Transferring Pretrained Networks to Small Data via Category Decorrelation

Ying Jin*
jiny18@mails.tsinghua.edu.cn
Zhangjie Cao*
caozhangjie14@gmail.com
Mingsheng Long
mingsheng@tsinghua.edu.cn
Jianmin Wang
jimwang@tsinghua.edu.cn

School of Software, BNRist
Research Center for Big Data
Tsinghua University, China

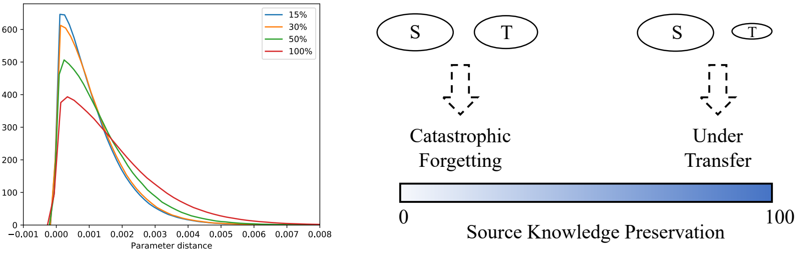
Abstract

Transfer learning by fine-tuning neural networks pre-trained on large-scale datasets excels at accelerating the training process and improving the model performance for the target task. Previous works have unveiled *catastrophic forgetting* in fine-tuning, where the model is over-transferred thus losing pre-trained knowledge, especially facing large-scale target dataset. However, when fine-tuning pre-trained networks to small data, ***under transfer*** emerges instead, where the model sticks to the pre-trained model and learns little target knowledge. Under transfer severely restricts the wide use of fine-tuning but is still under-investigated. In this paper, we conduct an in-depth study of under transfer problem in fine-tuning and observe that when we finetune model to small data, ***redundant category correlation*** becomes stronger in the model prediction, which is a potential cause of under transfer. Based on the observation, we propose a novel regularization approach, **Category Decorrelation (CatDec)**, to minimize category correlation in the model, which introduces a new inductive bias to strengthen the model transfer. CatDec is orthogonal to existing fine-tuning approaches and can collaborate with them to address the dilemma of catastrophic forgetting and under transfer. Experiment results demonstrate that the proposed approach can consistently improve the fine-tuning performance of various mainstream methods. Further analyses prove that CatDec alleviates redundant category correlation and helps transfer.

1 Introduction

Deep learning has achieved revolutionary success in computer vision [8, 9, 23]. Though obtaining significant improvement over previous shallow learning or rule-based methods, deep learning requires a large amount of labeled data to train a highly generalizable model. For practical problems, collecting enough labeled data for deep learning is laborious or even prohibited, *e.g.* in medical image [12]. Apart from the burden of data collection, training deep networks from scratch usually needs intense computational resources and takes tremendous time. Due to the limitations above, the wide use of deep learning is constrained.

[*] Equal contribution



(a) Parameter Distance (b) Catastrophic Forgetting V.S. Under Transfer

Figure 1: (a) Distribution of the distance between model parameters and the pre-trained ones. Different curves correspond to different portions of target labeled data for fine-tuning. As the number of target labeled data decreases, the parameter distance from the pre-trained model becomes smaller, which means under transfer. (b) When data is abundant, the model suffers from catastrophic forgetting, preserving little source knowledge. While when data is scarce, the model switches to under transfer, which means that the model sticks to source knowledge and learns little target knowledge.

A simple yet effective approach to mitigate the limitations is *fine-tuning*, which is an effective approach of transfer learning to deep networks. Fine-tuning typically initializes the parameters as the model pre-trained on large-scale datasets and then tunes the parameters with the data in the target domain. It exploits the advantage of deep learning models that the deep representations learned on large-scale datasets are transferable across various tasks and domains [5, 20, 33].

