Understanding and Improving Transfer Learning of Deep Models via Neural Collapse

Anonymous authors
Paper under double-blind review

Abstract

With the ever-increasing complexity of large-scale pre-trained models coupled with a shortage of labeled data for downstream training, transfer learning has become the primary approach in many fields, including natural language processing, computer vision, and multimodal learning. Despite recent progress, the fine-tuning process for large-scale pre-trained models in vision still mostly relies on trial and error. This work investigates the relationship between neural collapse (NC) and transfer learning for classification problems. NC is an intriguing while prevalent phenomenon that has been recently discovered in terms of the final-layer features and linear classifiers of trained neural networks. Specifically, during the terminal phase of training, NC implies that the variability of the features within each class diminishes to zero, while the means of features between classes are maximally and equally distanced. In this work, we examine the NC attributes of pre-trained models on both downstream and training data for transfer learning, and we find strong correlation between feature collapse and downstream performance. In particular, we discovered a systematic pattern that emerges when linear probing pre-trained models on downstream training data: the more feature collapse of pre-trained models on downstream data, the higher the transfer accuracy. Additionally, we also studied the relationship between NC and transfer accuracy on the training data. We validate these findings through a series of comprehensive experiments under a range of conditions. Moreover, these findings allow us to develop a principled, parameter-efficient fine-tuning method that employs skip-connection to induce the last-layer feature collapse on downstream data. When compared to the full model fine-tuning, our proposed fine-tuning method delivers comparable or even superior performance, while reducing fine-tuning parameters by at least 90% and mitigating overfitting in situations especially when the downstream data is scarce.

1 Introduction

Transfer learning has gained widespread popularity in computer vision, medical imaging, and natural language processing (Zhuang et al., 2020; Devlin et al., 2019; Cheplygina et al., 2019). By taking advantage of domain similarity, a pre-trained, large model from source datasets is reused as a starting point or feature extractor for fine-tuning a new model on a smaller downstream task (Zhuang et al., 2020). The reuse of the pre-trained model during fine-tuning reduces computational costs significantly and leads to superior performance on problems with limited training data. Despite of recent advances, the underlying mechanism of transfer learning is still far from understood, and many of existing approaches lack principled guidance.

In the past, the bulk of research has focused on improving data representations during the phase of model pre-training, with various metrics developed to assess representation quality on source data, such as validation accuracy on ImageNet (Kornblith et al., 2019) and feature diversity (Kornblith et al., 2021; Nayman et al., 2022). However, the evaluation of representation quality has become increasingly challenging due to the scale and limited accessibility of source data. Additionally, with the emergence of large-scale foundation models (Bommasani et al., 2021; Zhou et al., 2023) such as GPT-4 (OpenAI, 2023), SAM (Kirillov et al., 2023), and CLIP (Radford et al., 2021), the research focus has shifted towards the evaluating and fine-tuning of these pre-trained foundation models (Kumar et al., 2022) on downstream tasks.

Therefore, in this work, we mainly focus on the downstream data for evaluating and fine-tuning large-scale pre-trained models. We examine the principles of transfer learning by drawing connections to an intriguing phenomenon prevalent across different network architectures and datasets, termed '*Neural Collapse*" ($\mathcal{NC}$) (Papyan et al., 2020; Han et al., 2022), where the last-layer features and classifiers "collapse" to simple while elegant mathematical structures on the training data. Specifically, during the terminal phase of training, it has been observed that (i) the within-class variability of last-layer features collapses to zero for each class and (ii) the between-class means and last-layer classifiers collapse to the vertices of a Simplex Equiangular Tight Frame (ETF) up to scaling.

While the initial discovery of $\mathcal{NC}$ was rooted in the analysis of the source training data, our study demonstrates a compelling relationship between $\mathcal{NC}$ and transferability on downstream datasets. We adapt the metrics for assessing $\mathcal{NC}$ to gauge the quality of learned representations concerning both *within-class diversity* and *between-class discrimination*. Remarkably, when applying linear probing to pre-trained models on downstream training data, we uncover an intriguing trend: as the features of pre-trained models collapse more significantly on the downstream training data, the transfer accuracy improves. Capitalizing on this insight, we design more principled and parameter-efficient fine-tuning approaches for large-scale pre-trained models, including foundation models like CLIP (Dosovitskiy et al., 2021; Radford et al., 2021).

Finally, for a more comprehensive study, we also investigate the relationship between transferability and $\mathcal{NC}$ metrics on the source training data. We considered different types of factors that affect transfer accuracy of pre-trained models, such as training losses, projection heads, and network architectures. Our work reveals the limitation of using source data for predicting transfer accuracy.

Contributions. In summary, we highlight our contributions below.

- **More collapsed features on downstream tasks lead to better transfer accuracy.** Through an extensive examination of pre-trained models using linear probing across various scenarios, our work shows the following relationship: a higher degree of feature collapse on downstream data tends to yield improved transfer accuracy. This phenomenon is supported by comprehensive experiments conducted on multiple downstream datasets (Krizhevsky et al., 2009; Maji et al., 2013; Cimpoi et al., 2014; Parkhi et al., 2012), diverse pre-trained models (He et al., 2016; Dosovitskiy et al., 2021; Huang et al., 2017; Sandler et al., 2018; Radford et al., 2021), and even within the context of few-shot learning (Tian et al., 2020). Notably, we find that this relationship holds across different layers when linear probing is applied to features of distinct layers from the same pre-trained model, thereby offering valuable insights for the principled design of more effective linear probing techniques.

- **More efficient fine-tuning of large pre-trained models via $\mathcal{NC}$.** Based on the aforementioned insights, we propose a simple and memory-efficient fine-tuning approach that achieves comparable or superior performance to full model fine-tuning in vision classification tasks. Our method revolves around the utilization of skip connections to fine-tune a critical layer in the pre-trained network, thereby maximizing the collapse of the last-layer feature. To validate the effectiveness of our strategy, we conduct experiments using publicly available pre-trained models, including ResNet, ViT, and CLIP (He et al., 2016; Dosovitskiy et al., 2021; Radford et al., 2021). Remarkably, our method substantially reduces the number of fine-tuning parameters by at least 90% compared to full model fine-tuning. Moreover, in the context of low-shot learning, our approach exhibits improved robustness to data scarcity, exhibiting reduced overfitting tendencies.

- **Limitations of studying transfer accuracy using source training data.** In addition to investigating the connection between downstream $\mathcal{NC}$ and transferability, our also studied the relationship between source $\mathcal{NC}$ and transferability. More specifically, we investigated different key factors that affects the transfer accuracy based upon the $\mathcal{NC}$ metrics on the source data, such as loss functions Hui & Belkin (2020), the projection head Chen et al. (2020), data augmentations Chen & He (2021); Khosla et al. (2020), and adversarial training Salman et al. (2020); Deng et al. (2021). Within a certain threshold, we found that the more diverse the features are, the better the transferability of the pre-trained model. However, as randomly generated features that are diverse do not generalize well, the relationship does not always hold and using source data diversity metrics has limitation for predicting transfer accuracy.

Relationship to Prior Art. As we conclude this section, we delve into the significance of our works in comparison to existing results on neural collapse, model pre-training, and parameter-efficient fine-tuning. Below, we provide a summary and brief discussion of these results.

