Towards Inadequately Pre-trained Models in Transfer Learning

Andong Deng[1,†], **Xingjian Li**[2,5,†], **Di Hu**[3,*], **Tianyang Wang**[4], **Haoyi Xiong**[2], **Cheng-Zhong Xu**[5]
[1]University of Central Florida [2]Baidu Research [3]Renmin University of China
[4]University of Alabama at Birmingham [5]University of Macau

Abstract

Transfer learning has been a popular learning paradigm in the deep learning era, especially in annotation-insufficient scenarios. Better ImageNet pre-trained models have been demonstrated, from the perspective of architecture, by previous research to have better transferability to downstream tasks[26]. However, in this paper, we find that during the same pre-training process, models at middle epochs, which are ***inadequately pre-trained****, can outperform fully trained models when used as feature extractors (FE), while the fine-tuning (FT) performance still grows with the source performance. This reveals that there is not a solid positive correlation between top-1 accuracy on ImageNet and the transferring result on target data. Based on the contradictory phenomenon between FE and FT that a better feature extractor fails to be fine-tuned better accordingly, we conduct comprehensive analyses on features before the softmax layer to provide insightful explanations. Our discoveries suggest that, during pre-training, models tend to first learn spectral components corresponding to large singular values and the residual components contribute more when fine-tuning.*

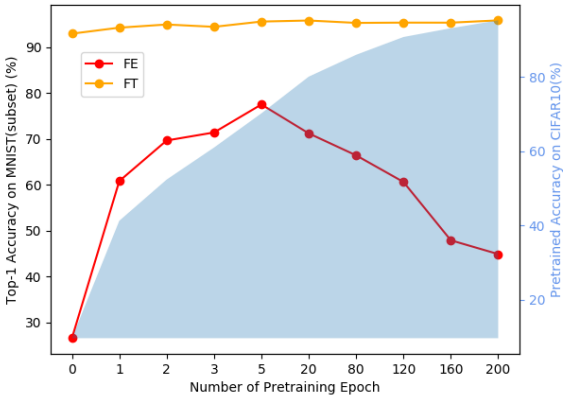

Figure 1. Toy experiment of transfer learning from a ResNet18[19] model pre-trained on CIFAR10[28] to a subset of MNIST[29]. FE means viewing the pre-trained model as a feature extractor, and FT means fine-tuning the whole model. It can be seen from the figure that the 5th-epoch model brings the best FE performance, which suggests that further pre-training on the source task would harm the feature quality for the target task. When fine-tuning the whole model, more adequate pre-training tends to deliver higher transfer learning performance.

1. Introduction

Deep learning has achieved tremendous success in modern computer vision with the aid of the strong supervision of well-labeled datasets, such as ImageNet[10]. However, data annotation is notoriously labor-extensive and time-consuming, especially in some specific domains where expertise is highly required. In such scenarios, transfer learning is of great interest for practitioners to train deep models with a small labeled dataset. Fortunately, existing efforts observe that when training on large-scale datasets, middle features of DNNs exhibit remarkable transferability to various downstream tasks [80, 75]. This facilitates popular deep transfer learning paradigms of fine-tuning a pre-trained model (FT) or simply employing the pre-trained model as a feature extractor (FE). With relatively sufficient labeled examples, fine-tuning the whole network usually achieves higher performance. Despite this, FE is still important when training resources are limited, or end-to-end training is not feasible. For example, some applications combine DNN features and other handcrafted features to obtain both accurate and explainable shallow classifiers [37, 55, 54].

Despite the ubiquitous utilization of pre-trained models, it still remains mysterious how such models benefit transfer learning. Several works pioneer to explore this plausible yet essential problem. [26] systematically investigates whether better-performing models on source tasks, e.g. ImageNet, necessarily yield better performances on downstream tasks. They confirm this hypothesis for both FE and FT, over deep architectures with different capacities. However, recent works in the domain of adversarial training discover that an

[†]Equal contribution. Work is partly done when Andong was an intern at Baidu Research. [*]Corresponding author.

adversarially pre-trained model, though performs worse on ImageNet due to additional adversarial regularizations, can still transfer better than its natural (following the naming practice in [63], referring to pre-training without adversarial methods) counterpart (with the same architecture)[57, 63]. In fact, these discoveries are to some extent in contradiction to the findings in [26], which argues that worse source models may transfer better.

Our work investigates the influence of pre-training on transfer effects from a different perspective. Specifically, we focus on the trajectory of the pre-training process, inspired by recent studies on the learning order of DNNs. Several works [2, 23, 35, 41] discover that DNNs tend to firstly learn simple and shallow features, e.g. colors and textures, which are regarded as more general and transferable across different data domains [80]. From the perspective of the frequency domain, such features lie in low-frequency spectrums. On the other hand, several other works reveal that high-frequency features obtained by a pre-trained model are likely to cause a negative transfer [9].

The aforementioned observations motivate a question that, does a fully pre-trained model definitely outperform its inadequately pre-trained version when transferring to target tasks (according to claims in [26]), or is there an intermediate pre-trained checkpoint that yields a better transfer effect than that of the fully pre-trained version? To our best knowledge, very little work manages to explore how the transferability of a model is impacted by the different stages in a pre-training process.

To investigate this question, we run a toy experiment using CIFAR10 as the source dataset and a subset of MNIST (we randomly choose 100 data points for each digit from the official training split, resulting in a 1000-sample training set) as the target. Briefly, we train a ResNet-18[19] on CIFAR10 for 200 epochs and choose a set of checkpoints to run transfer learning in two different settings. In one setting, we treat the pre-trained model as a feature extractor (FE) and only retrain a softmax classifier, while in the other we fine-tune the whole model (FT). The retraining or fine-tuning continues for 100 epochs on the target dataset. As shown in Figure 1, the best performance of FE comes from the early 5th-epoch model, while the FT performance is higher for later checkpoints.

Two counter-intuitive facts can be observed from our results. One is that, a pre-trained model with higher accuracy on the source task is not necessarily better on the target task, especially when used as a feature extractor (FE). Among the checkpoints on the pre-training trajectory, there is no positive correlation between the source and target accuracy. The other observation shows inconsistent behaviors between FE and FT, indicating that a good starting point (FE) does not guarantee a good final result (FT). In order to explain the observed phenomenons, we investigate the spectral components of deep features before the FC layer (in Section 4.4), and observe that different parts of components contribute diversely for different pre-trained checkpoints within the same pre-training process.

