrt{k+1}}{R_0 L'} \log \frac{4TB\sqrt{k+1}}{R_0 L'}$ correspondingly, we have $F_k(\theta_k) \geq F(\theta_k) + \frac{R_0 L'}{2\sqrt{k+1}}$ for all $k = 0, \ldots, K-1$.

We then analyze the dynamics of the algorithm and give the total convergence rate. For $k = 0, \ldots, K-1$, it holds that

$$
\begin{aligned}
F_k(\theta_k) &\overset{(A)}{\leq} F_k(\theta^*) + \langle \nabla F_k(\theta_k), \theta_k - \theta^* \rangle \\
&\overset{(B)}{=} F_k(\theta^*) + \frac{1}{2\eta}\left(\|\theta_{k+1} - \theta_k\|^2 + \|\theta_k - \theta^*\|^2 - \|\theta_{k+1} - \theta^*\|^2\right),
\end{aligned} \tag{1}
$$

where $\theta^* \in \arg\max_{\theta \in \Theta} F(\theta)$. The inequality A is due to the convexity of $F_k(\theta)$. The equality B is due to the update steps $\theta_{k+1} = \theta_k - \eta \nabla F_k(\theta_k)$ in Algorithm 1 and the fact $\langle a, b \rangle = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a - b\|^2\right)$.

Plugging the approximation error $F_k(\theta_k) \geq F(\theta_k) + \frac{R_0 L'}{2\sqrt{k+1}}$ and the inequality $F_k(\theta) \leq F(\theta)$ for $k = 0, \ldots, K-1$ and all $\theta \in \Theta$ into (1), we have

$$
F(\theta_k) \leq F(\theta^*) + \frac{1}{2\eta}\left(\|\theta_{k+1} - \theta_k\|^2 + \|\theta_k - \theta^*\|^2 - \|\theta_{k+1} - \theta^*\|^2\right) + \frac{R_0 L'}{2\sqrt{k+1}}. \tag{2}
$$

Taking the average over $k = 0, \ldots, K-1$, we further have

$$
\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1} F(\theta_k) &\leq F(\theta^*) + \frac{1}{2\eta}\left(\sum_{k=0}^{K-1}\|\theta_{k+1} - \theta_k\|^2 + \|\theta_0 - \theta^*\|^2 - \|\theta_K - \theta^*\|^2\right) + \frac{1}{K}\sum_{k=0}^{K-1}\frac{R_0 L'}{2\sqrt{k+1}} \\
&\leq F(\theta^*) + \frac{1}{2\eta}\left(\sum_{k=0}^{K-1}\|\theta_{k+1} - \theta_k\|^2 + \|\theta_0 - \theta^*\|^2\right) + \frac{1}{K}\sum_{k=0}^{K-1}\frac{R_0 L'}{2\sqrt{k+1}} \\
&\overset{(A)}{\leq} F(\theta^*) + \frac{1}{2\eta}\left(\sum_{k=0}^{K-1}\|\theta_{k+1} - \theta_k\|^2 + \|\theta_0 - \theta^*\|^2\right) + \frac{R_0 L'}{\sqrt{K}} \\
&\overset{(B)}{\leq} F(\theta^*) + \frac{K L'^2 \eta}{2} + \frac{R_0^2}{2\eta} + \frac{R_0 L'}{\sqrt{K}} \\
&\overset{(C)}{=} F(\theta^*) + \frac{2 R_0 L'}{\sqrt{K}}.
\end{aligned} \tag{3}
$$

The inequality A is due to the fact $\sum_{k=1}^{K} \frac{1}{\sqrt{k}} < 2\sqrt{K}$. The inequality B is because for $k = 0, \ldots, K-1$, the function $F_k(\theta)$ is L'-Lipschitz continuous, which implies $\|\theta_{k+1} - \theta_k\|^2 = \eta^2 \|\nabla F_k(\theta_k)\|^2 \leq \eta^2 L'^2$. The equality C is due to our choice for the step sizes, i.e., $\eta_k = \eta = \frac{R_0}{L'\sqrt{K}}$ for all $k = 0, \ldots, K-1$.

By the convexity of $F(\theta)$, we have $F(\bar{\theta}_K) \leq \frac{1}{K}\sum_{k=0}^{K-1} F(\theta_k)$. Combined with (3), we attain the desired result.