- **Studies of the Neural Collapse Phenomenon.** The occurrence of the $\mathcal{NC}$ phenomenon has recently gained significant attention in both practical and theoretical fields (Papyan et al., 2020; lu2, 2022; Zhu et al., 2021; Fang et al., 2021; Han et al., 2022; Kothapalli, 2023), with studies exploring its implications on training, generalization, and transferability of deep networks; see a recent review paper (Kothapalli, 2023). Most of these existing studies focused on training. For instance, various loss functions, such as cross-entropy (Papyan et al., 2020; lu2, 2022; Zhu et al., 2021; Fang et al., 2021; Ji et al., 2022; Yaras et al., 2022), mean-squared error (Mixon et al., 2020; Han et al., 2022; Zhou et al., 2022; Tirer & Bruna, 2022; Rangamani & Banburski-Fahey, 2022; Wang et al., 2022; Dang et al., 2023), and supervised contrastive (Graf et al., 2021) losses, have been shown to exhibit $\mathcal{NC}$ during training. Other lines of works have investigated the relationship between imbalanced training and $\mathcal{NC}$ (Fang et al., 2021; Xie et al., 2022; Yang et al., 2022; Thrampoulidis et al., 2022; Behnia et al., 2022; Zhong et al., 2023; Sharma et al., 2023; Behnia et al., 2023), generalization to the settings of large classes (Liu et al., 2023b; Gao et al., 2023; Jiang et al., 2023), as well as the progressive feature variability collapse across network layers (Hui et al., 2022; Papyan, 2020; He & Su, 2022; Rangamani et al., 2023). Recent works have also investigated the relationships between $\mathcal{NC}$ and both generalization and transferability (Hui et al., 2022; Galanti et al., 2022b;a; Galanti, 2022; Chen et al., 2022a; Wang et al., 2023). In particular, the study in Galanti et al. (2022b;a) reveal that $\mathcal{NC}$ occurs on test data when drawn from the same distribution as training data, but the collapse is less severe for finite test samples (Hui et al., 2022). While Galanti et al. (2022b;a) also investigated the correlation between $\mathcal{NC}$ and transfer learning, their study focused solely on the influence of the number of source classes. Our research, on the other hand, is much more comprehensive and demonstrates a universal correlation between $\mathcal{NC}$ and transfer accuracy on downstream data [1]. We examine a broad range of settings, including various tasks, pre-trained model architectures, training setups, and even different layers within individual models. Furthermore, our investigation into $\mathcal{NC}$ has contributed to the principled development of more efficient fine-tuning methods.
- **Model Pre-training & Fine-Tuning.** While various studies have explored factors influencing transferability during model pre-training (Kornblith et al., 2019; 2021; Nayman et al., 2022), their findings largely remain ambiguous. Additionally, these studies primarily focus on characterizing transferability based on the source dataset, which is often inaccessible in the era of large-scale pre-trained foundation models (Dosovitskiy et al., 2021; Radford et al., 2021; OpenAI, 2023; Liu et al., 2023a; Touvron et al., 2023). In contrast, our work primarily concentrates on evaluating the transferability of pre-trained models using downstream data. This approach not only enables us to establish a universal relationship between $\mathcal{NC}$ and transferability, but also aligns with the current landscape of large-scale pre-trained models that have limited access to source data. In the field of vision model fine-tuning, two commonly used approaches are linear probing and full model fine-tuning (Kumar et al., 2022). Linear probing focuses on training only the classifier, resulting in parameter-efficient models that often underperform compared to full model fine-tuning. Previous studies have attempted to strike a balance between parameter efficiency and performance by incorporating "adaptors" as auxiliary subnetworks between pre-trained layers (Rebuffi et al., 2017; Houlsby et al., 2019). While effective for transformers in NLP and recently extended to multi-modality and vision domains (Gao et al., 2021; Zhang et al., 2021; Chen et al., 2022b), this approach requires additional structures and is specific to certain network architectures. In contrast, our fine-tuning method does not introduce extra components and can be easily applied to any network architecture. Another line of work involves storing all intermediate layer features and linear probing a subset of them (Evci et al., 2022; Adler et al., 2020). While this approach shows promise when training data for downstream tasks is scarce, it fails to compete with full model fine-tuning when ample training data is available. In contrast, our approach is applicable regardless of data volume and is more memory-efficient since it does not require storing all intermediate features. Furthermore, certain methods aim to adapt pre-trained models to downstream tasks by updating low-rank counterparts of weight matrices (Hu et al.,

[1] The work (Wang et al., 2023) also attempt to establish connections between $\mathcal{NC}$ on downstream data and transfer accuracy.However, their metric is more intricate than ours. Moreover, their study is exclusively concentrated on $\mathcal{NC}$ in downstream tasks of various pre-trained models. In contrast, our research encompasses a more comprehensive scope, examining both inter-model and intra-model $\mathcal{NC}$ across diverse training configurations.

2021; He et al., 2022; Zhang et al., 2023; Chavan et al., 2023). These methods often involve modifying multiple components of pre-trained models at different layers, whereas our approach focuses on updating one single layer of the overall model.

2 Evaluating Pre-trained Models via NC

In this section, we provide a brief overview of the $\mathcal{NC}$ phenomenon, followed by the introduction of metrics for evaluating the quality of learned representations for transfer learning.

Basics of Deep Neural Networks. Let us first introduce some basic notations by considering a multi-class (e.g., K class) classification problem with finite training samples. Let $\{n_k\}_{k=1}^K$ be the number of training samples in each class. Let $\boldsymbol{x}_{k,i}$ denote the ith input data in the kth class ($i \in [n_k]$, $k \in [K]$), and the corresponding one-hot training label is represented by $\mathbf{y}_k \in \mathbb{R}^K$, with only the kth entry equal to 1. Thus, given any input data $\boldsymbol{x}_{k,i}$, we learn a deep network to fit the corresponding (one-hot) training label $\boldsymbol{y}_k$ such that

$$\boldsymbol{y}_k \;\approx\; \psi_{\boldsymbol{\Theta}}(\boldsymbol{x}_{k,i}) \;=\; \underbrace{\boldsymbol{W}_L}_{\textbf{linear classifier } \boldsymbol{W}} \cdot\; \sigma\left(\boldsymbol{W}_{L-1}\cdots\sigma\left(\boldsymbol{W}_1\boldsymbol{x}_{k,i}+\boldsymbol{b}_1\right)+\boldsymbol{b}_{L-1}\right)+\boldsymbol{b}_L, \tag{1}$$

where $\boldsymbol{W} = \boldsymbol{W}_L$ represents the last-layer linear classifier and $\boldsymbol{h}_{k,i} = \boldsymbol{h}(\boldsymbol{x}_{k,i}) = \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,i})$ is a deep hierarchical representation (or feature) of the input $\boldsymbol{x}_{k,i}$. Here, for a L-layer deep network $\psi_{\boldsymbol{\Theta}}(\boldsymbol{x})$, each layer is composed of an affine transformation, followed by a nonlinear activation $\sigma(\cdot)$ and normalization functions (e.g., BatchNorm (Ioffe & Szegedy, 2015)). We use $\boldsymbol{\Theta}$ to denote all the network parameters of $\psi_{\boldsymbol{\Theta}}(\boldsymbol{x})$ and $\boldsymbol{\theta}$ to denote the network parameters of $\phi_{\boldsymbol{\theta}}(\boldsymbol{x})$. Additionally, we use

$$\boldsymbol{H} \;=\; \begin{bmatrix}\boldsymbol{H}_1 & \boldsymbol{H}_2 & \cdots & \boldsymbol{H}_K\end{bmatrix} \in \mathbb{R}^{d\times N}, \quad \boldsymbol{H}_k \;=\; \begin{bmatrix}\boldsymbol{h}_{k,1} & \cdots & \boldsymbol{h}_{k,n}\end{bmatrix} \in \mathbb{R}^{d\times n},\ \forall\, k \in [K],$$

to represent all the features in matrix form. Additionally, the class mean for each class is written as

$$\overline{\boldsymbol{H}} \;:=\; \begin{bmatrix}\overline{\boldsymbol{h}}_1 & \cdots & \overline{\boldsymbol{h}}_K\end{bmatrix} \in \mathbb{R}^{d\times K} \quad \text{and} \quad \overline{\boldsymbol{h}}_k \;:=\; \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{h}_{k,i}, \quad ;\forall\, k \in [K].$$

Accordingly, we denote the global mean of $\boldsymbol{H}$ as $\boldsymbol{h}_G = \frac{1}{K}\sum_{k=1}^N \overline{\boldsymbol{h}}_k$.