As the existed fine-tuning tasks include relatively abundant target data, the model is likely to lose knowledge learned in pre-training when incorporating new information from the target training data. Such a problem is called catastrophic forgetting, where the model tends to over-fit the target data, which deteriorates the generalization performance. Most of the previous fine-tuning works [16, 17] focus on catastrophic forgetting and introduce new regularization methods to avoid catastrophic forgetting, where significant progress has been made to alleviate the challenge.

Table 1: Comparison of different fine-tuning methods.

Method	Challenge		Dataset Size	
	Catastrophic Forgetting	Under Transfer	Small	Large
L2 (original)	×	×	×	✓
L2-SP	✓	×	×	✓
DELTA	✓	×	×	✓
CatDec (our method)	×	✓	✓	✓
CatDec + existing methods	✓	✓	✓	✓

However, little attention has been paid to the other side of the coin: the situation when target data is relatively small, which is the focus of this paper. Note that it is also different from few-shot learning [26] which assumes that target labeled data is too scarce to enable fine-tuning while we assume the data amount is still eligible to perform fine-tuning. We conduct analysis on the model parameters before and after vanilla fine-tuning. As shown in Figure 1, with sufficient target labeled data, the fine-tuned model deviates from the pre-trained model with a large divergence. While with insufficient target labeled data, the parameter divergence drops dramatically, which means that the model sticks to the pre-trained model and learns little new target knowledge. As catastrophic forgetting can be viewed as *over transfer*, we can consider this phenomenon as **under transfer**. Under transfer is caused by two reasons: 1) the small target data contains too insufficient target knowledge to draw the model to deviate from the source pre-trained model and approach the target domain; 2) regularization terms for catastrophic forgetting in previous works exacerbate the under transfer problem by constraining the fine-tuned model to stay near the pre-trained model. To address the dilemma of catastrophic forgetting and under transfer, we cannot simply ask for more target data or remove the regularization terms, since the former violates the small data prerequisite while the latter may bring back catastrophic forgetting.

Human beings also face the problem of under transfer when there is little data in the target domain. For example, we are shown a few images of several new species and then asked to recognize these species. We usually address the problem by using **inductive bias**, such as the fact that one instance cannot belong to two species and different species should not have too strong correlation or they are likely to be the same species.

In this paper, we also leverage inductive bias to introduce an 'extra force' to strengthen the model transfer. We observe that **heavy category correlations** exist in the model prediction, especially in the target domain, which are mostly redundant. Such redundant correlations are learned by the source pre-trained model and preserved due to under transfer, which causes the probability of some classes to be simultaneously high and thus harms classification. Based on the observation, we propose **Category Decorrelation (CatDec)** to address the under transfer challenge in fine-tuning. We design a new regularization term to minimize the correlation between classes, which introduces new inductive bias into transfer learning to remove redundant correlations of the pre-trained model and enhance the model transfer. Note that our work is orthogonal to existing methods and can be embedded into current methods to mitigate both under transfer and catastrophic forgetting. A comparison of previous fine-tuning methods and our CatDec is presented in Table 1. The contributions of the paper can be summarized as:

- We conduct an in-depth study of the **under transfer** problem occurring in fine-tuning with small data, which is overlooked by previous works;

- We propose **Category Decorrelation (CatDec)** to alleviate under transfer by minimizing redundant category correlation, which introduces a new inductive bias to strengthen the model transfer.

- We conduct experiments on several benchmarks with both small target data and full target data. Experiment results prove that the proposed CatDec approach consistently improves previous fine-tuning methods. Further results demonstrate that CatDec removes redundant category correlation and boosts model transfer under small data.

2 Related Works

Transfer learning is an important machine learning paradigm which transfers knowledge from a source domain to a target domain [2, 22]. Different transfer learning settings are proposed such as inductive transfer learning [30], multi-task learning [2] and domain adaptation [25]. In this paper, we focus on inductive transfer learning for deep networks, where the target label space is different from the source one and target labeled data is available.

