Charting the Right Manifold: Manifold Mixup for Few-Shot Learning

Puneet Mangla[1,2*], Mayank Singh[1*], Abhishek Sinha[1*], Nupur Kumari[1*], Balaji Krishnamurthy[1], Vineeth N Balasubramanian[2]

[1]Media and Data Science Research, Adobe Inc. Noida, INDIA
[2]Indian Institute of Technology, Hyderabad, INDIA

13 Dec 2019
MetaLearn Workshop, NeurIPS 2019

[*] Authors contributed equally

Few-Shot Learning

The model is trained on a set of classes (base classes) with abundant examples in a fashion that promotes the model to classify unseen classes (novel classes) using few labeled instances

Existing Approaches

- **Meta-learning based methods:**
  aim to learn an optimizer or a good model initialization that can adapt for novel classes in few gradient steps and limited labelled examples. E.g. Ravi & Larochelle, 2017; Andrychowicz, Marcin, et al. 2016; Finn et al. 2017

- **Distance metric based methods:**
  leverage the information about similarity between images. E.g. Vinyals, Oriol, et al. 2016; Snell, J. et al. 2017

- **Hallucination based methods:**
  augment the limited training data for the new task by generating or hallucinating new data points. E.g. Hariharan, B., & Girshick, R. 2017; Wang, Yu-Xiong, et al. 2018

Key Contributions

- We observe that applying Manifold Mixup (Verma, V, et al. 2018) regularization over the feature manifold enriched via rotation self-supervision task of (Gidaris, S. et al. 2018) significantly improves the performance in few-shot tasks in comparison with Baseline++ (Wei-Yu Chen et al. 2019).

- The proposed methodology outperforms state-of-the-art methods by 3-8% over CIFAR-FS, CUB and mini-ImageNet datasets.

- We show that the improvements made by our methodology become more pronounced in the cross-domain few-shot task evaluation and on increasing N from standard value of 5 in the N-way K-shot evaluation.

Manifold Mixup (Verma, V, et al. 2018)

leverages linear interpolations in hidden layers of neural network to help the trained model generalize better.

$$L_{mm} = \mathbb{E}_{(x,y)\in\mathcal{D}_b}\left[L\Big(Mix_\lambda(f_\theta^l(\mathbf{x}), f_\theta^l(\mathbf{x}')), Mix_\lambda(y, y')\Big)\right]$$

where

$$Mix_\lambda(a, b) = \lambda \cdot a + (1 - \lambda) \cdot b$$

D_b is the training data and λ is sampled from a β(α,α) distribution and L is standard cross entropy loss

Rotation Self-Supervision (Gidaris, S. et al. 2018)

The input image is rotated, and the auxiliary task of the model is to predict the rotation. Training loss is $L_{rot} + L_{class}$

$$L_{rot} = \frac{1}{|C_R|} * \sum_{\mathbf{x} \in \mathcal{D}_b} \sum_{r \in C_R} L(c_{W_r}(f_\theta(\mathbf{x}^r)), r)$$

$$L_{class} = \mathbb{E}_{(x,y) \in \mathcal{D}_b, r \in C_R} \left[L(x^r, y) \right]$$

D_b is the training data; $|C_R|$ is the number of rotated images; c_{W_r} is a 4-way linear classifier

Proposed Method: $S2M2_R$

1. Self-supervised training: train with rotation self-supervision as an auxiliary task

2. Fine-tuning with Manifold Mixup: fine-tune the above model with Manifold-Mixup for a few more epochs i.e. $L = L_{mm} + 0.5(L_{rot} + L_{class})$

After obtaining the backbone, a cosine classifier is learned over the feature representation of novel classes for each few-shot task.

Comparison with prior state-of-the-art methods

Method	*mini*-Imagenet		CUB		CIFAR-FS	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
MAML [2]	54.69 ± 0.89	66.62 ± 0.83	71.29 ± 0.95	80.33 ± 0.70	58.9 ± 1.9	71.5 ± 1.0
ProtoNet [5]	54.16 ± 0.82	73.68 ± 0.65	71.88 ± 0.91	87.42 ± 0.48	55.5 ± 0.7	72.0 ± 0.6
RelationNet [6]	52.19 ± 0.83	70.20 ± 0.66	68.65 ± 0.91	81.12 ± 0.63	55.0 ± 1.0	69.3 ± 0.8
LEO [4]	61.76 ± 0.08	77.59 ± 0.12	68.22 ± 0.22	78.27 ± 0.16	-	-
DCO [3]	62.64 ± 0.61	78.63 ± 0.46	-	-	72.0 ± 0.7	84.2 ± 0.5
Baseline++ [1]*	57.53 ± 0.10	72.99 ± 0.43	70.4 ± 0.81	82.92 ± 0.78	67.50 ± 0.64	80.08 ± 0.32
Manifold Mixup	57.16 ± 0.17	75.89 ± 0.13	73.47 ± 0.89	85.42 ± 0.53	69.20 ± 0.2	83.42 ± 0.15
Rotation	63.9 ± 0.18	81.03 ± 0.11	77.61 ± 0.86	89.32 ± 0.46	70.66 ± 0.2	84.15 ± 0.14
$S2M2_R$	$\mathbf{64.93 \pm 0.18}$	$\mathbf{83.18 \pm 0.11}$	$\mathbf{73.71 \pm 0.22}$	$\mathbf{88.59 \pm 0.14}$	$\mathbf{74.81 \pm 0.19}$	$\mathbf{87.47 \pm 0.13}$