In this paper, we conduct extensive transfer learning experiments, including ImageNet and the other 9 benchmark datasets. The results suggest that, when retraining a new classifier on top of the features extracted from pre-trained models, inadequately pre-trained ImageNet models yield significantly better performance than that of the standard 90-epoch pre-trained version, but the performance still highly correlates with the source performance when fine-tuning. Further, we present insightful analyses to explain such a difference from the perspective of spectral components of the extracted features and find that there are specific components corresponding to pre-trained models at different pre-training stages. In summary, our main contributions are as follows:

- Our work is the first to investigate how **different checkpoints** in the same pre-training process perform on transfer learning tasks. This contributes to a broader and deeper understanding of the transferability of neural networks.
- We discover that in the same pre-training process, an **inadquately pre-trained model** tends to transfer better than its fully pre-trained counterpart, especially when the pre-trained model is used as a frozen feature extractor. We also further experimentally consolidate this claim beyond image classification.
- We observe that FT prefers later pre-training checkpoints, compared with FE. Our analyses based on spectrum decomposition indicate that the learning order of different feature components leads to different preferences of pre-trained checkpoints between FE and FT.
- We also point out the risk of utilizing transferability assessment approaches as a general tool to select pre-trained models. We evaluate LogME [76], LEEP[44] and NCE[62], which are dependent on frozen pre-trained models. Aiming to select the best pre-trained model among different checkpoints, scores obtained by these algorithms often show poor correlations with the actual fine-tuning performance.

2. Related Work

Pre-training on large datasets, such as ImageNet[10], has long been a common method for transfer learning in various kinds of downstream tasks. Due to the huge effort brought by data annotation, researchers have reached a consensus that supervised or unsupervised pre-training as a parameter initialization or even an important medium for representation learning on existing large datasets is beneficial[12, 22]

for general downstream tasks[46, 43, 47, 60, 17, 6, 5, 16, 8, 79] or specific ones[68, 67, 72, 32]. Zeiler et al. [80] have found that retraining a softmax classifier on top of a fixed pre-trained feature would benefit the classification of target data by a large margin compared with training from scratch. In recent years, designing different kinds of pretext tasks (e.g. jigsaw puzzle[46], rotation angle prediction[14], temporal order prediction[43]) as a self-supervised pre-training method became a popular trend in this community. Later on, contrastive learning[47, 60, 17, 6] has also been demonstrated as a better self-supervised pre-training approach. Beyond the vision domain, large-scale unsupervised pre-training in speech and natural language[58, 11, 73, 52, 4] is appealing as well. Furthermore, learning universal representation and capturing cross-modal correspondence by pre-training in a multimodality setting[51, 38, 31, 30] and its downstream applications[82, 77] play an important role in the development of artificial general intelligence[15].

With such a powerful impact on deep learning, in the computer vision community, researchers have also been trying to understand the mechanism behind the success of pre-training, especially the ImageNet case, since ImageNet indeed has strong transferring power even to different data domains (e.g., in geoscience[42] and biomedical science[53]). Erhan et al.[13, 12] experimentally validated the role of unsupervised pre-training as a regularizer for the following supervised learning. Huh et al. [21], via designing thorough experiments, answer a series of questions about the performance difference of transferring brought by different aspects (e.g., number of training samples, number of training classes, fine-grained or coarse-grained pre-training, etc.) of ImageNet. Using the proper normalization method and extending the training time, He et al.[18] challenge this well-established paradigm and argue that it is possible to obtain better performance on target data from random initialization in detection and segmentation tasks. Following this work, Zoph et al. [83] further point out that self-training, with stronger data augmentation, can also lead to better transferring performance than pre-training. Nonetheless, pre-training is also viewed as a helpful training fashion for downstream tasks from different perspectives. Hendrycks et al.[20] have discovered that, in task-specific methods (e.g., label corruption, class imbalance, adversarial examples, etc.), pre-training enhances model robustness and brings consistent improvement compared with regular approaches.

Aiming to investigate what kinds of pre-training models could bring better transferring performance, Kornblith et al.[26] conduct extensive experiments on 16 different network architectures and suggest that models with higher top-1 accuracy on ImageNet could learn better transferable representations for target tasks. From the perspective of adversarial training, Utrera et al.[63] found that adversarially-trained models, though perform poorer on source data, actually have stronger transferability than natural models. And they further claimed that adversarially-trained models can learn more human-identifiable semantic information. Later, focusing more on model architecture, Salman et al.[57] drew the same conclusion, which further consolidates this viewpoint. In this work, we further investigate the relationship between top-1 accuracy on ImageNet and the transfer performance and found that some suboptimal models during pre-training transfer better when viewed as feature extractors, which is an analogous phenomenon with early-stopping[48, 74] in supervised learning that higher accuracy on training set does not mean higher test performance.

In order to further boost the performance of transfer learning, in several previous publications[71, 34, 9], new regularizers have been comprehensively investigated w.r.t. both model parameters and features. In [71], the convolutional weights are penalized to be closer to the source parameters rather than zero to avoid information loss from source data. Li et al. [34] utilize an attention mechanism to restrict the difference between the convolutional features at the same hierarchy from the source model and target one, respectively. Further, Chen et al.[9] claim that feature components corresponding to small singular values would be an impediment to knowledge transferring and then propose to suppress such components as regularization during fine-tuning. In this work, we also take advantage of Singular Value Decomposition on the features before the softmax layer and provide empirical analysis of the learning mechanism during the learning process.

3. Experimental Setup

We conduct extensive experiments on 8 representative natural image classification datasets (CIFAR10[28], CIFAR100[28], Food-101[3], FGVC Aircraft[40], Stanford Cars[27], CUB-200-2011[65], Oxford 102 Flowers[45] and MIT Indoor 67[50]) and one medical dataset (MURA [56]) based on standard pre-training on both ResNet50 [19] (90 epochs) and T2T-ViT_t-14 [78] (300 epochs), which are representative architectures for ConvNets and Transformers in image classification. The top-1 accuracies are 76.06% and 81.55% for ResNet50 and T2T-ViT_t-14, respectively. For pre-training details, we follow the standard ImageNet training configuration and the official T2T-ViT implementation for the two models, respectively.