Fine-tuning is a promising approach to inductive transfer learning for deep networks. Fine-tuning firstly pre-trains the deep network on existing large-scale datasets and then tunes the network with target labeled data. Donahue *et al.* [5] fix the feature extractor weight and train a label predictor to classify the features. Yosinki *et al.* [33] demonstrate that representations learned by deep networks are transferable and quantify the transferability. Huh *et al.* [11] dig deeper into deep transfer learning by analyzing features extracted by different networks trained on ImageNet. Recently, plentiful works emerge to improve the fine-tuning from varieties of perspectives, including filter distribution constraining [1], sparse transfer [18], and filter subset selection [4, 7]. Kornblith *et al.* [14] further investigates the influential factors on deep inductive transfer.

Catastrophic forgetting is an important challenge in inductive transfer learning, which is originated from incremental learning [28] and lifelong learning [29]. In inductive transfer learning, the pre-trained networks may lose previously learned knowledge when being tuned to the target task and obtaining knowledge specific to the target task. L2-SP [17] prevents catastrophic forgetting by constraining the divergence between the current model parameters and pre-trained parameters. DELTA [16] proposes a feature map regularization with attention motivated by knowledge distillation for model compression [10, 24, 32, 34]. BSS [3] studies and addresses negative transfer in fine-tuning.

Few-shot learning aims at classifying examples from new classes (called query instances) with only a few labeled instances in each class (called support instances) [6, 21, 27]. In few-shot learning, labeled data is very scarce, e.g. only one or several pieces in each class. So typical few-shot learning methods do not re-train the whole network. But for fine-tuning, we have much more data in each class, e.g. 10%, which are enough to re-train the whole network. As shown in fine-tuning works [3, 16], even naive fine-tuning outperforms few-shot methods by a large margin.

Under Transfer means that the pre-trained model is not transferred to the target domain enough, which usually happens with small data. The previous works on fine-tuning emphasize the hazard of catastrophic forgetting while neglecting the under transfer problem. Apart from inadequate target knowledge caused by small data, their regularization terms make the under transfer problem worse since they generally reduce the divergence between the fine-tuned model parameters and the pre-trained model parameters. Aiming to avoid under transfer, we propose Category Decorrelation (CatDec) to remove redundant category correlation, which produces a new inductive bias to bolster model transfer. Note that our work is orthogonal to existing methods on catastrophic forgetting and can enhance fine-tuning performance of these methods.

3 Method

In fine-tuning, as shown in Figure 3, we have a pre-trained model consisting of a feature extractor (F_0) and a classifier (C_0) and a labeled target dataset. We train the model to fit the target dataset, where the fine-tuned feature extractor and classifier are denoted by F and C. The main difference between fine-tuning and domain adaptation is that the former has labeled target datasets, while the latter has purely unlabeled target datasets. We focus on small data fine-tuning in this paper. In this section, we first investigate 1) under transfer phenomenon in fine-tuning and 2) category correlation caused by under transfer. Based on this, we propose Category Decorrelation to alleviate under transfer.

3.1 Under Transfer in Fine-tuning

When fine-tuning with large data, the model parameters are tuned to fit the target data well and the model is transferred to the target domain adequately. Under such a situation, as shown in Figure 1, the parameter divergence between the fine-tuned model and the pre-trained model is large. To achieve adequate model transfer when only small data is available, the parameter divergence should be at least comparable to the divergence for large data. However, with smaller data, the parameter divergence actually decreases significantly, which demonstrates that under transfer occurs in previous fine-tuning methods.

Under transfer is attributed to the inadequate target knowledge and the regularization terms designed for catastrophic forgetting, which constrain the parameter divergence between the fine-tuned and the pre-trained model [16, 17]. Actually, under transfer and catastrophic forgetting are two extremes of the model transfer. Catastrophic forgetting over-transfers the model and lose all pre-trained knowledge while under transfer learns little new target knowledge and preserves too much pre-trained knowledge. To address the dilemma, the solution should alleviate under transfer with small target data while at the same time not cause severer catastrophic forgetting. Therefore, a new inductive bias, which requires no new target data, is suitable for strengthening model transfer.