B.2 Analysis for the Minimax Pre-training Method

B.2.1 Proof of Proposition 5.1

By the assumption on the downstream task losses ℓ_λ and the task space Λ, the equation

$$
\max_{\lambda \in \Lambda} \mathbb{E}_{z \sim P}\left[\ell_\lambda(\theta, z)\right] = \max_{t \in [T]} \mathbb{E}_{z \sim P}\left[\ell_t(\theta, z)\right]
$$

holds for all $\theta \in \Theta$.

By the definition of θ^*_{max}, we have

$$\begin{aligned}\max_{\lambda\in\Lambda}\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta^*_{\text{max}}, z\right)\right] &= \max_{t\in[T]}\mathbb{E}_{z\sim P}\left[\ell_t\left(\theta^*_{\text{max}}, z\right)\right]\\ &\le \max_{t\in[T]}\mathbb{E}_{z\sim P}\left[\ell_t\left(\theta^*_{\text{average}}, z\right)\right]\\ &= \max_{\lambda\in\Lambda}\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta^*_{\text{average}}, z\right)\right],\end{aligned}$$

which is the result to prove.

B.2.2 Proof of Proposition 5.3

Proposition 5.3 is a standard result of gradient descent for strongly-convex and smooth functions. We include the proof here for completeness. By the convexity of $f(x)$ and the choice of the step size η we have

$$\begin{aligned}f(x_{k+1}) &\le f(x_k) + \langle\nabla f(x_k), x_{k+1}-x_k\rangle + \frac{L_f}{2}\|x_{k+1}-x_k\|^2\\ &= f(x_k) - \left(\frac{1}{\eta}-\frac{L}{2}\right)\|x_{k+1}-x_k\|^2\\ &\le f(x_k),\end{aligned}$$

which means the objective values are nonincreasing and implies $f(x_k)\le f(x_0)$ for all $k\in[K]$.

By the μ_f-strongly convexity at the point x^*, it holds for all $x\in\mathbb{R}^d$ that

$$\|x-x^*\|^2 \le \frac{2}{\mu_f}\left(f(x)-f(x^*)\right). \tag{4}$$

Combining (4) and the sequence $\{f(x_k)\}_{k=0}^K$ decreasing, we obtain

$$\|x_k-x^*\|^2 \le \frac{2}{\mu_f}\left(f(x_0)-f(x^*)\right), \text{ for all } k=0,1,\ldots,K-1,$$

as desired.

B.2.3 Proof of Theorem 5.4

The following proposition characterizes the sample complexity to find an ϵ-approximately optimal parameter by ERM for a downstream task λ within the parameter space $\Theta_\lambda(\theta_0)$. Theorem 5.4 follows directly from Proposition B.1 by considering the worst-case downstream task $\lambda\in\Lambda$. The remaining is to prove Proposition B.1.

Proposition B.1. *For a given task $\lambda\in\Lambda$ and a parameter space $\Theta_\lambda(\theta_0)$, let the parameter $\hat\theta^*_\lambda\in\Theta_\lambda(\theta_0)$ be the minimizer in of the empirical risk for N_λ i.i.d. samples $\{z_i\}_{i=1}^N$ from a distribution P, i.e., $\hat\theta^*_\lambda=\arg\min_{\theta\in\Theta_\lambda(\theta_0)}=\frac{1}{N_\lambda}\sum_{i=1}^{N_\lambda}\ell_\lambda(\theta,z_i)$. The parameter $\hat\theta^*_\lambda$ is ϵ-approximately optimal with probability at least $1-\delta$ if*

$$N_\lambda \ge \frac{8dB^2}{\epsilon^2}\log\left(1+\frac{16L'}{\epsilon}\sqrt{\frac{2}{\mu}\mathbb{E}^*}\right)+\frac{8B^2}{\epsilon^2}\log\frac{2}{\delta}, \tag{5}$$

where $\mathbb{E}^*=\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta_0,z\right)\right]$.