*denotes our implementation

Effect of Varying N in N-way K-shot Evaluation

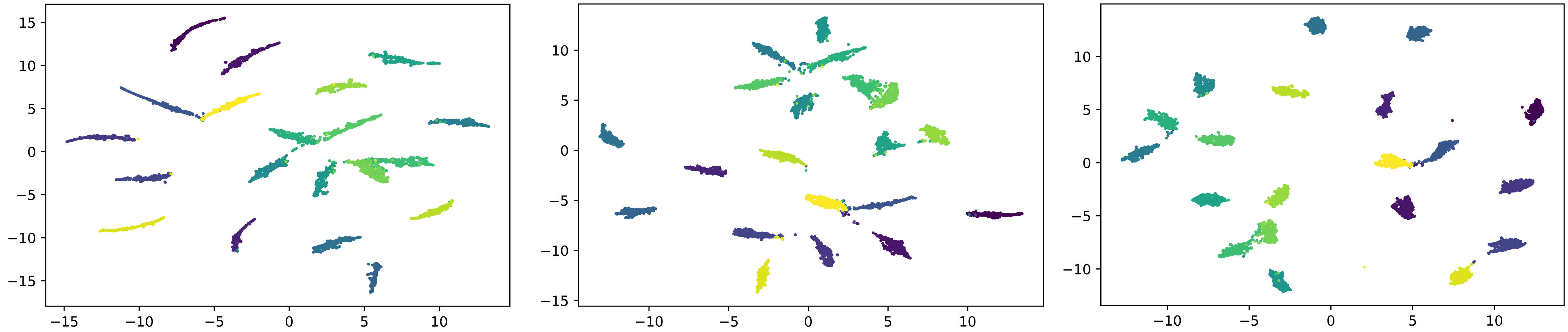
Method	10-way		15-way		20-way	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Baseline++*	40.43	56.89	31.96	48.2	26.92	42.8
LEO*	45.26	64.36	36.74	56.26	31.42	50.48
DCO*	44.83	64.49	36.88	57.04	31.5	51.25
Manifold Mixup	42.46	62.48	34.32	54.9	29.24	48.74
Rotation	47.77	67.2	38.4	59.59	33.21	54.16
$S2M2_R$	**50.4**	**70.93**	**41.65**	**63.32**	**36.5**	**58.36**

*denotes our implementation

Cross Domain Few-Shot Learning

Method	*mini*-Imagenet ⟹ CUB	
	1-Shot	5-Shot
DCO	44.79 ± 0.75	64.98 ± 0.68
Baseline++	40.44 ± 0.75	56.64 ± 0.72
Manifold Mixup	46.21 ± 0.77	66.03 ± 0.71
Rotation	$\mathbf{48.42} \pm \mathbf{0.84}$	68.40 ± 0.75
$S2M2_R$	48.24 ± 0.84	$\mathbf{70.44} \pm \mathbf{0.75}$

Visualization of Feature Representations



UMAP (McInnes, L. et al. 2018) 2-dim plot of feature vectors of novel classes in mini-Imagenet dataset using Baseline++, Rotation, $S2M2_R$ (left to right)

Summary

- learning feature representation with relevant regularization and self-supervision techniques lead to consistent improvement in few-shot learning tasks on a diverse set of image classification datasets.

- feature representation learning using both self-supervision and classification loss and then applying Manifold-mixup over it, outperforms prior state-of-the-art approaches in few-shot learning.

Thank You!
Questions?

Code: https://github.com/nupurkmr9/S2M2_fewshot

kbalaji@adobe.com; vineethnb@iith.ac.in

References

1. W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Wang, and J.-B. Huang.A closer look at few-shot classification. InInternationalConference on Learning Representations, 2019. W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Wang, and J.-B. Huang.A closer look at few-shot classification. In InternationalConference on Learning Representations, 2019.
2. C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. InProceedingsof the 34th International Conference on Machine Learning-Volume 70, pages 1126–1135. JMLR. org, 2017.
3. K. Lee, S. Maji, A. Ravichandran, and S. Soatto. Meta-learning with differentiable convex optimization.CoRR,abs/1904.03758, 2019.
4. A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu,S. Osindero, and R. Hadsell. Meta-learning with latent em-bedding optimization. InInternational Conference on Learn-ing Representations, 2019.
5. J. Snell, K. Swersky, and R. Zemel. Prototypical networksfor few-shot learning. InAdvances in Neural InformationProcessing Systems, pages 4077–4087, 2017.
6. F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, andT. M. Hospedales. Learning to compare: Relation networkfor few-shot learning.CoRR, abs/1711.06025, 2017.