3.2 Category Correlation when Under Transfer

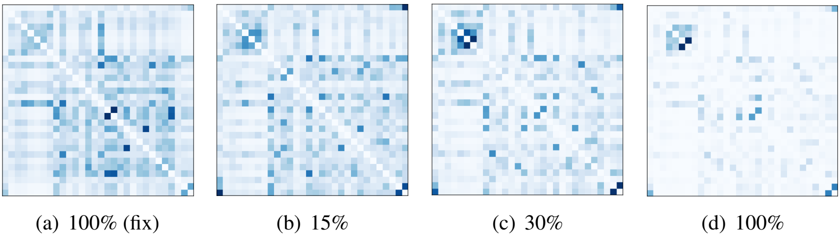

(a) 100% (fix) (b) 15% (c) 30% (d) 100%

Figure 2: Class dependence based on the predictions of fine-tuned models, darker color means higher dependence (the diagonal entries are removed for clearer demonstration). Both models are fine-tuned with L2-SP. The 4 figures, from left to right, correspond to models fine-tuned by 100%, 15%, 30%, 100% target labeled data while we fix the parameters of layers except the classifier for the first figure and tunes all parameters for other figures.

To design a proper inductive bias, we conduct experiments to seek factors related to under transfer. We evaluate the category correlation in the predictions made by fine-tuned models, which is computed by the Harmonic mean of the probability of two classes. As shown in Figure 2, with large data, there is only little category correlation, while fixing model parameters, i.e. zero parameter divergence, has highly redundant category correlations. With 15% and 30% target labeled data, the redundant category correlations also exist. The redundant correlations are learned by the source pre-trained model and preserved in the fine-tuning process, which is part of under transfer. Such correlations add unnecessary constraints on the probability of different classes, which influence the model performance. Therefore, in this paper, we propose **Category Decorrelation (CatDec)** to reduce redundant category correlation, which introduces a new inductive bias to strengthen model transfer. Our approach is orthogonal to existing fine-tuning methods and easy to implement.

3.3 Category Decorrelation

We propose our category decorrelation loss in this section. Suppose N is the number of examples, we use $Z \in \mathbb{R}^{N \times |\mathcal{C}|}$ to denote the raw logit output of the classifier and $\widehat{P} \in \mathbb{R}^{N \times |\mathcal{C}|}$ to denote the model prediction, i.e. a probability distribution over all classes, for all the samples, where both Z and $\widehat{P}$ are matrices and each line corresponds to a data sample. For the i^{th} instance, the probability that it belongs to the j^{th} class $\widehat{P}_{ij}$ can be derived by Z_{ij} as

$$\widehat{P}_{ij} = \frac{\exp\left(Z_{ij}/T\right)}{\sum_{j'=1}^{|\mathcal{C}|} \exp\left(Z_{ij'}/T\right)}, \tag{1}$$

where T is the temperature for scaling. $\widehat{P}$ indicates the temperatured softmax output, which shrinks to original softmax prediction when $T = 1.0$. We can tune T to achieve smoother or sharper probability distributions.

To reduce the redundant category correlation, we need to prevent class pairs from being predicted with a high probability simultaneously. Thus, we maximize the distance between the probabilities of class pairs. Let us take the columns of $\widehat{P}$ into consideration. For column $\widehat{P}_{\cdot,j}$ and $\widehat{P}_{\cdot,k}$, the distance between the j-th and k-th classes R_{jk} can be derived by

$$R_{jk} = \left\| \widehat{P}_{\cdot,j} - \widehat{P}_{\cdot,k} \right\|_2. \tag{2}$$

Maximizing R_{jk} for all the classe pairs can remove the redundant category correlation, which enables us to strengthen model transfer.