Denote $\mathbb{E}_{z\sim P}\left[\ell_\lambda\left(\theta,z\right)\right]$ as $f_\lambda(\theta)$ and $\frac{1}{N}\sum_{i=1}^N\ell_\lambda(\theta,z_i)$ as $\hat f_\lambda(\theta)$. First, we note

$$\begin{aligned}&\Pr\left(f_\lambda\left(\hat\theta^*_\lambda\right)-f_\lambda\left(\theta^*_\lambda\right)\ge\epsilon\right)\\ =&\Pr\left(\left[f_\lambda\left(\hat\theta^*_\lambda\right)-\hat f_\lambda\left(\hat\theta^*_\lambda\right)\right]+\left[\hat f_\lambda\left(\hat\theta^*_\lambda\right)-\hat f_\lambda\left(\theta^*_\lambda\right)\right]+\left[\hat f_\lambda\left(\theta^*_\lambda\right)-f_\lambda\left(\theta^*_\lambda\right)\right]\ge\epsilon\right)\\ \overset{(A)}{\le}&\Pr\left(\left[f_\lambda\left(\hat\theta^*_\lambda\right)-\hat f_\lambda\left(\hat\theta^*_\lambda\right)\right]+\left[\hat f_\lambda\left(\theta^*_\lambda\right)-f_\lambda\left(\theta^*_\lambda\right)\right]\ge\epsilon\right)\\ \le&\Pr\left(2\sup_{\theta\in\Theta_\lambda(\theta_0)}\left|f_\lambda\left(\theta\right)-\hat f_\lambda\left(\theta\right)\right|\ge\epsilon\right)\\ =&\Pr\left(\sup_{\theta\in\Theta_\lambda(\theta_0)}\left|f_\lambda\left(\theta\right)-\hat f_\lambda\left(\theta\right)\right|\ge\frac{\epsilon}{2}\right).\end{aligned}\tag{6}$$

The inequality A is because $\hat{f}_\lambda\left(\hat{\theta}^*_\lambda\right) - \hat{f}_\lambda\left(\theta^*_\lambda\right) \le 0$ by the definition of $\hat{\theta}^*_\lambda$.

We derive the upper bound by covering numbers. We only consider Euclidean space for simplicity.

Definition B.2 (Covering numbers [31, Chapter 4])**.** Consider a subset $S \subset \mathbb{R}^d$ and let $\epsilon > 0$. A subset $\mathcal{N} \subset S$ is called an ϵ-net of S if every point in S is within distance ϵ of some points of $\mathcal{N}$, i.e., for all $x \in S$, there exists $x_0 \in \mathcal{N}$ such that $\|x - x_0\| \le \epsilon$. The smallest possible cardinality of an ϵ-net of S is called the covering number of S and is denoted $C(S, \epsilon)$, i.e.,

$$C(S,\epsilon) := \min\left\{|\mathcal{N}| \mid \mathcal{N} \text{ is an } \epsilon\text{-net of } S\right\}.$$

Consider an ϵ'-net $\mathcal{N}(\Theta_\lambda(\theta_0), \epsilon')$ of $\Theta_\lambda(\theta_0)$ where $\epsilon' = \frac{\epsilon}{8L'}$. By the definition of ϵ'-net, we have

$$\sup_{\theta\in\Theta_\lambda(\theta_0)} \left|f_\lambda(\theta) - \hat{f}_\lambda(\theta)\right| \le \sup_{\theta\in\mathcal{N}(\Theta_\lambda(\theta_0),\epsilon')} \left|f_\lambda(\theta) - \hat{f}_\lambda(\theta)\right| + \frac{\epsilon}{4}. \tag{7}$$

Combining (6) and (7), we obtain

$$\begin{aligned}
&\Pr\left(f_\lambda\left(\hat{\theta}^*_\lambda\right) - f_\lambda\left(\theta^*_\lambda\right) \ge \epsilon\right) \\
\le &\Pr\left(\sup_{\theta\in\mathcal{N}(\Theta_\lambda(\theta_0),\epsilon')} \left|f_\lambda(\theta) - \hat{f}_\lambda(\theta)\right| \ge \frac{\epsilon}{4}\right) \\
\le &C\left(\Theta_\lambda(\theta_0),\epsilon'\right) \sup_{\theta\in\mathcal{N}(\Theta_\lambda(\theta_0),\epsilon')} \Pr\left(\left|f_\lambda(\theta) - \hat{f}_\lambda(\theta)\right| \ge \frac{\epsilon}{4}\right)
\end{aligned} \tag{8}$$

We leverage the upper bounds of covering numbers of balls [31, Chapter 4].

