SMILE: Sample-to-feature MIxup for Efficient Transfer LEarning

Xingjian Li[*]
Big Data Lab, Baidu Research, Baidu Inc.
University of Macau
lixingjian@baidu.com

Haoyi Xiong[†]
Big Data Lab, Baidu Research
Baidu Inc.
xionghaoyi@baidu.com

Chengzhong Xu
State Key Lab of IOTSC
Department of Computer Science
University of Macau
czxu@um.edu.mo

Dengjing Dou
Big Data Lab, Baidu Research
Baidu Inc.
doudejing@baidu.com

Abstract

To improve the performance of deep learning, mixup has been proposed to force the neural networks favoring simple linear behaviors in-between training samples. Performing mixup for transfer learning with pre-trained models however is not that simple, a high capacity pre-trained model with a large fully-connected (FC) layer could easily overfit to the target dataset even with samples-to-labels mixed up. In this work, we propose SMILE— Sample-to-feature Mixup for EffIcient Transfer LEarning. With mixed images as inputs, SMILE regularizes the outputs of CNN feature extractors to learn from the mixed feature vectors of inputs, in addition to the mixed labels. SMILE incorporates a mean teacher to provide the surrogate "ground truth" for mixed feature vectors. Extensive experiments have been done to verify the performance improvement made by SMILE, in comparisons with a wide spectrum of transfer learning algorithms, including fine-tuning, L2-SP, DELTA, BSS, RIFLE, Co-Tuning and RegSL, even with mixup strategies combined.

1 Introduction

Mixup [1] is an effective strategy for improve generalization performance of DNN , where the objective is to have DNNs in the learning procedure favor the *linear behaviors* in-between training samples. To achieve the goal, the mixup strategy picks up multiple images from the training set, mixes the samples and labels proportionally to generate a new pair of sample and label for data augmentation. The regularization effects brought by mixup could help control the complexity of DNN models [2, 3] while largely improving the robustness and generalization performance [4].

As a kind of data augmentation, mixup is naturally expected to be more essential when training data are insufficient. However, we surprisingly find the contrary in the scenario of deep transfer learning: when fine-tuning a pre-trained model with a few target examples, mixup improves the performance with reduced margins or even downgrades the performance when the target dataset is small. See Figure 1 for detailed results. Considering the strong capacity of pre-trained models and the limited training dataset, our research yields the concern that, *can fine-tuning with pre-trained models overfit to the mixed-up samples and labels?*

[*]Xingjian Li and Haoyi Xiong contributed equally.
[†]Correspondence to Haoyi Xiong.

INTERPOLATE Workshop at 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

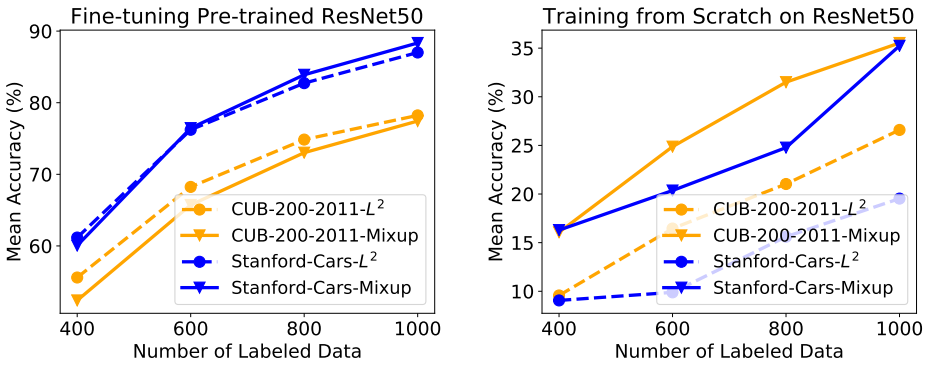

Figure 1: Performance comparison between the L^2 regularization and mixup for fine-tuning an ImageNet pre-trained model (left) and training from scratch (right). To simulate scenarios with limited target datasets, we randomly select 50 classes from two transfer learning benchmarks, which are CUB-200-2011 and Stanford-Cars. As seen, Mixup brings remarkable improvements if training from scratch (right), but this does not apply to transfer learning (left).