4. Results and Analyses

In this section, we showcase all the experimental results of the transfer learning in two different settings: 1. Utilizing the pre-trained models as a feature extractor (FE) and re-training a softmax classifier; 2. Fine-tuning (FT) the whole model. We present experimental results of FE and FT in Section 4.1 and 4.2 respectively. A key observation is that

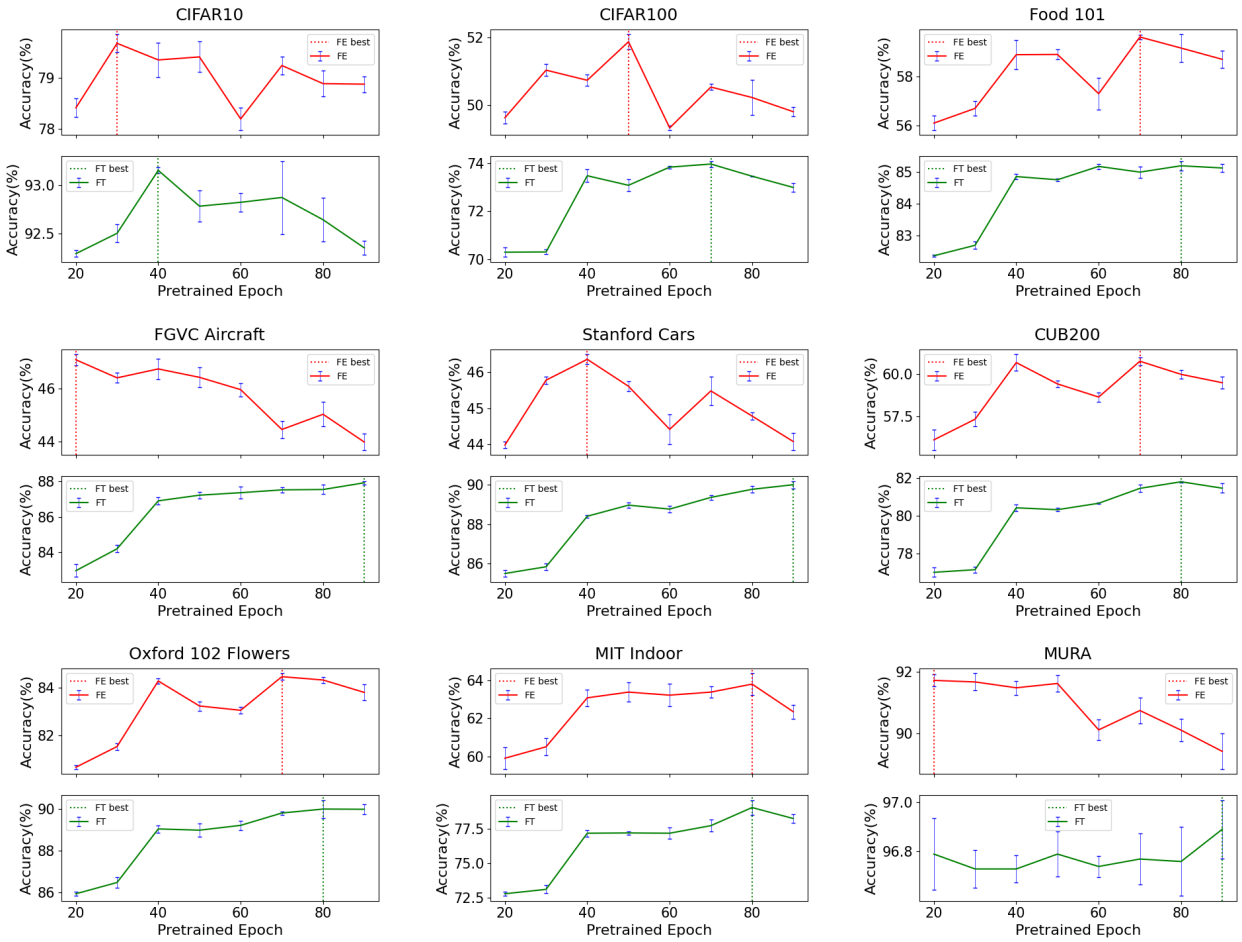

Figure 2. Transfer learning performance on selected datasets. We can observe obvious different trends w.r.t. pre-trained epoch for FE and FT. The FT generally grows with the pre-training epochs increasing, while FE regularly reaches the peak at a middle epoch.

inadequately pre-trained checkpoints transfer better for FE. Besides, we find that a better FE, which can be viewed as a better initialization for the target model, does not yield a better fine-tuning result. This is confirmed in Section 4.3 by t-SNE[64] visualization of deep features before the classifier. In Section 4.4, we manage to discover the in-depth learning mechanism during fine-tuning and empirically explain the aforementioned paradox, with the help of spectral components analysis.

4.1. Inadequately Pre-training Brings Better Feature Extractors

Concretely, in Figure 2 and Figure 3, we can easily observe that there exists a best transferring spot before the model is fully pre-trained when viewed as a feature extractor for different datasets, which means that the correlation between the accuracy of the pre-trained model and the quality of the feature of the penultimate layer is not as positive as claimed in [26]. We can also notice that the general trace of the FE performance is roughly a U-form curve with respect to the source performance, implying a potential trade-off between multiple factors during the pre-training process. Some curves exhibit a form of double-U, e.g. Stanford Cars and CUB-200-2011, and the FE performance at pre-trained epoch 40 and 70 is more likely to increase. We suspect this phenomenon may relate to the learning rate annealing after the 30-th and 60-th pre-training epoch [35]. In 4.6, we will showcase some scenarios where inadequately pre-training brings advantages.