Lemma B.3. *The covering number of a ball of radius R, denoted as B_R, in $\mathbb{R}^d$ satisfies*

$$C(B_R,\epsilon) \le \left(\frac{2R}{\epsilon} + 1\right)^d.$$

By Lemma B.3, we have

$$C\left(\Theta_\lambda(\theta_0),\epsilon'\right) \le \left(\frac{2R_\lambda}{\epsilon'} + 1\right)^d, \tag{9}$$

where $R_\lambda = \sqrt{\frac{2}{\mu}\mathbb{E}_{z\sim P}\left[\ell_\lambda(\theta_0, z)\right]}$.

By Hoefferding's inequality, for each $\theta \in \mathcal{N}(\Theta_\lambda(\theta_0), \epsilon')$, we have

$$\Pr\left(\left|f_\lambda(\theta) - \hat{f}_\lambda(\theta)\right| \ge \frac{\epsilon}{4}\right) \le 2\exp\left(-\frac{n\epsilon^2}{8B^2}\right) \tag{10}$$

Plugging (9) and (10) into (8), we have

$$\Pr\left(f_\lambda\left(\hat{\theta}^*_\lambda\right) - f_\lambda\left(\theta^*_\lambda\right) \ge \epsilon\right) \le 2\left(\frac{2R_\lambda}{\epsilon'} + 1\right)^d \exp\left(-\frac{N_\lambda\epsilon^2}{8B^2}\right). \tag{11}$$

By (11), when the number of samples N_λ satisfies

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

we have $\Pr\left(f_\lambda\left(\hat{\theta}^*_\lambda\right) - f_\lambda\left(\theta^*_\lambda\right) \ge \epsilon\right) \le \delta$.

C Training details

C.1 Part-of-Speech Mask BERT Training Setting

In this work, we denote the number of layers (i.e., Transformer blocks) as L, the hidden size as H, and the number of self-attention heads as A. For comparison purposes, we primarily report results on two models with the same size: PoS-BERT$_{\text{BASE}}$ and BERT$_{\text{BASE}}$(L=12, H=768, A=12, Total Parameters=110M). The model is trained with AdamW [22] by setting β_1 = 0.9, β_2 = 0.999, ϵ = 1e-6, and L_2 weight decay of 0.01. The learning rate is warmed up over the first 10K steps to a peak value of 1e-4, then linearly decayed. We duplicate training data ten times to avoid using the same mask for each training instance in every epoch so that each sequence is masked in 10 different ways over the 40 training epochs. Thus, each training sequence was seen with the same mask four times. The hyperparameters for experiments are shown as Table 1 and Table 2.

Hyperparam	Part-of-Speech Mask BERT
Number of Layers	12
Hidden size	768
Attention heads	12
Attention heads size	64
Dropout	0.1
Warmup steps	10K
Peak Learning Rate	2e-4
Batch Size	256
Weight Decay	0.01
Max Steps	1000K
Learning Rate Decay	Linear
Adam ϵ	1e-6
Adam β_1	0.9
Adam β_2	0.999
Gradient Clipping	0.0

Table 1: Hyperparameters for pre-training Part-of-Speech Mask BERT.

Hyperparam	GLUE
Learning Rate	2e-5
Batch Size	32
Weight Decay	0.1
Learning Rate Decay	Linear
Warmup Ratio	0.06

Table 2: Hyperparameters for fine-tuning Part-of-Speech Mask BERT on GLUE.

C.2 Multi-Modal Mask MAE Training Setting

We use Vit-B [11] with a patch size of 16×16 pixels as the backbone for our MAE experiments, and estimate the model's performance under different pre-training epochs, i.e., 400 and 1,600 epochs on ImageNet1K and 800 epochs on ImageNetS50. We choose AdamW as the optimizer with a base learning rate of 1e-4 and weight decay of 0.05. We first warm up the learning rate with 40 epochs and then decay it with cosine decay [21]. We set the batch size to 2048 and trained the models using 8×A100 GPUs with automatic mixed precision enabled. Our data augmentations are straightforward. We randomly crop the images, setting the random scale between 0.2 and 1.0 and the random aspect ratio between 0.75 and 1.33. Afterward, we resize the crops to 224×224 pixels and apply a random horizontal flip with a probability of 0.5. The hyperparameters for experiments are shown as Table 3 and Table 4.