Our Observations We find fine-tuning with high capacity pre-trained models CAN overfit to the mixup samples/labels. From mixup, we simply derive a linear interpolation (IL) loss to measure the error of linear interpolation between a pair of samples (x_1, y_1) and (x_2, y_2) for the model $f(\cdot)$,

$$\text{IL}(f) = \mathbb{E}_\lambda[\|f(\lambda x_1 + (1-\lambda)x_2) - (\lambda f(x_1) + (1-\lambda)f(x_2))\|_2^2] \,, \tag{1}$$

where a lower linear interpolation loss indicates stronger linear behaviors in-between the samples and usually better generalization performance [4]. Our experiments however find that fine-tuning with mixup could obtain a low interpolation loss in the training set while suffering a high interpolation loss in the test set ($\geq$25% higher interpolation loss on the testing set than the one on training set, please see also in **Section 4**). This observation indicates that the *linear behaviors* gained by vanilla mixup could not well generalize to the testing dataset and overfit to the mixup samples/labels from the training dataset. These inaccurately interpolated deep features may help less or even harm the learning of the target task. Thus, our research intends to study a way to *make mixup strategies generalizable in deep transfer learning settings while significantly improving the performance of DNNs.*

Our Work To achieve the above goal, in this work, we propose `SMILE`—Sample-to-feature Mixup strategies for Efficient Transfer Learning. One of the major challenges lies in that, unlike the vanilla mixup, there are actually no ground-truth labels for mixed features. To tackle this problem, we introduce two kinds of *pseudo feature labels*, by deeply exploiting the generalizability of the pre-trained model. Specifically, given two samples drawn from the target domain as the input, `SMILE` first linearly combines two samples proportionally and sends the mixed-up sample to the target model. Then, the following two regularizers are employed on the mixed output.

- Mixup on target deep features. It constrains the Euclidean distance between the output of target model's CNN feature extractor and a mixed-up feature vector (i.e., linear combination of a mean teacher model's outputs for the two samples).
- Mixup on adapted source labels. An additional FC classifier for the target network to adapt the target dataset but in the source label domain. It is regularized to learn from the linear combination of classification results given by the teacher classifier.

We carry out extensive experiments using a wide range of source and target datasets. Results show that `SMILE` can outperform a number of state-of-the-art baseline algorithms including L^2-SP [5], DELTA [6], BSS [7], RIFLE [8], Co-Tuning [9] and RegSL [10] with/without vanilla mixup strategies.

2 Related work

The related works to this study include [11, 12] for mixup strategies and [6–9] for transfer learning through regularizing feature space. We here particularly focus on the discussion on the manifold

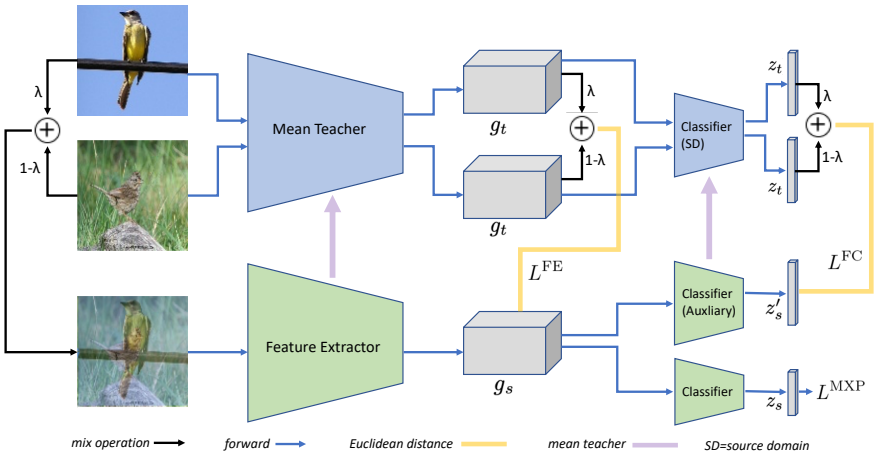

Figure 2: The Architecture of SMILE: Deep Transfer Learning with Sample-to-feature Mixup.