A Review of Neural Collapse. It has been widely observed that the last-layer features $\mathbf{H}$ and classifiers $\mathbf{W}$ of a trained network on a balanced training dataset $\{\boldsymbol{x}_{k,i}, \boldsymbol{y}_k\}$ with $n = n_1 = n_2 = \cdots = n_K$ exhibit simple but elegant mathematical structures (Papyan et al., 2020; Papyan, 2020). Here, we highlight two key properties below.[2]

- **Within-class variability collapse.** For each class, the last-layer features collapse to their means,

$$\boldsymbol{h}_{k,i} \to \overline{\boldsymbol{h}}_k, \quad \forall\, 1 \le i \le n,\ 1 \le k \le K. \tag{2}$$

- **Maximum between-class separation.** The class-means $\overline{\boldsymbol{h}}_k$ centered at their global mean $\mathbf{h}_G$ are not only linearly separable but also maximally distant and form a Simplex Equiangular Tight Frame (ETF): for some $c > 0$, $\overline{\boldsymbol{H}} = \begin{bmatrix}\overline{\boldsymbol{h}}_1 - \boldsymbol{h}_G & \cdots & \overline{\boldsymbol{h}}_K - \boldsymbol{h}_G\end{bmatrix}$ satisfies

$$\overline{\boldsymbol{H}}^\top\overline{\boldsymbol{H}} \;=\; \frac{cK}{K-1}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right). \tag{3}$$

Recent studies have shown that $\mathcal{NC}$ is prevalent across a wide range of classification problems (Papyan et al., 2020; Mixon et al., 2020; Zhou et al., 2022; Han et al., 2022; Fang et al., 2021; Graf et al., 2021; Zhou

[2]Additionally, self-duality convergence has also been observed in the sense that $\boldsymbol{w}_k = c'\overline{\boldsymbol{h}}_k$ for some $c' > 0$. We omit them here because they are not the main focus of this work.

et al.) on the source training data, regardless of the loss function used, the neural network architecture and the dataset. Intuitively, the prevalence of $\mathcal{NC}$ phenomenon implies that the features in each class are maximally separated on the training data, and the network learns a max-margin linear classifier in the last-layer. Additionally, if $\mathcal{NC}$ would also occur on downstream data, it could serve as an indicator of the transferability of pre-trained models that we study below.

Measuring the Transferability of Pre-trained Models via $\mathcal{NC}$ Metrics. Based on the above discussion, we can assess the transferability of pre-trained models on the down by measuring the feature diversity and separation using metrics for evaluating $\mathcal{NC}$ (Papyan et al., 2020; Zhu et al., 2021) defined as follows:

$$\mathcal{NC}_1 \;:=\; \frac{1}{K}\operatorname{trace}\left(\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^{\dagger}\right). \tag{4}$$

More specifically, it measures the magnitude of the within-class covariance matrix $\boldsymbol{\Sigma}_W \in \mathbb{R}^{d\times d}$ of the learned features compared to the between-class covariance matrix $\boldsymbol{\Sigma}_B \in \mathbb{R}^{d\times d}$, where

$$\boldsymbol{\Sigma}_W \;:=\; \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(\boldsymbol{h}_{k,i}-\overline{\boldsymbol{h}}_k\right)\left(\boldsymbol{h}_{k,i}-\overline{\boldsymbol{h}}_k\right)^{\top}, \quad \boldsymbol{\Sigma}_B \;:=\; \frac{1}{K}\sum_{k=1}^{K}\left(\overline{\boldsymbol{h}}_k-\boldsymbol{h}_G\right)\left(\overline{\boldsymbol{h}}_k-\boldsymbol{h}_G\right)^{\top}.$$

Here, $\boldsymbol{\Sigma}_B^{\dagger}$ represents the pseudo-inverse of $\boldsymbol{\Sigma}_B$, which normalizes $\boldsymbol{\Sigma}_W$ to capture the relative relationship between the two covariances. Essentially, if the features of each class are more closely clustered around their corresponding class means, $\mathcal{NC}_1$ will be smaller. Conversely, if the class means are more separated, $\mathcal{NC}_1$ will also be smaller for the same $\boldsymbol{\Sigma}_W$. However, computing this metric in (4) can be computationally expensive due to the pseudo-inverse, which is often the case for large models. To address this issue, alternative metrics such as class-distance normalized variance (CDNV) (Galanti et al., 2022b) and numerical rank (Zhou et al., 2022) have been introduced recently; we refer interested readers to Appendix B for more details.

Based upon the above, in the following we propose to evaluate the quality of learned representations for transfer learning using the metric in equation 4 on the *downstream/source data*, providing new insights into model fine-tuning. Specifically, in Section 3, we focus on downstream tasks and find that the $\mathcal{NC}$ metric and transfer accuracy are negatively correlated: smaller $\mathcal{NC}_1$ on downstream data leads to better transfer accuracy. Based on the findings, we will propose a simple and parameter efficient fine-tuning method in Section 4 that outperforms full model fine-tuning. Additionally, we also investigate the relationship between $\mathcal{NC}$ and pre-training in Section 5. The experimental setup is detailed in Appendix C.

3 Study of $\mathcal{NC}$ & Transfer Accuracy on Downstream Tasks

First, we explore the relationship between the $\mathcal{NC}_1$ metric and transfer accuracy by evaluating pre-trained models on downstream data. The practice of transferring pre-trained large models to smaller downstream tasks has become a common approach in NLP (Devlin et al., 2019; Houlsby et al., 2019; Hu et al., 2021), vision (Dosovitskiy et al., 2021; Evci et al., 2022; Adler et al., 2020), and multi-modal learning. In the meanwhile, evaluating $\mathcal{NC}$ metrics of pre-trained models on downstream data is more feasible, as the availability and size of source data are often limited and prohibitive[3]. To maintain control over the factors influencing our study, we focus on *linear probing* pre-trained models, wherein we freeze the entire model and solely train a linear classifier on top of it using the downstream data. Our findings reveal a negative correlation between transfer accuracy and the $\mathcal{NC}_1$ metric on the downstream data. This phenomenon is universal, as it holds true across multiple downstream datasets (Krizhevsky et al., 2009; Maji et al., 2013; Cimpoi et al., 2014; Parkhi et al., 2012), different pre-trained models (He et al., 2016; Dosovitskiy et al., 2021; Huang et al., 2017; Sandler et al., 2018; Radford et al., 2021), the few-shot learning regime, and even across different layers within individual models.

Pre-trained models with more collapsed last-layer features exhibit better transferability. To support our claim, we pre-train ResNet50 models on the Cifar-100 dataset using different levels of data aug-

[3]For instance, the JFT dataset (Sun et al., 2017), used in pretraining the Vision Transformer, is extensive and not publicly accessible.

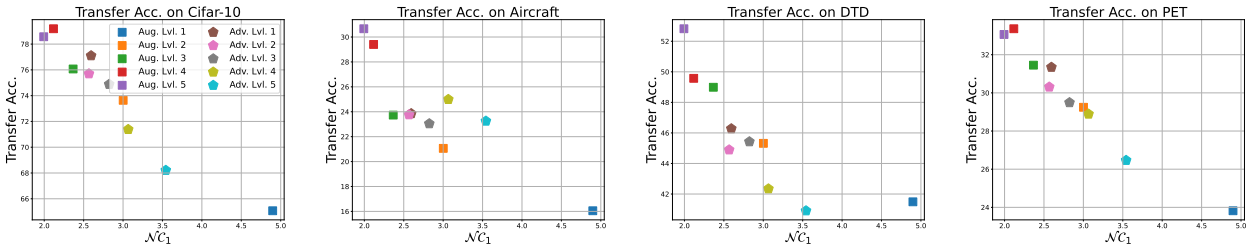


Figure 1: **Transfer accuracy and $\mathcal{NC}_1$ of Cifar-100 pre-trained models on different downstream tasks.** We pre-train ResNet50 models on Cifar-100 using different levels of data augmentation or adversarial training. Here, $\mathcal{NC}_1$ is measured on the downstream Cifar-10 dataset.