D Limitation

Contemporary pre-training models are consistently enlarging in size. However, due to limitations associated with computational power and the non-disclosure tendency of large-scale models, we were unable to conduct our experimentation directly on ultra-large models such as LLaMA [29], GPT3 [5], and V-MoE [26]. Notwithstanding, we have validated our hypothesis on two commonly encountered domains and model frameworks, thus illustrating the extensive applicability of our proposed methodology.

Hyperparam	**{None, GradNorm, DWA}**	**Uncertainty**	**Minimax**
Batch Size	2048	2048	2048
Learning Rate	8e-4	8e-4	8e-4
Min Learning Rate	1e-6	1e-6	1e-6
Weight Decay	0.05	0.05	0.05
Adamw ϵ	1e-8	1e-8	1e-8
Adamw β_1	0.9	0.9	0.9
Adamw β_2	0.95	0.95	0.95
Epoch	{800}	{400, 800, 1600}	{400, 800, 1600}
Warm up Epoch	40	40	40
Learning Rate Schedule	cosine decay	cosine decay	cosine decay
Non-masked tokens	98	98	98
Input resolution	224×224	224×224	224×224
Augmentation	RandomResizeCrop	RandomResizeCrop	RandomResizeCrop
Dropout	0.0	0.0	0.0
Patch Size	16	16	16

Table 3: Hyperparameters for pre-training Multi-Modal Mask MAE. We only pre-train 800 epochs on ImageNetS50, and pre-train both 400 and 1600 epochs on ImageNet1K.

Hyperparam	**Classification**		**Semantic Segmentation**		**Depth**
	ImageNet1K	**ImageNetS50**	**ImageNetS50**	**NYUv2**	**NYUv2**
Epoch	100	100	100	100	2000
Warm up Epoch	5	5	20	20	100
Batch Size	1024	1024	1024	1024	2048
Learning Rate	4e-3	4e-3	1e-4	1e-4	1e-4
Min Learning Rate	1e-6	1e-6	1e-6	1e-6	0
Weight Decay	0.05	0.05	0.05	0.05	1e-4
Adamw β_1	0.9	0.9	0.9	0.9	0.9
Adamw β_2	0.999	0.999	0.999	0.999	0.999
Layer Decay	0.65	0.65	0.75	0.75	0.75
Patch Size	16	16	16	16	16
Drop path	0.1	0.1	0.1	0.1	/
LR Schedule	cosine decay	cosine decay	cosine decay	cosine decay	cosine decay
Input resolution	224×224	224×224	224×224	224×224	256×256
Augmentation	Rand(9, 0.5)	Rand(9, 0.5)	LSJ	LSJ	LSJ

Table 4: Hyperparameters for fine-tuning Multi-Modal Mask MAE on various downtasks. The augmentation strategy LSJ is large scale jittering [14]. And we use drop path [18] in classification and semantic segmentation tasks.

References

[1] Sara Atito Ali Ahmed, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. Computing Research Repository, 2021.

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

[3] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. Computing Research Repository, abs/2202.03555, 2022.

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In International Conference on Learning Representations, 2021.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[6] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. Computing Research Repository, abs/1902.06705, 2019.

[7] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In International conference on machine learning, pages 1691–1703. PMLR, 2020.

[8] 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.

[9] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In International Conference on Machine Learning, pages 1310–1320. PMLR, 2019.

[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] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021.

[12] 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.

[13] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jégou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? Computing Research Repository, 2021.

[14] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2918–2928, 2021.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020.

[16] 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.

[17] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. Advances in Neural Information Processing Systems, 29, 2016.

[18] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In European conference on computer vision, pages 646–661. Springer, 2016.

[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.

[20] 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.

[21] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In International Conference on Learning Representations, 2017.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019.

[23] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. OpenAI, 2018.

[24] Alec Radford and Jeffrey Wu. Language models are unsupervised multitask learners. OpenAI, 1(8):9, 2019.

[25] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140):1–67, 2020.

[26] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. Advances in Neural Information Processing Systems, 34:8583–8595, 2021.

[27] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. International Conference on Learning Representations, 2020.

[28] Wilson L Taylor. “cloze procedure”: A new tool for measuring readability. Journalism quarterly, 30(4):415–433, 1953.

[29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[31] Roman Vershynin. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge university press, 2018.

[32] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14668–14678, 2022.

[33] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9653–9663, 2022.

[34] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In International Conference on Learning Representations, 2021.

[35] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 8, pages 1433–1438, 2008.