mixup strategy [11] and CutMix [12]. Compared to the *feature-to-label* mixup used in [11], SMILE proposed to a *sample-to-feature* strategy. Though MixCut [12] also uses *sample-to-feature* mixup to spatially fuse multiple images into one to form new feature maps accordingly, SMILE regularizes the CNN feature extractor to learn from a surrogate of "fine-tuned" feature vectors of mixed-up images even when the CNN has not yet been well-tuned in the target domain. Similar ideas have been studied in [13], where authors proposed MixSKD as an alternative approach to the sample-to-feature mixup. Please note that the arXiv version of [13] first appeared in August 2022, while an earlier version [14] of our work was put on arXiv at March 2021.

3 Experiments

We first present the experiment results based on Image Classification tasks on three popular object recognition datasets: CUB-200-2011 [17], Stanford Cars [18] and FGVC-Aircraft [19], which are intensively used in state-of-the-art transfer learning literatures [7–9]. Each of these datasets contains about 6k - 8k training samples. We use ImageNet [20] pre-trained ResNet-50 [21] as the source model. For each dataset, we create four random subsets with different number of categories and training examples.

To further confirm the performance improvement by SMILE is independent with the choice of pre-trained datasets and model architectures, we conduct additional experiments comparing our method with competitive baselines. Specifically, we use the Places365 [22] pre-trained ResNet-50 to perform fine-tuning on MIT-Indoors-67 [23], which is a scene classification task. We also evaluate our method on a more powerful model EfficientNet-B4 [24] designed by NAS over a large scale dataset Food-101 [25]. The descriptions about these benchmarks are summarized in Table 1.

SMILE is compared against multiple state-of-the-art fine-tuning algorithms including L^2 [15], L^2-SP [5], DELTA [6], BSS [7], RIFLE [8], Co-Tuning [9], RegSL [10] and Mixup [1]. As observed in Table 2 and 3, our proposed SMILE achieves remarkable improvements to vanilla fine-tuning on three standard benchmarks, and outperforms all state-of-the-art methods. As the number of training

Table 1: Characteristics of the target tasks.

target dataset	task category	source task	architecture	# training	# classes
CUB-200-2011	Object Recognition	ImageNet	ResNet-50	5,994	200
Stanford-Cars	Object Recognition	ImageNet	ResNet-50	8,144	196
FGVC-Aircraft	Object Recognition	ImageNet	ResNet-50	6,677	100
MIT-Indoor-67	Scene Classification	Places365	ResNet-50	5,356	76
Food-101	Object Recognition	ImageNet	EfficientNet-B4	75,000	101

Table 2: Comparison of top-1 accuracy (%) on transfer learning benchmarks. The notation C:X/N:Y refers to using Y examples from X selected categories.