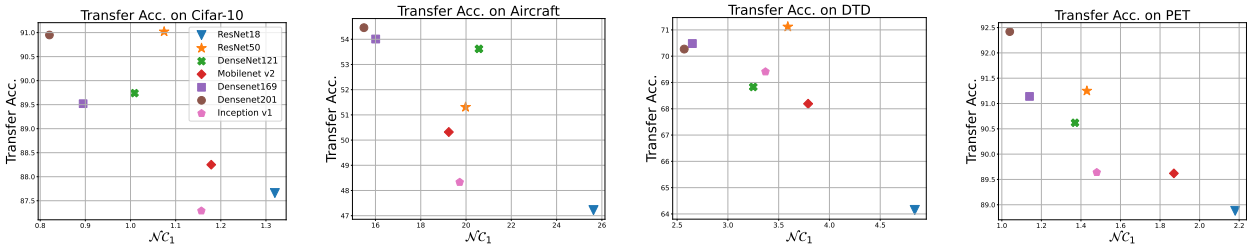


Figure 2: **Transfer accuracy and $\mathcal{NC}_1$ of public ImageNet-1k pre-trained models on different downstream tasks.** We evaluate transfer accuracy and $\mathcal{NC}_1$ on multiple downstream datasets using various ImageNet-1k pre-trained models, such as ResNet He et al. (2016), DenseNet Huang et al. (2017) and MobileNetV2 Sandler et al. (2018). The $\mathcal{NC}_1$ is measured on the corresponding downstream dataset.

mentations and adversarial training (Madry et al., 2018; Salman et al., 2020; Deng et al., 2021) strength,[4]. Once a model is pre-trained, we evaluate its transfer accuracy on four downstream datasets: Cifar-10 (Krizhevsky et al., 2009), FGVC-Aircraft (Maji et al., 2013), DTD (Cimpoi et al., 2014) and Oxford-IIIT-Pet (Parkhi et al., 2012). As shown in Figure 1, we find a negative (near linear) correlation between $\mathcal{NC}_1$ on Cifar-10 and transfer accuracy on the downstream tasks, where lower $\mathcal{NC}_1$ on Cifar-10 corresponds to higher transfer accuracy.[5] Thus, the $\mathcal{NC}_1$ metric on Cifar-10 can serve as a reliable indicator of transfer accuracy on downstream tasks. To further reinforce our argument, we conduct experiments on the same set of downstream tasks using publicly available pre-trained models on ImageNet-1k (Deng et al., 2009), such as ResNet (He et al., 2016), DenseNet (Huang et al., 2017) and MobileNetV2 (Sandler et al., 2018). In Figure 2, we observe the same negative correlation between $\mathcal{NC}_1$ on the downstream data and transfer accuracy, demonstrating that this relationship is not limited to a specific training scenario or network architecture.

Moreover, this negative correlation between downstream $\mathcal{NC}_1$ and transfer accuracy also applies to the few-shot (FS) learning settings, as shown in Table 1 for miniImageNet and CIFAR-FS datasets. Following Tian et al. (2020), we pre-train different models on the merged meta-training data, and then freeze the models, and learn a linear classifier at meta-testing time. During meta-testing time, support images and query images are transformed into embeddings using the fixed neural network. The linear classifier is trained on the support embeddings. On the query images, we compute the $\mathcal{NC}_1$ of the embeddings and record the few-shot classification accuracies using the linear model.

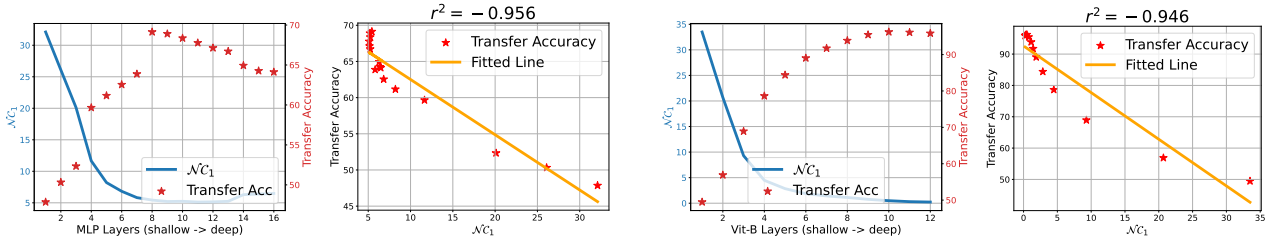
Layers with more collapsed output features result in better transferability. Furthermore, we find the same correlation between the $\mathcal{NC}_1$ metric and transfer accuracy also occurs across different layers of the same pre-trained model. Specifically, as depicted in Figure 4, we linear probe each individual layer of the same pre-trained network, where we use the output of each individual layer as a "feature extractor" and assess

[4]We use 5 levels of data augmentations, each level represents adding one additional type of augmentation, e.g., Level 1 means Normalization, level 2 means Normalization + RandomCrop, etc. For adversarial training strength, we follow the framework in Madry et al. (2018) and consider 5 different attack sizes. Please refer to Appendix C for more details.

[5]When evaluating the correlation between $\mathcal{NC}_1$ and transfer accuracy on the same downstream dataset, the correlation is not as strong as we find on Cifar-10, which we discuss in Appendix D.2.

Table 1: **Average few-shot classification accuracy and $\mathcal{NC}_1$ metrics on CIFAR-FS and Mini-ImageNet.** The test few-shot accuracies and $\mathcal{NC}_1$ are evaluated on miniImageNet and tieredImageNet meta-test splits. ConvNet4 denotes a 4-layer convolutional network with 64 filters in each layer.

Fewshot Accuracy						
Architecture	ConvNet4		ResNet12He et al. (2016)		SEResNet12Hu et al. (2019)	
	1 shot	5 shots	1 shot	5 shots	1 shot	5 shots
CIFAR-FS	61.59	77.45	68.61	82.81	69.99	83.34
Mini-ImageNet	52.04	69.07	59.52	75.92	60.21	77.17
$\mathcal{NC}_1$ of models pre-trained on meta-train splits						
CIFAR-FS	38.12		29.13		28.40	
Mini-ImageNet	51.82		28.24		25.58	



(a) ResNet34 (b) ViT-B

Figure 3: **$\mathcal{NC}_1$ and transfer learning accuracy of different layers from a pre-trained model (Left) and nearly linear relationship between transfer learning accuracy and $\mathcal{NC}_1$ (Right).** We use (a) an ImageNet-1k dataset pre-trained ResNet34 model and (b) a released pre-trained ViT-B model. We use the Cifar-10 dataset for transfer learning and measuring the corresponding $\mathcal{NC}_1$.

its transfer accuracy by training a linear classifier on top of it. Surprisingly, regardless of the layer's depth, if the layer's outputs are more collapsed (smaller $\mathcal{NC}_1$), the resulting features lead to better transfer accuracy. To support our claim, we carried out experiments using the ImageNet-1k pre-trained ResNet34 model (Deng et al., 2009; He et al., 2016). We evaluated the $\mathcal{NC}_1$ metric on each residual block's output feature for the downstream data, as shown in Figure 3 (a). The results indicate that the output features with a smaller $\mathcal{NC}_1$ lead to better transfer accuracy, and there is a near-linear relationship between $\mathcal{NC}_1$ and transfer accuracy, suggesting that transfer accuracy is more closely related to the degree of variability collapse in the layer than the layer's depth. This phenomenon is observed across different network architectures. Figure 3 (b) shows similar results with experiments on the ViT-B (vision transformer base model) (Dosovitskiy et al., 2021) using a pre-trained checkpoint available online.[6] As a result, our findings offer valuable insights into the linear probing of pre-trained models. The conventional practice of solely utilizing last-layer features for linear probing may not be optimal for transfer learning. Instead, employing the $\mathcal{NC}$ metric to identify the optimal layer for linear probing can result in improved performance.