Furthermore, we discover that different samples are not equally important when modeling category correlation. The sample that is classified correctly with high confidence, showing high 'peak' on the right class, is more suitable than the sample that is classified incorrectly. Therefore, we use the prediction probability of the ground-truth class as a measure of importance and propose the following weight function on samples,

$$W_i = \frac{N\left(1 + \exp(\widehat{P}_{i,y_i})\right)}{\sum\limits_{i'=1}^{N}\left(1 + \exp(\widehat{P}_{i',y_{i'}})\right)}, \tag{3}$$

where N is the number of samples, y_i is the ground-truth label of the i^{th} sample. Samples that the model gives wrong predictions to will be suppressed with this weighting mechanism. We adopt Laplace Smoothing and sqrt function, forming a *heavy-tailed* distribution to avoid over-penalization. Then the weighted classifier prediction matrix becomes

$$R'_{j,k} = R_{j,k} * (W_i)^{1/2}, \tag{4}$$

With the above R', we can derive our regularization loss as follows,

$$L = -\frac{1}{|\mathcal{C}|^2} \sum_{i=1}^{|\mathcal{C}|} \sum_{j \neq i}^{|\mathcal{C}|} R'_{i,j}, \tag{5}$$

where $\mathcal{C}$ is the label space of the target domain. Minimizing the loss actually maximizes the distance R', so we add a negative in the loss. Our loss function can eliminate all the non-negligible probability pairs, which removes the correlation between any two categories.

The computation of R in Eq.(2) can be derived in the matrix form. The computation burden can be reduced by advanced matrix multiplication algorithms such as Strassen algorithm. So our method brings little computation burden.

Our overall optimization objective is

$$\min_{F,C} E_{(\mathbf{x},\mathbf{y})} L_{\mathrm{CE}}\left(\widehat{P}, y\right) + \alpha L_{\mathrm{CatDec}}\left(\widehat{P}, y\right) + \beta L_{\mathrm{Reg}}, \tag{6}$$

where L_{Reg} is the potential regularization terms of previous methods [16, 17] and can also be $L_{\mathrm{Reg}} = 0$, meaning using no other regularization. α and β are trade-offs for our and other regularization terms. In the experiments, we perform experiments with different regularization terms and our CatDec consistently improves the performance.

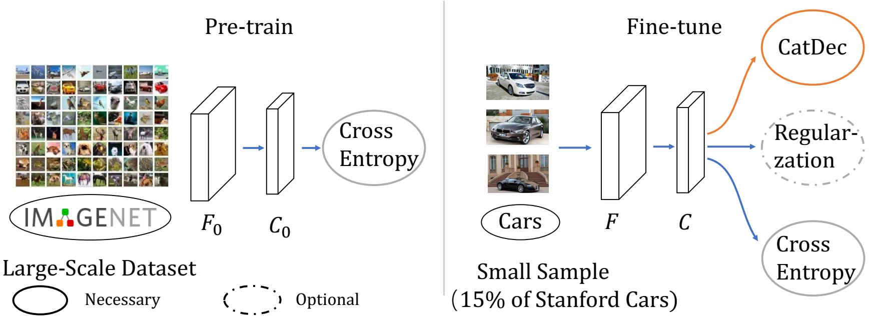


Figure 3: The architecture for fine-tuning. The left figure is the pre-training phase, where the model is trained with large-scale datasets such as ImageNet with cross-entropy loss. The right figure is the fine-tuning phase. We initialize the feature extractor F with the pre-trained feature extractor but random initialize the classifier C. Then the model is trained end-to-end by our CatDec loss, cross-entropy loss and optional regularization terms of previous fine-tuning methods.

3.4 Training Process

As shown in Figure 3, the overall training is two-phased, the first phase is the pre-training phase where the model is trained with source data. This phase can be ignored if the pre-trained model is directly available. The second phase is the fine-tuning phase, where the model is initialized with the parameters of the pre-trained model except the classifier. The last fully connected layer is randomly initialized since the class space of the target domain is always different from the source domain. Then the overall model is trained end-to-end with target labeled data with the objective function in Equation 6.

Our Category Decorrelation reduces the redundant category correlation in the prediction to introduce a new inductive bias to address under transfer, which is orthogonal but compatible with current fine-tuning methods. Plugged into current fine-tuning methods for catastrophic forgetting, the overall approach can mitigate the dilemma of catastrophic forgetting and under transfer.