Dataset	Method	Dataset			
		C:25%/N:400	C:25%/N:800	C:All/N:15%	C:All/N:100%
CUB-200-2011	L^2 [15]	55.59±1.02	74.85±0.12	44.70±0.17	80.64±0.30
	Mixup [16]	52.39±0.68	73.02±0.11	44.27±0.31	81.86±0.20
	L^2-SP [5]	54.38±0.32	73.90±0.22	45.30±0.23	81.58±0.10
	DELTA [6]	58.15±0.26	75.84±0.08	47.88±0.15	82.21±0.15
	BSS [7]	54.99±0.73	74.14±0.34	46.41±0.09	81.10±0.04
	RIFLE [8]	53.68±0.89	73.05±1.09	44.13±0.38	81.94±0.06
	Co-Tuning [9]	57.98±0.08	75.11±0.47	49.98±0.23	82.60±0.03
	RegSL [10]	57.62±0.88	75.51±0.44	46.92±0.28	80.20±0.17
	SMILE	**62.13±0.55**	**77.27±0.35**	**51.73±0.04**	**83.62±0.07**
Stanford-Cars	L^2 [15]	61.17±0.36	82.73±0.59	43.01±0.53	90.14±0.12
	Mixup [16]	60.25±0.68	83.60±0.02	45.73±0.15	91.51±0.18
	L^2-SP [5]	61.00±0.28	82.05±0.05	44.12±0.33	90.61±0.12
	DELTA [6]	62.05±0.13	82.1±0.44	43.27±0.27	90.86±0.08
	BSS [7]	64.97±0.69	83.81±0.39	47.45±0.23	91.14±0.04
	RIFLE [8]	62.85±0.22	83.57±0.43	43.61±0.07	91.08±0.12
	Co-Tuning [9]	**66.05±0.41**	81.05±0.39	44.29±0.42	91.19±0.11
	RegSL [10]	60.12±0.63	82.91±0.08	42.52±0.37	91.02±0.05
	SMILE	65.17±1.11	**85.90±0.16**	**50.93±0.17**	**92.21±0.05**
FGVC-Aircraft	L^2 [15]	59.63±1.11	79.57±0.18	51.13±0.45	88.27±0.51
	Mixup [16]	65.20±0.80	**84.53±0.62**	54.42±0.55	89.33±0.17
	L^2-SP [5]	54.70±0.73	76.13±0.82	48.85±0.70	87.97±0.66
	DELTA [6]	53.47±0.24	71.73±1.02	51.05±0.38	88.92±0.25
	BSS [7]	61.40±1.13	81.47±0.24	52.61±0.11	88.47±0.16
	RIFLE [8]	60.97±0.49	79.87±0.38	52.13±0.31	89.45±0.44
	Co-Tuning [9]	62.98±0.72	80.03±0.04	52.05±0.43	88.19±0.33
	RegSL [10]	61.87±0.37	79.40±0.92	51.64±0.43	88.87±0.26
	SMILE	**68.40±0.33**	**84.57±0.29**	**60.04±0.33**	**90.16±0.15**

Table 3: Comparison of top-1 accuracy (%) with different transfer learning algorithms on more task types and architectures.

Dataset	Method	Sampling Rates		
		30%	50%	100%
MIT-Indoor-67	L^2 [15]	78.68±0.20	80.80±0.18	82.00±0.21
	Mixup [16]	77.44±0.44	80.28±0.28	82.87±0.50
	DELTA [6]	80.80±0.22	82.80±0.25	83.67±0.18
	BSS [7]	78.23±0.50	80.35±0.28	82.15±0.22
	RIFLE [8]	76.76±0.08	78.71±0.33	81.78±0.07
	SMILE	**82.00±0.14**	**83.54±0.20**	**85.37±0.16**
Food-101	L^2 [15]	80.25±0.28	83.43±0.15	86.77±0.03
	Mixup [16]	82.63±0.11	84.93±0.06	87.82±0.06
	DELTA [6]	81.38±0.08	84.07±0.06	87.34±0.07
	BSS [7]	81.13±0.04	83.96±0.09	87.33±0.03
	RIFLE [8]	81.13±0.04	83.82±0.02	87.29±0.11
	SMILE	**82.84±0.16**	**85.25±0.09**	**88.20±0.10**

Table 4: Feature-IL and Label-IL for different fine-tuning methods over the training (sampling the CUB-200-2011 training set by 30%) and testing dataset. Lower is better. Add. Data refers to involving the remaining 70% training examples for fine-tuning. However, the interpolation loss for the training set is still calculated on the original 30%.

Method	Label-IL		Feature-IL	
	Train	Test	Train	Test
Finetune	1.80	1.92	1.92	1.93
Finetune + Add. Data	1.85	1.88	1.58	1.63
Finetune + MXP	**1.65**	2.00	1.98	2.02
SMILE	1.75	**1.82**	**1.48**	**1.53**

examples becomes smaller, our method yields more significant benefits, e.g. SMILE outperforms vanilla fine-tuning by more than 8% on FGVC-Aircraft when 15% training samples are used.

4 Linear Interpolation Effects and Generalization

We use Eq 1 to measure Label-IL using the classifier outputs and Feature-IL using the last hidden layer of ResNet-50 for different transfer learning methods, with CUB-200-2011 (with 30% sampling rates) as the training set for all methods. Several arguments can be deduced from results in Table 4.