4 Parameter Efficient Fine-tuning via $\mathcal{NC}$ on Downstream Tasks

Furthermore, our observations, as detailed in the previous section, can provide valuable guidance in designing efficient fine-tuning strategies. In the context of adapting large-scale pre-trained models to downstream vision tasks, there are two common approaches: (i) **linear probing** (Khosla et al., 2020; Kornblith et al., 2021; Deng et al., 2021), which uses the pre-trained model as a feature extractor and only learns a linear classifier for the downstream task, and (ii) **full model fine-tuning** (Dosovitskiy et al., 2021; Kornblith et al., 2019; Salman et al., 2020), which adjusts the entire pre-trained model using downstream training data. Linear probing is a highly parameter-efficient method, while full model fine-tuning, albeit more costly particularly for large-scale foundation models, typically yields superior results. However, in scenarios with limited learning data, full model fine-tuning may lead to overfitting; see Figure 7. To balance these trade-offs, recent attention has been given to parameter-efficient fine-tuning in NLP, which selectively adjusts a

[6]The checkpoint used for ViT-B can be found here.

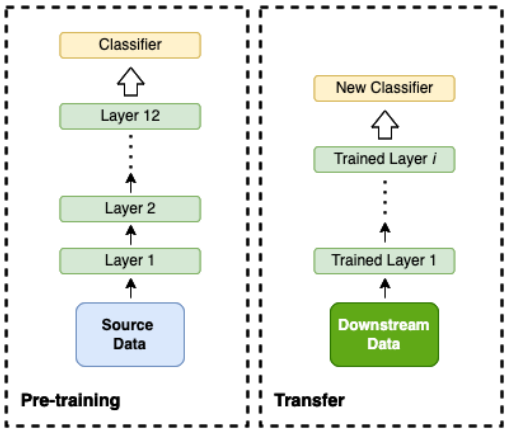


Figure 4: **An illustration of layer-wise transfer learning.** We use a pre-trained model up to the intermediate i-th layer as a feature extractor for transfer learning on the downstream tasks.

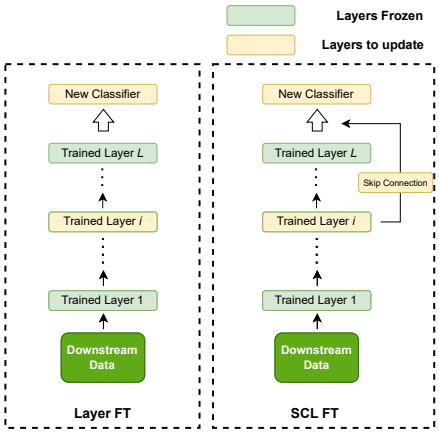


Figure 5: **Illustrations of Layer FL (left) and SCL FT (right).** Layer FL simply fine-tune one intermediate layer, while SCL FT adds a skip connection to the final layer.

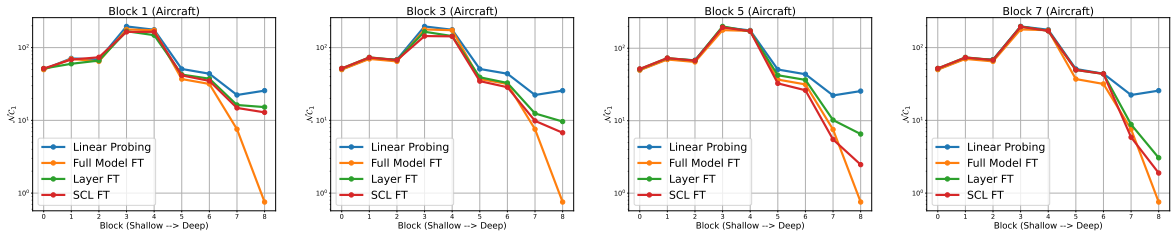


Figure 6: **Layer-wise $\mathcal{NC}_1$ for different fine-tuning methods on ImageNet-1k pre-trained ResNet18.** We compare the layer-wise $\mathcal{NC}_1$ across different fine-tuning methods, including linear probing, Layer FT, and SCL FT, using the ImageNet-1k pre-trained ResNet18 backbone. We evaluate the models on the downstream dataset of FGVC-Aircraft. For the figures from left to right, we plot the results of fine-tuning only Block 1, Block 3, Block 5, and Block 7 in both Layer FT and SCL FT.

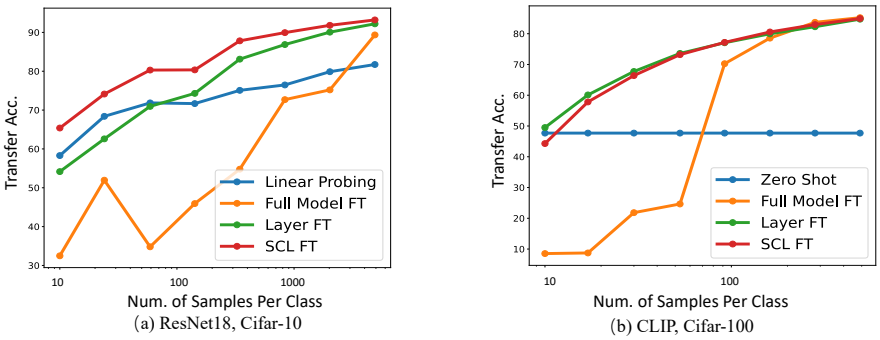
small portion of network parameters (Hu et al., 2021; Rebuffi et al., 2017; Houlsby et al., 2019). Given the growing trend of foundation models beyond NLP (Dosovitskiy et al., 2021; Radford et al., 2021; Zhai et al., 2021), our work focuses on vision classification tasks, exploring parameter-efficient fine-tuning based on the correlation between last-layer feature variability on downstream data and transfer accuracy. Based upon our extensive experiments in Section 3, we conjecture that

The optimal transfer accuracy can be achieved by selectively fine-tuning layers to make the last-layer features as collapsed as possible on the downstream training data.

Based on this, we introduce a simple fine-tuning strategy aimed at increasing feature collapse levels in the last layer, with our results presented in Table 2 and Figure 7. These results show that our proposed method attains comparable or often superior transfer accuracy compared to full model fine-tuning, saving over 90% of parameters (see Table 2) and reducing overfitting when training data is scarce (see Figure 7). In the following, we provide a detailed description of our fine-tuning approach.

Proposed Method: Skip Connection Layer (SCL) Fine-tuning (FT). Inspired by the observation in Section 3, we propose Skip Connection Layer (SCL) Fine-tuning (FT), which consists of two key components:

- **Fine-tuning one key intermediate layer.** To be parameter efficient, we *only* fine-tune one of the intermediate layers while keeping the rest of the network frozen, which we called it *layer fine-tuning*



(a) ResNet18, Cifar-10 (b) CLIP, Cifar-100

Figure 7: **Transfer accuracy for different fine-tuning methods with varying size of downstream training dataset.** We fine-tune ImageNet-1k pre-trained ResNet18 models and CLIP using subsets of the Cifar-10 and Cifar-100 downstream datasets, respectively with varying sizes.

