Supplementary Materials: Exploring Complementary Strengths of Invariant and Equivariant Representations for Few-Shot Learning

Anonymous CVPR 2021 submission

Paper ID 10643

1. Overview

In the supplementary materials we include the following: additional details about the applied geometric transformations (Section 2), additional results with the transformations sampled from the complete space of affine transformations (Section 3), ablation study on the coefficient of inductive loss (Section 4), ablation study on the temperature of knowledge distillation (Section 5), effect of successive self knowledge distillation (Section 6), and effect of enforcing invariance and equivariance for supervised classification (Section 7).

2. Geometric Transformations

For our geometric transformations, we sample from a complete space of similarity transformation and use four rotation transformations: $\{0°, 90°, 180°, 270°\}$, two scaling transformations: $\{0.67, 1.0\}$ and three aspect ratio transformations: $\{0.67, 1.0, 1.33\}$. Different combinations of these transformations lead to different values of M (total number of applied transformations). An ablation study on the value of M is included in section 4.2 of the main paper. In Table 1 we include the complete description of different values of M that we use in our experiments.

3. Additonal Resutls with Affine Transformations

We perform a set of experiments where the objective is to sample geometric transformation from the complete space of affine transformations. To this end, we quantize the affine transformation space according to Table 2. This leads to 972 distinct geometric transformations. Since it's not feasible to apply all the 972 transformations on an input image $\mathbf{x}$ to obtain the input tensor $\mathbf{x}_{all} = \{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{971}\}$, we randomly sample 10 geometric transformations from the set of 972 transformations. We apply these randomly sampled 10 geometric transformations on an input image $\mathbf{x}$ and generate the input tensor $\mathbf{x}_{all}$. The results of these experiments are presented in Table 3. From Table 3 it's evident that training with either invariance or equivariance improves over the baseline training for both 1 and 5 shot tasks (2.5-3.7% improvement). Joint optimization for both invariance and equivariance provides additional improvement of $\sim 1\%$. Even though the experiments with geometric transformations sampled from the complete affine transformation space do not improve over the training with $M = 16$ (description of $M = 16$ is available in Table 1), the experiments demonstrate consistent improvement when both invariance and equivariance are enforced simultaneously. This provides additional support for our claim that enforcing both invariance and equivariance is beneficial for learning good general representations for solving challenging FSL tasks.

4. Ablation Study for Coefficient of Inductive Loss

We conduct an ablation study to measure the effect of different values of the coefficient of inductive loss (without multi-head distillation) on the CIFAR-FS [1] validation set; the results of 5-way 1-shot FSL tasks are presented in fig. 1. From fig. 1 it is evident that the proposed method is fairly robust to the different values of the coefficient of the inductive loss. However, the best performance is obtained when we set the loss coefficient to 1.0. Based on this ablation study, we use a loss coefficient of 1.0 for the inductive loss in all of our experiments.

5. Ablation Study for Knowledge Distillation Temperature

To analyse the effect of knowledge distillation temperature (for Kullback Leibler (KL) divergence losses) we conduct an ablation study on the validation set of CIFAR-FS [1] dataset. From fig. 2 we can observe that the proposed method with multi-head distillation objective is not very sensitive to the temperature coefficient of knowledge distillation. The proposed method achieves similar performance on the CIFAR-FS validation set when the value of distillation temperature is set to 4.0 and 5.0. Based on this ablation

M	Description
3	AR: $\{0.67, 1.0, 1.33\}$
4	ROT: $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$
8	ROT: $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\} \times$ S: $\{0.67, 1.0\}$
12	AR: $\{0.67, 1.0, 1.33\} \times$ ROT: $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$
16	$\big($AR: $\{0.67, 1.0, 1.33\} \times$ ROT: $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}\big) \bigcup \big($ROT: $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\} \times$ S: $\{0.67\}\big)$
20	$\big($AR: $\{0.67, 1.0, 1.33\} \times$ ROT: $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}\big) \bigcup \big($ROT: $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\} \times$ S: $\{0.67\} \times$ AR: $\{0.67, 1.33\}\big)$
24	AR: $\{0.67, 1.0, 1.33\} \times$ ROT: $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\} \times$ S: $\{0.67, 1.0\}$

Table 1. Complete description of different values of M based on different combination of aspect ratio (AR), rotation (ROT), and scaling (S) transformations.

Transformation	**Quantized Values**
Rotation	$\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$
Translation$_X$	$\{-0.2, 0.0, 0.2\}$
Translation$_Y$	$\{-0.2, 0.0, 0.2\}$
Scale	$\{0.67, 1.0, 1.33\}$
Aspect-Ratio	$\{0.67, 1.0, 1.33\}$
Shear	$\{-20^\circ, 0^\circ, 20^\circ\}$

Table 2. Quantization of the space of Affine transformations.

Method	**1-Shot**	**5-Shot**
Baseline Training	62.02 ± 0.63	79.64 ± 0.44
Ours with only Invar (affine)	65.55 ± 0.81	82.17 ± 0.52
Ours with only Equi (affine)	65.70 ± 0.79	82.47 ± 0.53
Ours with Equi and Invar (affine)	66.82 ± 0.79	82.96 ± 0.53
Ours with Equi and Invar (M=16)	66.82 ± 0.80	84.35 ± 0.51

Table 3. Average 5-way few-shot classification accuracy with 95% confidence intervals on **miniImageNet** dataset; trained with different geometric transformations.