4.2. Fully Pre-training Brings Better Fine-tuning Performance

The case for FT is quite different compared with FE. The general evolution trend for fine-tuning is still positively correlated with the source performance, though the fully pre-trained checkpoint is not always the best. And we can also find that the best FT model emerges later than the best FE model. This asynchronization is actually surprising because

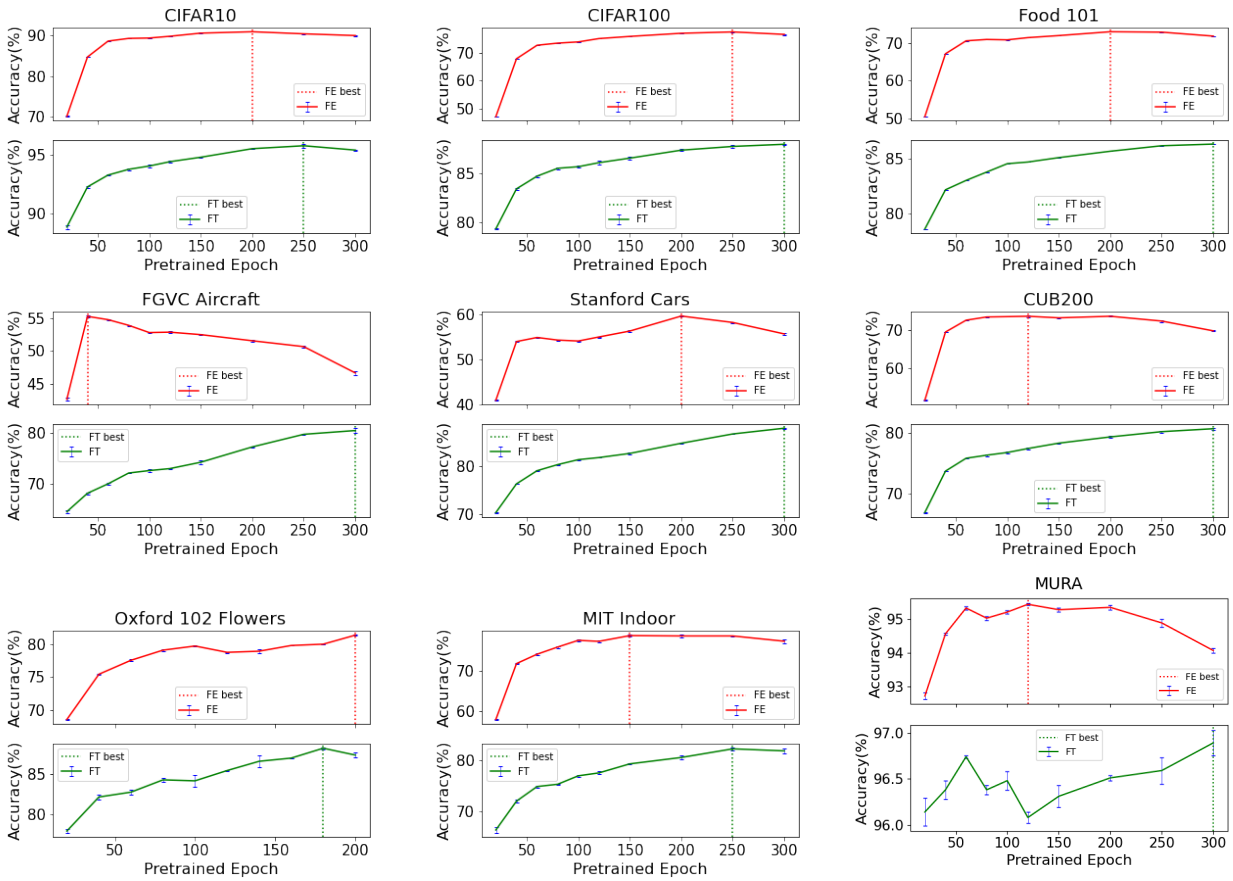

Figure 3. Transfer learning performance on selected datasets on T2T-ViT [78]. The trends are similar to those on ResNet50.

it is common sense that a better initialization should bring better fine-tuning results. However, our work is not the only one that challenges this intuition. Recent empirical studies [81, 33] propose to improve fine-tuning by re-initializing top layers, i.e. employing a worse feature extractor as the starting point of fine-tuning.

4.3. Visualization Analysis

In this subsection, we empirically visualize deep features of the best FE model (at the 5th epoch) and the fully pre-trained one (at the 200th epoch). The model is pre-trained on CIFAR10 and then transferred to MNIST, by both FE and FT. Deep features on the last convolutional layer of ResNet-18, produced by MNIST images, are extracted and dimensionally reduced to a 2-d space with t-SNE[64]. The FE performances are 96.47% and 88.47%, and the FT performances are 99.30% and 99.46%(in this experiment we use the full version of MNIST). As can be seen from the top two plots in Figure 4, the visualization result is consistent with the transferring performance. When directly using the pre-trained model to extract features, data points in the embedding space of the 5-epoch model are clustered better, especially for categories corresponding to index 1 and 6; while the fully pre-trained model produces a more chaotic feature distribution that many data points are entangled with their incongruent neighbors. However, the situation becomes reversely when the whole model is fine-tuned. There exist a couple of misclassified data points in the feature space of the 5-epoch model, while the fully pre-trained model provides highly tight and discriminative features. This phenomenon is somehow surprising because this indicates that a better initialization, i.e., more discriminative features, might lead to worse fine-tuning performance.

4.4. Spectral Component Analysis

Based on the observations from Figure 2 and Figure 3, two questions naturally arise: **What makes an inadequately pre-trained model a better feature extractor? What makes a better initialization (FE) perform worse than a fully pre-trained model which could not produce more discriminative features at the beginning?** To answer these questions, we resort to spectral analysis by Singular Value Decomposition (SVD) for an in-depth investigation. Specifically, we first obtain the batched feature matrix before the classification layer, which we denote as $\boldsymbol{F} \in R^{b \times d}$, where b is batch size and d is feature dimen-

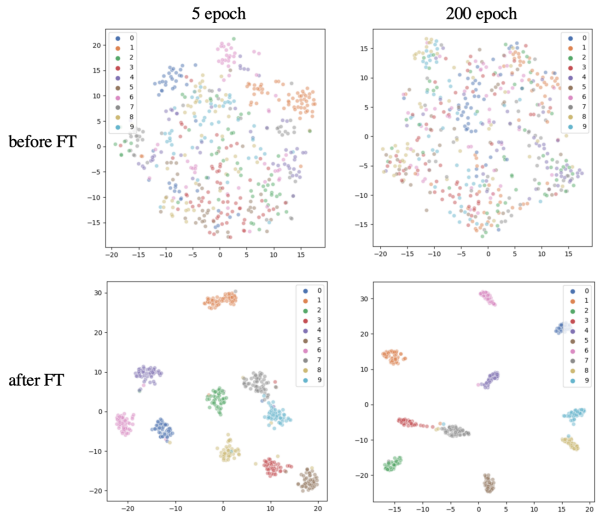

Figure 4. T-SNE visualization of features extracted before the classifier, before and after fine-tuning. Models are pre-trained on CIFAR10 and transferred to MNIST. The 5-epoch pre-trained model provides a better feature distribution on MNIST than the fully pre-trained one, but after fine-tuning for 50 epochs, the fully pre-trained model surpasses the 5-epoch one. Best viewed in color.

sion. After this, we decompose the matrix using SVD as:

$$\boldsymbol{F} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}}, \tag{1}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are left and right singular vectors respectively, and $\boldsymbol{\Sigma}$ is a rectangular diagonal matrix with the singular values on the diagonal. For convenience, we assume that all singular values are sorted in descending order.