(1) More Data, Better Generalization, and Lower Label-IL and Feature-IL. There is no doubt to assume that, in practice, a model trained with more data should enjoy better generalization performance. In addition to improve the testing accuracy , we find that, when we involve additional training samples, both Label-IL and Feature-IL would be lower on the testing sets, compared to vanilla fine-tuning.

(2 Fine-tuning with vanilla mixup is NOT generalizable even in the label space, due to the lack of linear interpolation in feature spaces. As shown in Table 4, although Label-IL of the vanilla mixup is significantly lower on the training set than other methods, its Label-IL is high on the testing set (not generalizable). Furthermore, compared to other methods on both training/testing sets (even Fine-tuning on the testing set), Feature-IL of the vanilla mixup is high, i.e., poor linear interpolation in feature spaces.

(3) Sample-to-Feature Mixup could ensure the generalizability of mixup effects in the label space, as SMILE is with low Feature-IL and Label-IL on both training and testing sets. While SMILE achieves the lowest Feature-IL on both training and testing datasets, it also achieves the lowest testing Label-IL. The comparisons with vanilla mixup suggest that doing mixup in the label space is just not enough for fine-tuning.

These arguments solidify our motivation of sample-to-feature mixup for fine-tuning.[3]

5 Conclusion

In this work, we figure out the difficulty of applying mixup in transfer learning, and introduce SMILE—Sample-to-feature Mixup strategies for Efficient Transfer Learning. Beyond a direct combination of fine-tuning and mixup, SMILE pursues generalizable linear behaviors through both features of the target domain and the label space of the source domain. We conduct extensive experiments using a wide spectrum of target datasets. Results show that SMILE can significantly promote the effectiveness of fine-tuning and outperform various competitive fine-tuning algorithms. Ablation studies and empirical discussions further backup our design intuition and purposes.

[3]Note that the over-fitting of linear behaviors may not be directly calibrated with training/test accuracy as there exists other factors influence the accuracy, e.g. mixup also benefits from the effect of label smoothing [26].

References

[1] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[2] B. Hanin and D. Rolnick, "Complexity of linear regions in deep networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2596–2604.

[3] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[4] L. Zhang, Z. Deng, K. Kawaguchi, A. Ghorbani, and J. Zou, "How does mixup help with robustness and generalization?" *arXiv preprint arXiv:2010.04819*, 2020.

[5] X. Li, Y. Grandvalet, and F. Davoine, "Explicit inductive bias for transfer learning with convolutional networks," *Thirty-fifth International Conference on Machine Learning*, 2018.

[6] X. Li, H. Xiong, H. Wang, Y. Rao, L. Liu, and J. Huan, "Delta: Deep learning transfer using feature map with attention for convolutional networks," *arXiv preprint arXiv:1901.09229*, 2019.

[7] X. Chen, S. Wang, B. Fu, M. Long, and J. Wang, "Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 1906–1916.

[8] X. Li, H. Xiong, H. An, C.-Z. Xu, and D. Dou, "Rifle: Backpropagation in depth for deep transfer learning through re-initializing the fully-connected layer," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6010–6019.

[9] K. You, Z. Kou, M. Long, and J. Wang, "Co-tuning for transfer learning," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[10] D. Li and H. Zhang, "Improved regularization and robustness for fine-tuning in neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 249–27 262, 2021.

[11] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6438–6447.

[12] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023–6032.

[13] C. Yang, Z. An, H. Zhou, L. Cai, X. Zhi, J. Wu, Y. Xu, and Q. Zhang, "Mixskd: Self-knowledge distillation from mixup for image recognition," in *European Conference on Computer Vision*, 2022.

[14] X. Li, H. Xiong, C. Xu, and D. Dou, "SMILE: self-distilled mixup for efficient transfer learning," *CoRR*, vol. abs/2103.13941, 2021. [Online]. Available: https://arxiv.org/abs/2103.13941

[15] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning*, 2014, pp. 647–655.

[16] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[17] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[18] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[19] S. Maji, E. Rahtu, J. Kannala, M. B. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *ArXiv*, vol. abs/1306.5151, 2013.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[23] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 413–420.

[24] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[25] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.

[26] A. Singh and A. Bay, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks neurips reproducibility challenge 2019," 2019.