4 Experiments

We plug CatDec to mainstream inductive transfer learning methods: L2, L2-SP [17] and DELTA [16] on several benchmarks for evaluation since CatDec is orthogonal to them.

4.1 Dataset

We choose the following four benchmark datasets in our experiments, among which the first one is highly similar to ImageNet (a subset), while the other three datasets are quite different from ImageNet.

Stanford Dogs [13] is a subset of ImageNet with over $20,000$ images from 120 breeds.

CUB-200-2011 [31] is a benchmark dataset for fine-grained image classification with $11,788$ images from 200 bird breeds.

Stanford Cars [15] is a dataset with $16,815$ images from 186 different kinds of cars.

FGVC Aircraft [19] is a benchmark dataset for aircraft classification with 102 different kinds of aircraft, each class has 100 images, so there are $10,200$ images in total.

4.2 Implementation Details

In our experiments, we sample 15%, 30%, 50%, 100% of the original train set respectively, and evaluate the model performance on the test set to observe whether our method can alleviate insufficient transfer. Since our work is orthogonal to previous methods, we compare the performance before and after adding our CatDec term to previous fine-tuning methods, including naive fine-tuning (L2), L2-SP [17] and DELTA [16]

For a fair comparison, we follow the fine-tuning protocol in mainstream methods [16], fine-tuning ResNet-50 pre-trained on ImageNet to other target datasets. We adopt the optimal hyper-parameter setting and training strategies in their original work. We note that the last fully connected layer has a learning rate which is 10 times bigger than other layers since it is trained from scratch. The trade-off α between CatDec loss and cross-entropy loss is set to 1.0, which is tuned through cross-validation on the target labeled data. For each sample rate, we sample 5 different subsets, run experiments 5 times on each dataset and report the average top-1 accuracy.

4.3 Results

Regularization Term. CatDec can serve as a regularization term to various existing fine-tuning methods. Table 2 shows the results of fine-tuning the ImageNet pre-trained models. We can observe that when the dataset has large divergence from ImageNet, such as CUB-200-2011, Stanford Cars and FGVC Aircraft, our method can consistently improve the model performance of existing methods, especially when the labeled data is limited (15% or 30%). This demonstrates that CatDec specially addresses under transfer when fine-tuning to small data. On the other hand, with sufficient labeled data, our method still has modest improvement, which demonstrates that CatDec is also generally a good regularization term for fine-tuning. For Stanford Dogs, CUB-200-2011, Stanford Cars and FGVC Aircraft, the average standard deviations for one sampled subset are 0.28, 0.21, 0.30, 0.31 and 0.30, 0.48, 0.38, 0.25 for different sampled subsets under one sampling rate respectively, showing the stability of our method and demonstrating the significance of the results.

Comparison with other methods. There are other regularization terms for fine-tuning, among them the latest method is BSS [3]. Here, we compare our method with BSS on FGVC Aircraft, the most difficult dataset we use in our experiments. Table 3 indicates that overall CatDec outperforms BSS, especially when the target data is small for fine-tuning.

Table 2: Comparison of Accuracy (ResNet-50) of Different Sampling Rate and Different Methods.

(a) Stanford Dogs

Method	Sampling Rate			
	15%	30%	50%	100%
L2	81.05	84.47	85.69	86.89
+ CatDec	**81.62**	**84.50**	**86.01**	**87.11**
L2-SP	81.41	84.88	85.99	86.72
+ CatDec	**81.52**	**85.06**	**86.31**	**86.77**
DELTA	81.46	83.66	84.73	86.01
+ CatDec	**81.98**	**83.75**	**84.80**	**86.34**

(b) CUB-200-2011

Method	Sampling Rate			
	15%	30%	50%	100%
L2	45.25	59.68	70.12	78.01
+ CatDec	**47.41**	**64.84**	**72.63**	**80.08**
L2-SP	45.08	57.78	69.47	78.44
+ CatDec	**46.77**	**60.48**	**70.19**	**78.50**
DELTA	46.83	60.37	71.38	78.63
+ CatDec	**52.99**	**64.41**	**72.49**	**78.68**