(Layer FT). To find the optimal layer for fine-tuning, we conduct an ablation study to compare the performances by fine-tuning different intermediate layers with rest of the network frozen. As shown in Figure 6, fine-tuning the layer closer to the final layer usually leads to more collapsed features in the last layer and, thus, the better transfer accuracy.[7] This is because the information from the inputs has been better extracted and distilled as getting closer to final layers. On the other hand, in many deep network architectures, layers closer to the output typically have more parameters.[8] Thus, to strike a balance between transfer performance and parameter efficiency, we opt to fine-tune one of the middle layers. More specifically, we fine-tune Block 5 of ResNet18, Block 8 of ResNet50 and Layer 8 for Vit-B32.

- **Improving feature collapse via skip connections.** As illustrated in Figure 5, building on the Layer FT, we introduce a second key component to further the collapse of last-layer feature by adding a skip connection from the key fine-tuned layer to the last layer. We then use the combined features (i.e., the sum of the two outputs) from these layers as the new feature for training the linear classifier and fine-tuning the selected layer. If the dimensions of the features differ (e.g., in CNN-based models), we zero-pad the lower-dimensional feature to match the dimension difference. Our proposed SCL FT method enables more effective fine-tuning of the selected layer by directly passing the data's information from the intermediate layer to the classifier, without losing information through other intermediate layers. Moreover, this approach leverages the depth of deep models, ensuring that the more refined features from the penultimate layer are also passed to the linear classifier. As demonstrated on ResNet in Figure 6, SCL FT leads the most collapsed feature in the last-layer other than full model fine-tuning, and best performance as shown in Figure 6. We also observe a similar phenomenon in the case of ViT, for which we postpone to Appendix D.

Advantages of Our Methods. Through comprehensive experiments, we demonstrate the superiority of our suggested SCL FT over conventional methods on vision classification problems, such as linear probing, full model fine-tuning, and zero-shot learning. Full model fine-tuning requires retraining all parameters, which amounts to 23 million for ResNet50 or 88 million for ViT-B32. In contrast, our methods, SCL FT and Layer FT, are highly efficient, achieving comparable or better results by fine-tuning only around 8% of the model's parameters, as shown in Table 2 and Appendix D.2. Moreover, our methods exhibit improved resistance to overfitting while achieving equal or superior performance compared to full model fine-tuning; see Figure 7. Below, we discuss two main advantages of our method in detail.

- **Surpassing full model fine-tuning with improved parameter efficiency.** As demonstrated in Table 2, our SCL FT technique, when applied with ResNet-based backbones and just fine-tuning 6% – 8% of parameters, delivers a significant performance boost compared to linear probing [9]. It frequently

[7] As shown in Tables 6 to 8, we find that fine-tuning Block 7 for ResNet18 (which has 8 blocks in total), Block 14 for ResNet50 (which has 16 blocks in total), and Layer 11 for ViT models (which have 12 layers in total) always yields the best or near-optimal results.

[8] For example, a ResNet18 model has $512 \times 512 \times 3 \times 3$ parameters in the penultimate layer, while only $64 \times 4 \times 3 \times 3$ parameters are in the input layer.

[9] Please refer to Table 6 for results on ResNet18.

Table 2: **Transfer learning results for linear probing / zero-shot, layer FT, SCL FT and full model FT on downstream datasets.** We use publicly available ResNet50, ViT-B and CLIP models.

Backbone	ResNet50					Vit-B			CLIP		
Dataset	Cifar-10	Cifar-100	Aircraft	DTD	PET	Aircraft	DTD	PET	Aircraft	DTD	PET
Transfer accuracy											
Linear Probe / Zero Shot	85.33	65.47	43.23	68.46	89.26	43.65	73.88	92.23	12.87	32.34	39.66
Layer FT	94.04	77.47	70.27	67.66	89.40	**65.83**	77.13	92.94	**67.63**	**79.47**	**91.09**
SCL FT	**94.94**	78.32	70.72	**72.87**	**91.69**	65.80	**77.34**	**93.13**	66.58	79.04	90.02
Full Model FT	85.51	**78.88**	**80.77**	76.12	73.24	64.66	76.54	93.02	59.11	72.82	84.44
$\mathcal{NC}_1$ **evaluated on the penultimate layer feature** $\boldsymbol{h}^{L-1}$											
Linear Probe / Zero Shot	1.84	18.36	20.36	3.52	1.45	17.91	1.99	0.66	17.47	2.96	3.77
Layer FT	0.28	3.22	3.37	1.68	0.68	6.98	1.62	0.44	1.30	0.42	0.32
SCL FT	0.22	2.61	1.02	0.64	0.39	7.48	1.33	0.35	1.65	0.61	0.46
Full Model FT	0.17	0.15	0.61	0.31	0.28	3.78	1.11	0.21	0.49	0.17	0.18
Percentage of parameters fine-tuned											
Linear Probe / Zero Shot	0.09%	0.86%	0.86%	0.41%	0.32%	0.09%	0.04%	0.03%	0.0%	0.0%	0.0%
Layer FT	6.52%	7.24%	7.24%	6.82%	6.73%	8.18%	8.14%	8.13%	8.16%	8.11%	8.10%
SCL FT	6.52%	7.24%	7.24%	6.82%	6.73%	8.18%	8.14%	8.13%	8.16%	8.11%	8.10%
Full Model FT	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

surpasses full model FT across an array of datasets such as Cifar, FGVC-Aircraft (Maji et al., 2013), DTD (Cimpoi et al., 2014), and Oxford-IIIT-Pet (Parkhi et al., 2012). Our method is easily adaptable, working effectively with both CNN-based (ResNets) and transformer-based (ViT and CLIP) network structures. In addition, our experiments in Table 2 show that when using transformer-based backbones like ViT and CLIP, fine-tuning only the key layer alone yields results that are as good or even superior to those obtained with SCL FT. Nonetheless, both Layer FT and SCL FT outperform full model FT, while utilizing considerably fewer parameters during the fine-tuning process.

- **Mitigating overfitting in the presence of downstream data scarcity.** When dealing with limited downstream data, performing fine-tuning on the entire large-scale pre-trained model can lead to severe overfitting, resulting in poor generalization performance. In contrast, our methods demonstrate much better resilience to data scarcity and decreased overfitting by specifically fine-tuning a small subset of model parameters. To demonstrate this, we conducted fine-tuning experiments on pre-trained ResNet18/CLIP models using varying sizes of subsets from the Cifar-10/Cifar-100 training samples. The outcomes are presented in Figure 7. Our findings reveal that full model fine-tuning is vulnerable to data scarcity, underperforming linear probing/zero-shot approaches when the downstream training data size is limited. In comparison, our SCL FT and Layer FT methods maintain their robustness and significantly surpass full model fine-tuning until a significant amount of downstream training data becomes available.

5 Study of $\mathcal{NC}$ & Transfer Accuracy on Model Pre-training

Finally, we compliment our study by examining the relationship between the $\mathcal{NC}$ metrics and transfer accuracy on the source training dataset. Compared to recent studies on the relationship between feature diversity and transfer accuracy Islam et al. (2021); Kornblith et al. (2021), our work not only examines the effect of different training losses but also (i) provides insights into several popular heuristics in model pre-training (e.g. projection head, data augmentation), and (ii) reveals the limitations of using feature diversity as the sole metric for evaluating the quality of pre-trained models. Specifically, we show that the positive correlation between feature diversity on source data (indicated by larger $\mathcal{NC}_1$) and transfer accuracy only holds within a certain extent.[10]

Impact of training losses and network architectures on feature diversity and transferability. Our investigation shows that the choice of training losses and the design of network architecture have a

[10]This positive correlation only holds up to a certain extent as random features do not collapse, but they do not generalize well. This is because random features are not discriminative. We conjecture that there could be a trade-off between feature diversity and discrimination.

Table 3: **Comparisons of $\mathcal{NC}_1$ and transfer accuracy with different training losses and settings.** We pre-train ResNet18 models on the Cifar-100 dataset, and then test the models on Cifar-10. We use proj. to denote projection head; w/o proj. means without projection head, w/linear proj. means adding one linear layer projection layer, and w/ mlp proj. means adding a two-layer MLP projection.