Then we divide the diagonal matrix $\boldsymbol{\Sigma}$ as the main matrix $\boldsymbol{\Sigma_m}$ and the residual matrix $\boldsymbol{\Sigma_r}$. To achieve this division, we first calculate the sum over all singular values as $S_\sigma^K = \sum_{i=1}^{K} \sigma_i$, and then determine the minimum k that satisfies $S_\sigma^k / S_\sigma^K \geq 0.8$. $\boldsymbol{\Sigma_m}$ preserves top k lines of $\boldsymbol{\Sigma}$ and fills the remaining elements with zero. $\boldsymbol{\Sigma_r}$ is then obtained by $\boldsymbol{\Sigma_r} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma_m}$. In this way, we can get two spectral components $\boldsymbol{F_m}$ and $\boldsymbol{F_r}$ of the original $\boldsymbol{F}$ by truncated SVD reconstruction as

$$\boldsymbol{F_m} = \boldsymbol{U}\boldsymbol{\Sigma_m}\boldsymbol{V}^{\boldsymbol{T}} \tag{2}$$

and

$$\boldsymbol{F_r} = \boldsymbol{U}\boldsymbol{\Sigma_r}\boldsymbol{V}^{\boldsymbol{T}}. \tag{3}$$

According to [9], $\boldsymbol{F_m}$, as the main components of the feature matrix, represent the majority of transferring knowledge of the extracted features, while $\boldsymbol{F_r}$ is untransferable components or is hard to transfer that may do harm to the learning process and further causes negative transfer[66]. To evaluate the two components, we retrain a softmax classifier with Gaussian initialization on top of $\boldsymbol{F_m}$ and $\boldsymbol{F_r}$ for 50 epochs. We set the batch size as 128, using Adam[25] optimizer, and the learning rate as 0.01.

Table 1. Results of Spectral Component Analysis for the best FE models and fully pre-trained one. SE denotes the pre-training epoch on CIFAR10, and SA means the pre-training accuracy. FE means viewing the pre-trained model as a feature extractor and only retraining a softmax classifier; FT means fine-tuning the whole model. The 96.47% means the MNIST accuracy of the 5-epoch model in the FE task, and the 88.24% means the classification accuracy on top of $\boldsymbol{F_m}$. We can observe that $\boldsymbol{F_m}$ and $\boldsymbol{F_r}$ perform differently no matter whether trained with more source information (pre-training) or more target one (fine-tuning):

Task \ SE(SA)	5 epochs (70.24%)			200 epochs (95.32%)		
FE	96.47%	$\boldsymbol{F_m}$ $\boldsymbol{F_r}$	88.24% 55.45%	88.47%	$\boldsymbol{F_m}$ $\boldsymbol{F_r}$	58.74% 71.28%
FT	99.30%	$\boldsymbol{F_m}$ $\boldsymbol{F_r}$	99.26% 27.77%	99.46%	$\boldsymbol{F_m}$ $\boldsymbol{F_r}$	99.08% 54.69%

For comparison, we choose the best FE model and the fully pre-trained model in this experiment. For convenience, we call the feature from the best FE model as BFE feature and the feature from the fully pre-trained model as FP feature. The first model pairs are from the CIFAR10-to-MNIST experiment, and the results are shown in Table 1. Since we only analyze the features before the softmax layer, the FE models are actually identical to the corresponding pre-trained models; the FT models are fine-tuned with MNIST for 50 epochs. The best FE model is the 5-epoch pre-trained model, whose accuracy is 96.47% and is 8% higher than the fully pre-trained one; however, after fine-tuning, the fully pre-trained model outperforms the 5-epoch one, even with less discriminative initial features. Thus, we decompose the BFE feature and FP feature to investigate which part of the components contributes to their higher performance in FE and FT, respectively.

As can be seen from Table 1, there are several interesting discoveries as followed.

- *The quality of $\boldsymbol{F_m}$ is responsible for the FE performance, while $\boldsymbol{F_r}$ is dominant when fine-tuning the whole model.* Specifically, we find that the 5-epoch model performs better as FE due to its remarkable superiority in $\boldsymbol{F_m}$. However, in the FT setting, the 5-epoch and 200-epoch models show similar performances in $\boldsymbol{F_m}$, and the higher $\boldsymbol{F_r}$ results in the higher overall performance of the 200-epoch model.

- *As pre-training fits source data, $\boldsymbol{F_m}$ becomes less discriminative on target data, but $\boldsymbol{F_r}$ transfers better* (observed from the line of FE in Table 1). The degeneration in transferability of $\boldsymbol{F_m}$ could be caused by domain discrepancy between source and target data, as fully fitting source data may convert general patterns to those specific to the source domain. On the contrary, since $\boldsymbol{F_r}$ can not be well learned at earlier pre-training

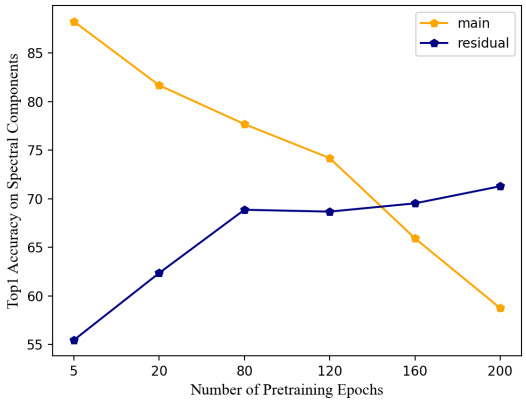

Figure 5. Evolution of the spectral components in pre-training from 5 epoch to 200 epoch. The orange curve represents the main components $\boldsymbol{F_m}$, and the blue one represents the residual $\boldsymbol{F_r}$. It is noticeable that $\boldsymbol{F_m}$ becomes less discriminative when the pre-training epoch grows.

stages, it generally becomes more informative by further pre-training.