(c) Stanford Cars

Method	Sampling Rate			
	15%	30%	50%	100%
L2	36.77	60.63	75.10	87.20
+ CatDec	**40.13**	**66.51**	**79.39**	**87.55**
L2-SP	36.10	60.30	75.48	86.58
+ CatDec	**41.21**	**66.15**	**76.46**	**86.93**
DELTA	39.37	63.28	76.53	86.32
+ CatDec	**43.13**	**67.83**	**78.70**	**87.89**

(d) FGVC Aircraft

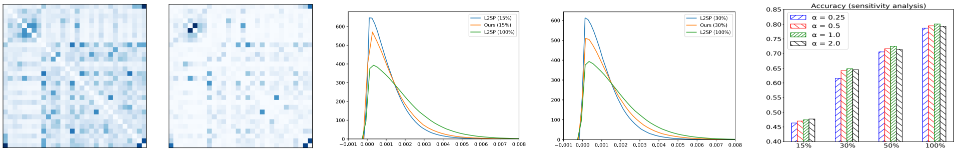
Method	Sampling Rate			
	15%	30%	50%	100%
L2	39.57	57.46	67.93	81.13
+ CatDec	**42.21**	**61.50**	**72.79**	**81.53**
L2-SP	39.27	57.12	67.46	80.98
+ CatDec	**46.59**	**63.31**	**71.41**	**81.22**
DELTA	42.16	58.60	68.51	80.44
+ CatDec	**48.54**	**65.29**	**72.64**	**81.70**

Table 3: Comparison of Accuracy with BSS.

Method	Sampling Rate			
	15%	30%	50%	100%
L2 + BSS	40.41	59.23	69.19	81.48
L2 + CatDec	**42.21**	**61.50**	**72.79**	**81.53**
L2-SP + BSS	40.02	58.78	68.96	**81.27**
L2-SP + CatDec	**46.59**	**63.31**	**71.41**	81.22
DELTA + BSS	43.79	61.58	69.46	80.85
DELTA + CatDec	**48.54**	**65.29**	**72.64**	**81.70**

Table 4: Accuracy with Different Metrics.

Metric	Sampling Rate			
	15%	30%	50%	100%
Baseline	45.25	59.68	70.12	78.01
Inner Product	46.67	63.95	71.71	79.67
Gaussian	46.22	64.39	72.13	78.97
Harmonic Mean	45.54	62.39	70.52	78.40
L2 Norm (CatDec)	**47.41**	**64.84**	**72.63**	**80.08**



(a) L2-SP (b) L2-SP + ours (c) 15% (d) 30% (e) Sensitivity analysis

Figure 4: (a)(b): Category Correlation of classifiers fine-tuned on 15% of Stanford Cars. Our method significantly alleviates the high correlation between categories. (c)(d): Parameter Distances from the pre-trained ones. Models are finetuned on 15% and 30% of Stanford Cars respectively. (e): Hyper-parameter sensitivity on α.

4.4 Analysis

Category Correlation. Figure 4 shows the category correlation of the classifier fine-tuned on 15% of data from Stanford Cars. We can observe that L2-SP shows heavy category correlation that does not exist in practice, while our method can obviously reduce such redundant correlation.

Parameter Distance, As discussed in Figure 1, parameters of the classifier fine-tuned on limited target data have smaller divergence from the pre-trained parameters, which is an indicator of under transfer. Here, we visualize such divergence of models fine-tuned by L2-SP and our method on limited data and L2-SP on the full set of data. As shown in Figure 4, our method can enlarge the parameter divergence from the pre-trained model. Also, our divergence is closer to the divergence of L2-SP on the full set, which is the optimal fine-tuning model. The results prove that the proposed CatDec can enhance model transfer.