Training	MSE (w/o proj.)	Cross-entropy (w/o proj.)	SupCon (w/ linear proj.)	SupCon (w/ mlp proj.)
$\mathcal{NC}_1$ **(Cifar-100)**	0.001	0.771	0.792	2.991
Transfer Acc.	53.96	71.2	69.89	79.51

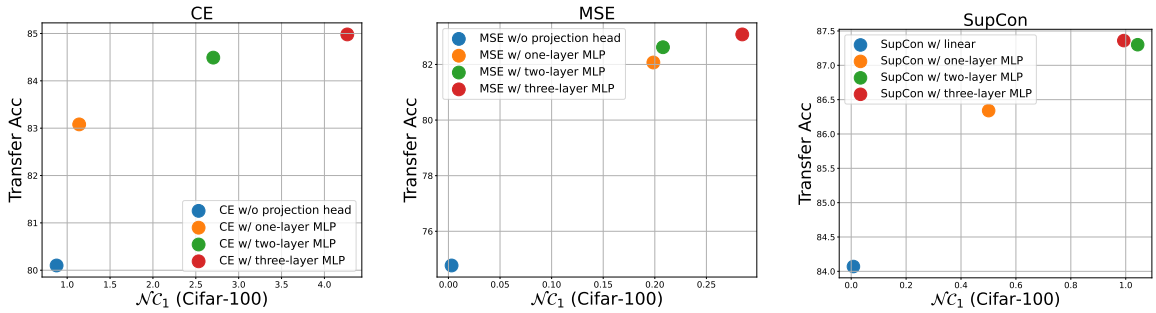


Figure 8: **Trend of $\mathcal{NC}_1$ during training and transfer learning accuracy of the pretrained models.** ResNet-50 models are pretrained using Cifar-100 dataset with CE loss (Left), MSE loss (Middle) and SupCon loss (Right). Models are pretrained with different numbers of layers for projection heads and transfered on the Cifar-10 dataset.

substantial impact on the levels of feature collapse on the penultimate layer and hence on the transfer accuracy. Similar observations were made in Islam et al. (2021); Kornblith et al. (2021) regarding the impact of different training losses on transfer performance. To demonstrate this, we pre-trained ResNet18 models on the Cifar-100 dataset using three different loss functions: CE, MSE (Hui & Belkin, 2020), and SupCon (Khosla et al., 2020). We then evaluated the test accuracy on the Cifar-10 dataset by training a linear classifier on the frozen pre-trained models. For the SupCon loss, we followed the setup described in Khosla et al. (2020), which uses an MLP as a projection head after the ResNet18 encoder. However, after pre-training, we only used the encoder network as the pre-trained model for downstream tasks and abandoned the projection head. The results of $\mathcal{NC}_1$ and the corresponding transfer accuracy for each scenario are summarized in Table 3, where the last two columns reports the results for SupCon with different layers of projection heads. From Table 3, we observe the following.

- *The choice of training loss impacts feature diversity, which in turn affects transfer accuracy.* Larger feature diversity, as measured by a larger $\mathcal{NC}_1$ value, generally leads to better transfer accuracy. For example, a model pre-trained with the MSE loss exhibits a severely collapsed representation on the source dataset, with the smallest $\mathcal{NC}_1$ value and the worst transfer accuracy.
- *The MLP projection head is crucial for improved transferability.* The model pre-trained with the SupCon loss and a multi-layer MLP projection head shows the least feature collapse compared to the other models and demonstrates superior transfer accuracy.[11] If we substitute the MLP with a linear projection layer, both the $\mathcal{NC}_1$ metric and transfer accuracy of SupCon decrease, resulting in performance comparable to the models pre-trained with the CE loss.

Building on the above observation, in the following we delve deeper into the role of the projection head in transfer learning by exploring the progressive decay of the $\mathcal{NC}_1$ metric across layers.

Projection layers in pre-training increase feature diversity for better transferability. The use of projection head for model pre-training was first introduced and then gained popularity in self-supervised learning (Chen et al., 2020; Chen & He, 2021), but the reason for its effectiveness is not fully understood.

[11]This observation aligns with recent work (Islam et al., 2021).

On this end, previous work He & Su (2022) has demonstrated that the feature within-class variability collapses progressively from shallow to deep layers, with the collapse becoming more severe in deeper layers (i.e., smaller $\mathcal{NC}_1$ values). Based upon this discovery, we conclude that the use of a projection layer in pre-training helps to prevent the variability collapse of the encoder network, thereby better preserving the information of the input data with improved transfer performance.

This can be demonstrated by our experiments in Figure 8, where we pre-train ResNet-50 models on the Cifar-100 dataset and report $\mathcal{NC}$ metrics and transfer accuracy for varying numbers of projection head layers (from one to three layers). Our results show that using projection heads significantly increases representation diversity and transfer accuracy – adding more layers of MLP projection leads to higher $\mathcal{NC}_1$ and improved transfer accuracy, although the performance improvement quickly plateaus at three layers of MLP. This suggests that the effectiveness of projection heads in model pre-training is not limited to contrastive losses but applies universally across various training loss types (e.g. CE and MSE).

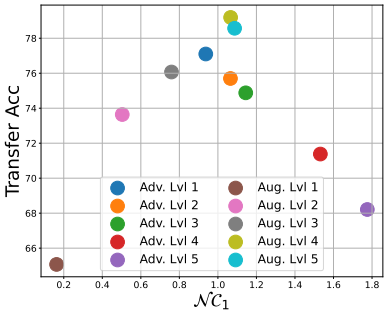


Figure 9: $\mathcal{NC}_1$ **vs. transfer learning accuracy.** Models are pre-trained on the Cifar-100 dataset with different levels of data augmentation and adversarial training strength, and then transfer accuracy of pre-trained models is reported on the Cifar-10 dataset.

Usage of the pre-trained $\mathcal{NC}$ metric for predicting transfer accuracy has limitations. So far, we have seemingly demonstrated an universal positive correlation between the $\mathcal{NC}_1$ and transferability. However, does the increase for the $\mathcal{NC}_1$ of learned features always lead to improved model transferability? To more comprehensively characterize the relationship between $\mathcal{NC}_1$ and transferability, we pre-train ResNet50 models on the Cifar-100 dataset using different levels of data augmentations and adversarial training (Madry et al., 2018; Salman et al., 2020; Deng et al., 2021) strength,[12] and then report the transfer accuracy of the pre-trained models on the Cifar-10 dataset. As shown in Figure 9, we observe that the positive relationship between the $\mathcal{NC}_1$ on the source dataset and the transfer accuracy only holds up to a certain threshold.[13] If the $\mathcal{NC}_1$ metric is larger than a certain threshold, the transfer accuracy decreases as the $\mathcal{NC}_1$ increases.

We believe the reason behind the limitation is that the magnitude of $\mathcal{NC}_1$ is affected by two factors of the learned features: (*i*) within class feature diversity and (*ii*) between class discrimination. When the $\mathcal{NC}_1$ of is too large, the features lose the between-class discrimination, which results in poor transfer accuracy. An extreme example would be an untrained deep model with randomly initialized weights. Obviously, it possesses large $\mathcal{NC}_1$ with large feature diversity, but its features have poor between-class discrimination. Therefore, random features have poor transferability. To better predict the model transferability, we need more precise metrics for measuring both the within-class feature diversity and between-class discrimination, where the two could have a tradeoff between each other. We leave the investigation as future work.