- *For FT, $\boldsymbol{F_m}$ is easily adapted to target data, but $\boldsymbol{F_r}$ becomes less discriminative on target data* (observed from each column in Table 1). Both the 5-epoch and 200-epoch models achieve very high $\boldsymbol{F_m}$ performance (99.26% and 99.08%) after fine-tuning. This implies the underlying learning mechanism that DNNs prefer a prior fitting with main spectrums rather than residual spectrums. The performance of $\boldsymbol{F_r}$ (from FE to FT) decreases due to the information capacity w.r.t. entire $\boldsymbol{F}$ is constant. Despite the degeneration, better $\boldsymbol{F_r}$ in FE still delivers better $\boldsymbol{F_r}$ after fine-tuning, indicating that the residual components learned from the source are not completely forgotten after fine-tuning on target.

There might exist another explanation for the phenomenon in this spectral components analysis, which is from the perspective of the frequency domain. We can view $\boldsymbol{F_m}$ as low-frequency components of the original $\boldsymbol{F}$, and $\boldsymbol{F_r}$ as the high-frequency one. A couple of previous publications have revealed that the neural networks are inclined to learn low-frequency information first in the training process[69, 70, 39]. In our case, during pre-training, the model rapidly learns low-frequency knowledge at the early 5 epochs, which makes it the best feature extractor for downstream tasks. When keep learning in the source domain, more high-frequency patterns, which are specific to the source domain, are gradually learned; therefore, the negative transfer happens.

We also illustrate the evolution of the classification performance of the two components for different pre-training epochs in the FE task (from CIFAR10 to MNIST) in Figure 5. It can be obviously noticed that $\boldsymbol{F_m}$ and $\boldsymbol{F_r}$ shows exactly opposite trends when pre-training epoch increases. With longer pre-training on CIFAR10, $\boldsymbol{F_m}$ becomes less discriminative since the model is prone to a deeper fitting to CIFAR10 with more high-frequency knowledge learned. Inversely, the residual components $\boldsymbol{F_r}$ becomes more informative for target data when memorizing more knowledge from the source domain.

4.5. Rethink Transferability Assessment Tools

In this subsection, we utilize several transferability assessment tools LogME[76], LEEP[44], and NCE[62] to validate whether it is possible to obtain the best checkpoint during pre-training without any training. We report these scores for three datasets at different pre-trained checkpoints and calculate the correlation coefficient and Kendall's τ coefficient[24] for FT performance. As can be seen in Figure 6, LogME shows good ability in selecting the best checkpoint on FGVC Aircraft and Flowers102, but a little bit poorer in CIFAR10; while the LEEP and NCE can hardly capture the correlation between the performance and the scores, especially in Flowers102.

4.6. Application of Inadequate Pre-training

Pre-trained backbone networks are frequently used as feature extractors in downstream tasks, e.g., image captioning[1], image retrieval[59], temporal action localization[61], few-shot image classification[49], etc. In this section, we leverage typical downstream tasks to validate the effectiveness of inadequately pre-trained models, demonstrating the universal advantages of our method in various scenarios.

We firstly focus on the image-text retrieval problem on the MSCOCO dataset[36] and choose the recent PVSE[59] to incorporate our pre-trained ResNet50 models for evaluation. As shown in Table 2, we obtain the best retrieval performance with the 70-epoch pre-trained ResNet50, which extends our conclusion beyond image classification. We also provide a simple yet effective method in this case for selecting checkpoints. We use 25% of the training data and the validation set to evaluate the models and select models according to rsum, which is the summation of the recall scores. The results in Table 3 are consistent with that obtained when the full data is in use, demonstrating the rationale and efficacy of such a method. Moreover, to further consolidate our claim, we use our pre-trained ResNet50 models to perform a few-shot image classification task. Specifically, we choose weight imprinting[49] on CUB-200-2011 in our experiment. In this method, the pre-trained ResNet50 models are firstly tuned on a subset of 100 classes, and then the classification weights are imprinted to fit unobserved classes. We directly take the accuracy of the 100-class subset as an indicator for model selection. In Ta-

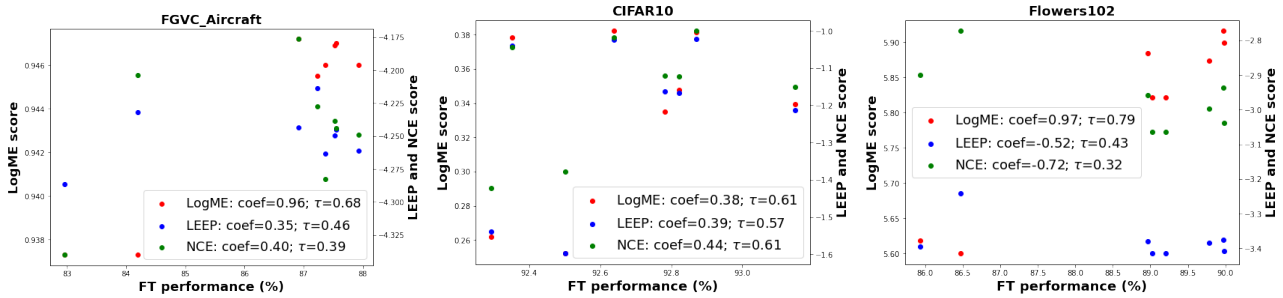

Figure 6. Transferability assessment for FGVC Aircraft, CIFAR10 and Flowers102 in and FT performance.

ble 4, we observe that inadequately pre-trained models are still reliable feature extractors for this task. More importantly, it demonstrates that using the 100-class accuracy for model selection can also lead to the best 70-epoch model.