Metrics. In CatDec, we use $R_{jk} = \left\|\widehat{P}_{\cdot,j} - \widehat{P}_{\cdot,k}\right\|_2$ to depict the category correlation between the j^{th} and k^{th} class. There are also other metrics that can reflect this correlation, such as negative inner product $R_{j,k} = -\widehat{P}_{\cdot,j} \cdot \widehat{P}_{\cdot,k}$, gaussian kernel function $R_{j,k} = e^{-\frac{1}{2}\|\widehat{P}_{\cdot,j}-\widehat{P}_{\cdot,k}\|_2^2}$, and harmonic mean $R_{j,k} = \frac{1}{N}\sum_{i=1}^{N}\frac{\widehat{P}_{i,j}\cdot\widehat{P}_{i,k}}{\widehat{P}_{i,j}\cdot+\widehat{P}_{i,k}}$. For the three metrics above, smaller $R_{j,k}$ means stronger correlation between the j^{th} and k^{th} category. To examine their effectiveness, we substitute the R_{ij} in Equation 2 with these metrics and evaluate their performance. Table 4 shows the results of different metrics when applying to L2 on CUB-200-2011. Alleviating category correlation, all these metrics can bring about improvements to the vanilla fine-tuning method, proving that category decorrelation is beneficial to knowledge transfer. Among them, the metric used in CatDec, L2 norm, achieves the highest performance, proving that it cooperates with the cross-entropy loss best.

Hyper-parameter Sensitivity. We conduct hyper-parameter sensitivity analysis on the trade-off α between our loss and cross-entropy loss. Figure 4(e) shows the accuracy of CatDec+L2 trained with different α values and different sampling rates on CUB-200-2011. CatDec works stably within a range of α across different portions of target data.

5 Conclusion

In this paper, we conduct an in-depth study of under transfer in fine-tuning with small data. Under transfer and catastrophic forgetting are two extremes of model transfer. To address under transfer, we propose Category Decorrelation (CatDec) to reduce redundant category correlation, which introduces new inductive bias to strengthen model transfer. Our CatDec is orthogonal to previous works on catastrophic forgetting and can be easily plugged into their methods to address the dilemma of catastrophic forgetting and under transfer. Experiment results prove that the proposed CatDec consistently improves the fine-tuning performance of previous fine-tuning methods on several benchmarks.

References

[1] Mehmet Aygun, Yusuf Aytar, and Hazim Kemal Ekenel. Exploiting convolution filter patterns for transfer learning. In *ICCV*, 2017.

[2] Rich Caruana. Multitask learning. *Machine learning*, 1997.

[3] Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In *NeurIPS*, 2019.

[4] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *CVPR*, 2018.

[5] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[6] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *ICLR*, 2018.

[7] Weifeng Ge and Yizhou Yu. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *CVPR*, 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[11] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[12] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 2019.

[13] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPR*, 2011.

[14] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? *CVPR*, 2019.

[15] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *3dRR*, Sydney, Australia, 2013.

[16] Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, and Jun Huan. Delta: Deep learning transfer using feature map with attention for convolutional networks. In *ICLR*, 2019.

[17] Xuhong Li, Grandvalet Yves, and Davoine Franck. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, 2018.

[18] Jiaming Liu, Yali Wang, and Yu Qiao. Sparse deep transfer learning for convolutional neural network. In *AAAI*, 2017.

[19] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *Technical report*, 2013.

[20] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[21] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, pages 721–731, 2018.

[22] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *TKDE*, 2009.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[24] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[25] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

[26] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[27] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208, 2018.

[28] Nadeem Ahmed Syed, Syed Huan, Liu Kah, and Kay Sung. Incremental learning with support vector machines. 1999.

[29] Sebastian Thrun. A lifelong learning perspective for mobile robot control. In *IROS*, 1995.

[30] Ricardo Vilalta, Christophe Giraud-Carrier, Pavel Brazdil, and Carlos Soares. *Inductive Transfer*, pages 545–548. Springer US, Boston, MA, 2010.

[31] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[32] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017.

[33] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014.

[34] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *ICLR*, 2017.