6 Conclusion

In this work, we have explored the relationship between the degree of feature collapse, as measured by the $\mathcal{NC}$, and transferability in transfer learning. Our findings show that there is a twofold relationship between $\mathcal{NC}$ and transferability: (*i*) more collapsed features on the downstream data leads to better transfer performance; and (*ii*) models that are less collapsed on the source data have better transferability up to a certain threshold. This relationship holds both across and within models. Based on these findings, we propose a simple yet effective model fine-tuning method with significantly reduced number of fine-tuning parameters. Our experiments show that our proposed method can achieve comparable or even superior performance compared to full model FT across various tasks and setups. Further discussions of future directions are deferred to Appendix A.

[12] We use 5 levels of data augmentations, each level represents adding one additional type of augmentation, e.g., Level 1 means Normalization, level 2 means Normalization + RandomCrop, etc. For adversarial training strength, we follow the framework in Madry et al. (2018) and consider 5 different attack sizes. Please refer to Appendix C for more details.

[13] The work Kornblith et al. (2021) studied the transferability based upon a notion called class separation, which is similar to our $\mathcal{NC}_1$ metric. They concludes that there is a positive correlation between class separation and transfer accuracy. However, the work only studied the relationship within a limited range, for the cases that the class separation is large.

References

Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.011. Special Issue on Harmonic Analysis and Machine Learning.

Thomas Adler, Johannes Brandstetter, Michael Widrich, Andreas Mayr, David P. Kreil, Michael Kopp, Günter Klambauer, and Sepp Hochreiter. Cross-domain few-shot learning by representation fusion. *ArXiv*, abs/2010.06498, 2020.

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016.

Tina Behnia, Ganesh Ramachandra Kini, Vala Vakilian, and Christos Thrampoulidis. On the implicit geometry of cross-entropy parameterizations for label-imbalanced data. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022. URL `https://openreview.net/forum?id=1piyfD_ictW`.

Tina Behnia, Ganesh Ramachandra Kini, Vala Vakilian, and Christos Thrampoulidis. On the implicit geometry of cross-entropy parameterizations for label-imbalanced data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10815–10838. PMLR, 2023.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258, 2021.

Kwan Ho Ryan Chan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. Redunet: A white-box deep network from the principle of maximizing rate reduction. *ArXiv*, abs/2105.10446, 2021.

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

Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pp. 3090–3122. PMLR, 2022a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2021.

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Y. Qiao. Vision transformer adapter for dense predictions. *ArXiv*, abs/2205.08534, 2022b.

Veronika Cheplygina, Marleen de Bruijne, and Josien P.W. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. pp. 280–296, 2019.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Hien Dang, Tan Nguyen, Tho Tran, Hung Tran, and Nhat Ho. Neural collapse in deep linear network: From balanced to imbalanced data. *arXiv preprint arXiv:2301.00437*, 2023.

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*, pp. 248–255. Ieee, 2009.

Zhun Deng, Linjun Zhang, Kailas Vodrahalli, Kenji Kawaguchi, and James Zou. Adversarial training helps transfer learning via better representations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=f8Dqhg0w-7i`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

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. *ArXiv*, abs/2010.11929, 2021.

Utku Evci, Vincent Dumoulin, H. Larochelle, and Michael Curtis Mozer. Head2toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning*, 2022.

Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

Tomer Galanti. A note on the implicit bias towards minimal depth of deep neural networks. *arXiv preprint arXiv:2202.09028*, 2022.

Tomer Galanti, András György, and Marcus Hutter. Generalization bounds for transfer learning with pre-trained classifiers. *arXiv preprint arXiv:2212.12532*, 2022a.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022b. URL `https://openreview.net/forum?id=SwIp410B6aQ`.

Peifeng Gao, Qianqian Xu, Peisong Wen, Huiyang Shao, Zhiyong Yang, and Qingming Huang. A study of neural collapse phenomenon: Grassmannian frame, symmetry, generalization. *arXiv preprint arXiv:2304.08914*, 2023.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Jiao Qiao. Clip-adapter: Better vision-language models with feature adapters. *ArXiv*, abs/2110.04544, 2021.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pp. 3821–3830. PMLR, 2021.

X.Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=w1UbdvWH_R3`.

Hangfeng He and Weijie J Su. A law of data separation in deep learning. *arXiv preprint arXiv:2210.17020*, 2022.

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*, pp. 770–778, 2016.

Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. Parameter-efficient model adaptation for vision transformers. 2022.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv*, abs/1502.03167, 2015.

Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8845–8855, 2021.

Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J Su. An unconstrained layer-peeled perspective on neural collapse. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=WZ3yjh8coDg`.

Jiachen Jiang, Jinxin Zhou, Peng Wang, Qing Qu, Dustin Mixon, Chong You, and Zhihui Zhu. Generalized neural collapse for a large number of classes. *arXiv preprint arXiv:2310.05351*, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

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*, pp. 2661–2671, 2019.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? In *NeurIPS*, 2021.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=QTXocpAP9p`.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10657–10665, 2019.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023a.

Weiyang Liu, Longhui Yu, Adrian Weller, and Bernhard Schölkopf. Generalizing and decoupling neural collapse via hyperspherical uniformity gap. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=inU2quhGdNU.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

Niv Nayman, Avram Golbert, Asaf Noy, Tan Ping, and Lihi Zelnik-Manor. Diverse imagenet models transfer better. *arXiv preprint arXiv:2204.09134*, 2022.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018.

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Farheen Ramzan, Muhammad Usman Ghani Khan, Asim Rehmat, Sajid Iqbal, Tanzila Saba, Amjad Rehman, and Zahid Mehmood. A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks. *Journal of Medical Systems*, 44(2), 2019.

Akshay Rangamani and Andrzej Banburski-Fahey. Neural collapse in deep homogeneous classifiers and the role of weight decay. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4243–4247. IEEE, 2022.

Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso Poggio. Feature learning in deep classifiers through intermediate neural collapse. Technical report, Center for Brains, Minds and Machines (CBMM), 2023.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *NIPS*, 2017.

T. Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *ArXiv*, abs/2104.10972, 2021.

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.

Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Saurabh Sharma, Yongqin Xian, Ning Yu, and Ambuj K. Singh. Learning prototype classifiers for long-tailed recognition. *ArXiv*, abs/2302.00491, 2023.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Kumar Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 843–852, 2017.

Christos Thrampoulidis, Ganesh R Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *arXiv preprint arXiv:2208.05512*, 2022.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pp. 266–282. Springer, 2020.

Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. *ArXiv*, abs/2201.12760, 2022.

Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. *arXiv preprint arXiv:2202.08087*, 2022.

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.

Peng Wang, Huikang Liu, Can Yaras, Laura Balzano, and Qing Qu. Linear convergence analysis of neural collapse with unconstrained features. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022. URL `https://openreview.net/forum?id=WC9im-M_y5`.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5549–5558, October 2023.

Liang Xie, Yibo Yang, Deng Cai, Dacheng Tao, and Xiaofei He. Neural collapse inspired attraction-repulsion-balanced loss for imbalanced learning. *arXiv preprint arXiv:2204.08735*, 2022.

Yibo Yang, Liang Xie, Shixiang Chen, Xiangtai Li, Zhouchen Lin, and Dacheng Tao. Do we really need a learnable classifier at the end of deep neural network? *arXiv preprint arXiv:2203.09081*, 2022.

Can Yaras, Peng Wang, Zhihui Zhu, Laura Balzano, and Qing Qu. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. *arXiv preprint arXiv:2209.09211*, 2022.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1204–1213, 2021.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=lq62uWRJjiY`.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Jiao Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *ArXiv*, abs/2111.03930, 2021.

Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. Understanding imbalanced semantic segmentation through neural collapse. 2023.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guan Wang, Kaichao Zhang, Cheng Ji, Qi Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, Lichao Sun Michigan State University, Beihang University, Lehigh University, Macquarie University, Nanyang Technological University, University of California at San Diego, Duke University, University of Chicago, and Salesforce AI Research. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *ArXiv*, abs/2302.09419, 2023.

Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. In *Advances in Neural Information Processing Systems*.

Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *International Conference on Machine Learning*, 2022.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34, 2021.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.