Table 2. Performance comparison of PVSE for different pre-training epochs. The 70-epoch pre-trained model obtain the best performance, which is consistent with our conclusion that inadequately pre-trained models could extract better visual features.

ep	Image-to-Text			Text-to-Image			rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
50	9.88	28.28	41.26	3.84	12.58	19.88	114.98
60	10.64	28.78	41.30	3.89	13.26	20.41	117.36
70	**12.28**	**33.46**	**45.82**	**5.20**	**16.46**	**25.44**	**138.66**
80	11.04	30.12	42.00	4.84	15.39	23.67	127.45
90	10.82	30.70	44.50	4.58	15.18	23.63	132.67

Table 3. The best rsum score exists in the 70-epoch models.

ep	Image-to-Text			Text-to-Image			rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
50	9.28	26.64	28.82	3.84	12.44	19.80	110.82
60	9.64	28.26	39.88	3.82	12.91	20.410	114.91
70	**12.22**	**31.38**	43.72	**4.86**	15.60	24.22	**132.00**
80	11.60	31.38	**43.98**	3.54	12.06	19.52	121.87
90	10.64	29.94	42.32	4.82	**16.36**	**25.06**	129.13

Table 4. Results of a classification task for different pre-training epochs. The table shows the performance for N-shot unseen classification and the accuracy on 100 seen classes.

ep	N=1	N=2	N=5	N=10	N=20	100-class
50	52.07	55.64	60.77	63.95	65.64	81.28
60	51.28	55.68	61.11	63.86	65.95	82.16
70	**52.92**	**57.44**	**62.20**	**65.08**	67.05	**82.82**
80	52.36	56.44	61.77	64.58	66.66	82.68
90	52.11	57.34	61.49	64.83	**67.17**	82.47

5. Discussion and Future Work

Better performance in source tasks has long been believed to be more beneficial in target tasks. However, in this paper, we find that when using pre-trained models as feature extractors and retraining a new softmax classifier, the transferring performance does not agree with the source accuracy. There always exists the best epoch in the pre-training process. Intuitively, this is possibly brought by the distribution gap between the source and target data, forming a trade-off between source and target knowledge. If pre-training is less, no sufficient (general) visual knowledge can be obtained and the feature is suboptimal, but negative transfer happens the other way around. Based on this observation, we can operate a more sophisticated checkpoint selection process when we need a good feature extractor trained from source data[37].

Moreover, the common sense that better initialization should bring better training results is challenged given our observations. As can be seen from the difference in the evolution along the pre-training epochs between FE (view pre-trained model as a feature extractor) and FT (fine-tune the whole model), the FT performance still has a high correlation with the source performance, regardless of the U-property of FE performance. This means that a better feature extractor, which can be viewed as a better model initialization, does not definitely brings a better fine-tuning result. Further, in order to provide a more insightful explanation, we conduct a comparative experiment between the best FE model and the fully pre-trained one. Specifically, we delve into the spectral components of the feature before the classification layer and find that the components from top singular values contribute most to the FE, while the components with small singular values play a more critical role in the FT performance. In previous research[9], spectral components corresponding to small singular values are criticized as hard to transfer or even untransferrable. Concretely, we reach consistent conclusions but take different operations. Unlike [9], we do not drop the residual component, but investigate its discriminativeness along with the main component. In this way, we empirically reveal the reason behind the paradox phenomenon that a better feature extractor fails to produce better fine-tuning results in the end. Consistent with an intuitive assumption that over-pre-training would undermine the performance of the pre-trained model as a feature extractor, we discover the main component of the

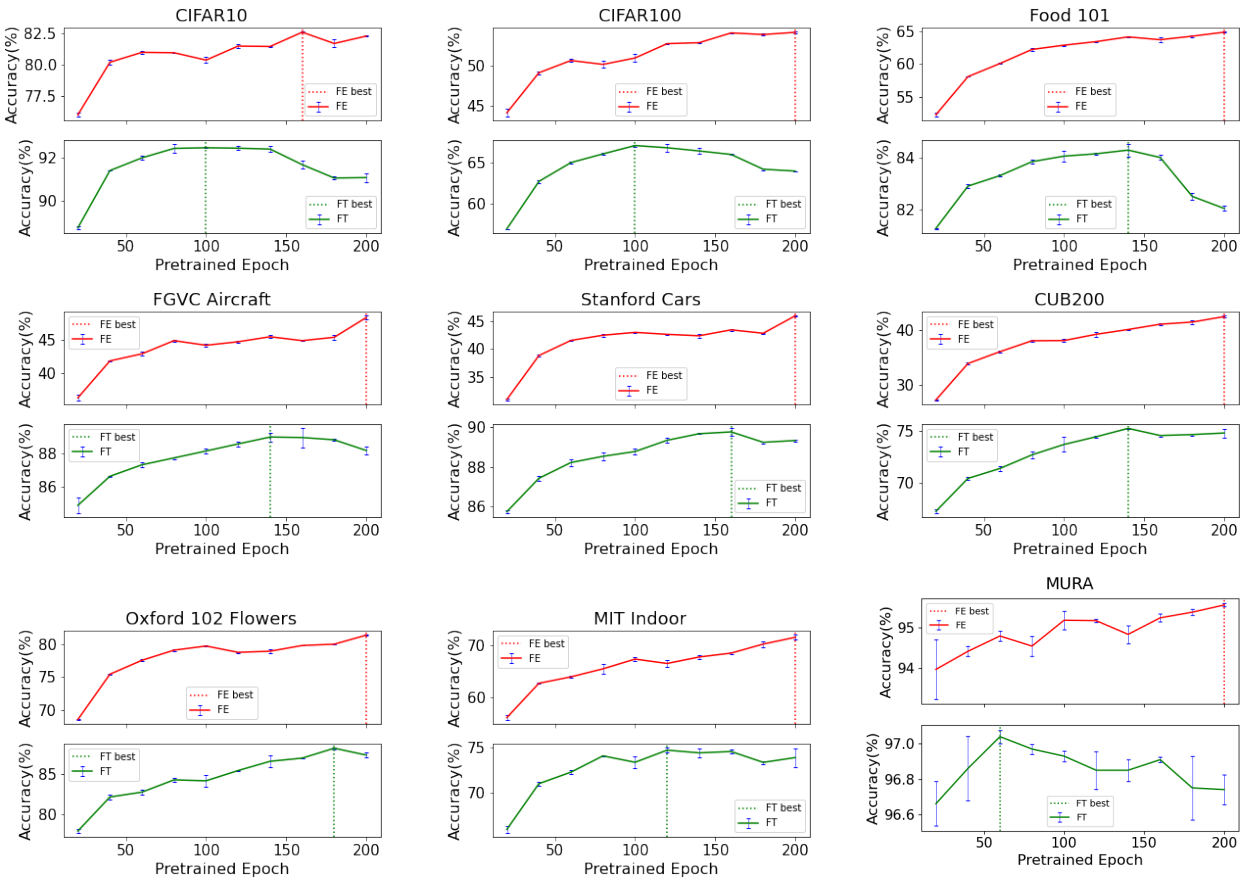

Figure 7. Transfer learning performance on selected datasets with Moco-v2 [7]. Unlike supervised pre-training, the FE performance on self-supervised pre-training can be enhanced with more pre-training epochs.

target feature becomes impaired due to the model overfitting to source data with the pre-training epoch increasing.

From a different perspective, we regard the main components as containing low-frequency knowledge of the feature, and the residual components as the carrier of high-frequency information. This makes sense since the residual components are generated from smaller 20% singular values, which are of high variation. In this way, our discoveries are also consistent with what has been well studied in the training mechanism of deep neural networks that the deep models learn low-frequency components before capturing high-frequency ones[69, 70].

However, there are still some phenomena beyond our explanation in Table 1. For example, since the performance of $\boldsymbol{F_m}$ decreases with more pre-training (from 88.24% to 58.74%), what makes it grow much faster (40.34% vs. 11.02%), though the accuracy is a little bit lower (99.08% vs. 99.30%) when trained with target data? It is attractive to keep investigating the correspondence between different spectral components and different learning stages (e.g., early or late in pre-training, pre-training, or fine-tuning). We believe such research is beneficial for designing new regularizers for better transfer learning. Meanwhile, new assessment tools should be developed in the future since recent advanced methods cannot precisely select the best pre-training checkpoint during the same pre-training process.

Furthermore, how such mechanism work in self-supervised learning is also an interesting topic. We provide the results of self-supervised pre-training with MoCo-v2 [7] in Figure 7. The results illustrate that, for FE, self-supervised pre-training does not obey the rules in the supervised case. We hypothesize that self-supervision, which is operated without explicit labels, alleviates the domain gap between source and target since it focuses more on learning an invariant mapping within the same training sample. Due to the page limit, we will investigate this difference between the two pre-training paradigms in future work.

Acknowledgements. This research was supported by National Natural Science Foundation of China (NO.62106272), the Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), in part by the Research Funds of Renmin University of China (NO. 21XNLG17) and Public Computing Cloud, Renmin University of China.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 7

[2] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR, 06–11 Aug 2017. 2

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 3

[4] 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. 3

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 3

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 9

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 3

[9] Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 6, 8

[10] 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, 2009. 1, 2

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[12] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010. 2, 3

[13] Dumitru Erhan, Pierre-Antoine Manzagol, Yoshua Bengio, Samy Bengio, and Pascal Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *Artificial Intelligence and Statistics*, pages 153–160. PMLR, 2009. 3

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 3

[15] Ben Goertzel and Cassio Pennachin. *Artificial general intelligence*, volume 2. Springer, 2007. 3

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 3

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[18] Kaiming He, Ross Girshick, and Piotr Dollar. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[19] 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. 1, 2, 3

[20] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019. 3

[21] Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. What makes imagenet good for transfer learning?, 2016. 3

[22] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020. 2

[23] Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019. 2

[24] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 7

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] 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 (CVPR)*, June 2019. 1, 2, 3, 4

[27] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In

4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia, 2013. 3

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 3

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 3

[32] Ming Li, Jie Wu, Xionghui Wang, Chen Chen, Jie Qin, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Aligndet: Aligning pre-training and fine-tuning in object detection. *arXiv preprint arXiv:2307.11077*, 2023. 3

[33] Xingjian Li, Haoyi Xiong, Haozhe An, Cheng-Zhong Xu, and Dejing Dou. Rifle: Backpropagation in depth for deep transfer learning through re-initializing the fully-connected layer. In *International Conference on Machine Learning*, pages 6010–6019. PMLR, 2020. 5

[34] Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, Zeyu Chen, and Jun Huan. Delta: Deep learning transfer using feature map with attention for convolutional networks. *arXiv preprint arXiv:1901.09229*, 2019. 3

[35] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 4

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[37] UK Lopes and João Francisco Valiati. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. *Computers in biology and medicine*, 89:135–143, 2017. 1, 8

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[39] Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. Theory of the frequency principle for general deep neural networks. *arXiv preprint arXiv:1906.09235*, 2019. 7

[40] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3

[41] Karttikeya Mangalam and Vinay Uday Prabhu. Do deep neural networks learn shallow learnable examples first? 2019. 2

[42] Dimitrios Marmanis, Mihai Datcu, Thomas Esch, and Uwe Stilla. Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109, 2015. 3

[43] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 3

[44] Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pages 7294–7305. PMLR, 2020. 2, 7

[45] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 3

[46] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 3

[47] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[48] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998. 3

[49] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018. 7

[50] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 413–420. IEEE, 2009. 3

[51] 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. 3

[52] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[53] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019. 3

[54] Muhammad Rahman, Yongzhong Cao, Xiaobing Sun, Bin Li, and Yameng Hao. Deep pre-trained networks as a feature extractor with xgboost to detect tuberculosis from chest x-ray. *Computers & Electrical Engineering*, 93:107252, 2021. 1

[55] Sivaramakrishnan Rajaraman, Sameer K Antani, Mahdieh Poostchi, Kamolrat Silamut, Md A Hossain, Richard J Maude, Stefan Jaeger, and George R Thoma. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568, 2018. 1

[56] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. *Hand*, 1(602):2–215. 3

[57] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *ArXiv preprint arXiv:2007.08489*, 2020. 2, 3

[58] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019. 3

[59] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019. 7

[60] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 3

[61] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 7

[62] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1395–1405, 2019. 2, 7

[63] Francisco Utrera, Evan Kravitz, N. Benjamin Erichson, Rajiv Khanna, and Michael W. Mahoney. Adversarially-trained deep nets transfer better: Illustration on image classification. In *ArXiv preprint arXiv:2007.05869*, 2021. 2, 3

[64] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4, 5

[65] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 3

[66] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11293–11302, 2019. 6

[67] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[68] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021. 3

[69] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019. 7, 9

[70] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *International Conference on Neural Information Processing*, pages 264–274. Springer, 2019. 7, 9

[71] LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, pages 2825–2834. PMLR, 2018. 3

[72] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2021. 3

[73] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 3

[74] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. 3

[75] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27:3320–3328, 2014. 1

[76] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *ICML*, 2021. 2, 7

[77] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023. 3

[78] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021. 3, 5

[79] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 3

[80] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1, 2, 3

[81] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. Revisiting few-sample bert fine-tuning. In *International Conference on Learning Representations*, 2021. 5

[82] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3

[83] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3833–3845. Curran Associates, Inc., 2020. 3