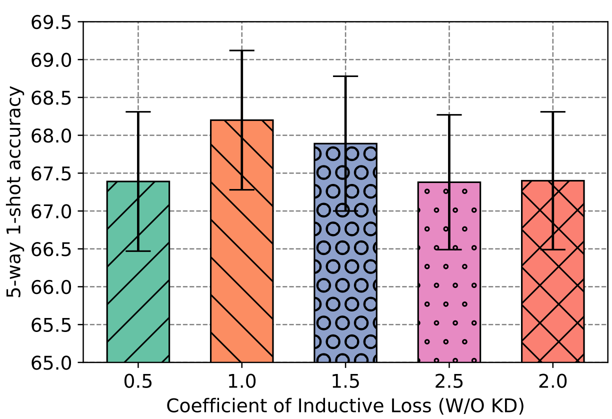

Figure 1. Ablation study on **CIFAR-FS** validation set with different coefficients of the inductive loss (W/O KD); the reported score is average 5-way 1-shot classification accuracy with 95% confidence intervals.

study and to be consistent with [4], we set the value of the coefficient of knowledge distillation temperature to 4.0 in all of our experiments.

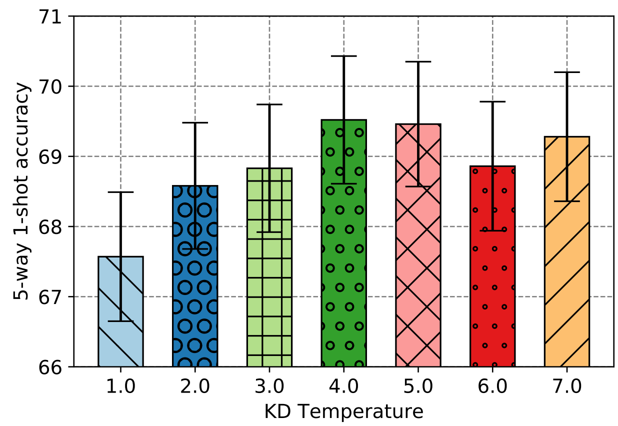

Figure 2. Ablation study on **CIFAR-FS** validation set with different values of knowledge distillation temperature; the reported score is average 5-way 1-shot classification accuracy with 95% confidence intervals.

6. Effect of Successive Distillation

In all of our experiments, we use only one stage of multi-head knowledge distillation. To further investigate the effect of knowledge distillation we perform multiple stages of self knowledge distillation on CIFAR-FS [1] dataset. The results are presented in fig. 3. Here, the 0^{th} distillation stage is the base learner trained with only the supervised baseline loss ($\mathcal{L}_{baseline}$), equivariant loss ($\mathcal{L}_{eq}$), and invariant loss ($\mathcal{L}_{in}$). From fig. 3, we observe that the performance in the FSL task improves for the first 2 stages of distillation, after that the performance saturates. Besides, the improvement from stage 1 to stage 2 is minimal ($\sim 0.1\%$). Therefore, to make the proposed method more computationally efficient we perform only one stage of distillation in all of our experiments.

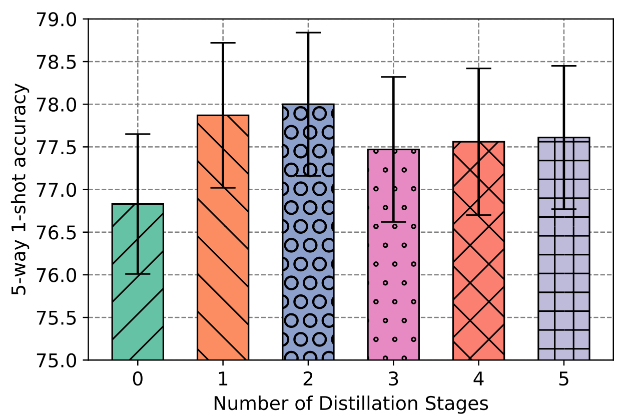

Figure 3. Evaluation of different knowledge distillation stages on **CIFAR-FS** dataset; the reported score is average 5-way 1-shot classification accuracy with 95% confidence intervals.

7. Invariance and Equivariance for Supervised Classification

To demonstrate the effectiveness of complementary strengths of invariant and equivariant representations we conduct fully supervised classification experiments on benchmark CIFAR-100 dataset [2]. For these experiments, we use the standard Wide-Resnet-28-10 [6] architecture as the backbone. For training, we use an SGD optimizer with an initial learning rate of 0.1. We set the momentum to 0.9 and use a weight decay of 5e$-$4. For all the experiments, the training is performed for 200 epochs where the learning rate is decayed by a factor of 5 at epochs 60, 120, and 160. We use a batch size of 128 for all the experiments as well as a dropout rate of 0.3. The training augmentations include standard data augmentations: random crop and random horizontal flip. For enforcing invariance and equivariance, we set the value of M to 12 for computational efficiency; description of $M = 12$ is available in Table 1. We do not perform knowledge distillation for these experiments. The results of these experiments are presented in Table 4.

From Table 4, we can notice that enforcing invariance provides little improvement (0.2%) over the supervised baseline. This is expected since the train and test data is coming from the same distribution and same set of classes; making the class boundaries compact (for seen classes) doesn't provide that much additional benefit. However, in the case of FSL we observe that enforcing invariance over baseline provides 2.62%, 2%, and 3.5% improvement for miniImageNet [5], CIFAR-FS [1], and FC100 [3] datasets respectively (section 4.2 of main text). On the other hand, enforcing equivariance for supervised classification provides better improvement (1.8%) since it helps the model to better learn the structure of data. Even though enforcing equivariance provides noticeable improvement for supervised classification, in the case of FSL we obtain a much bigger improvement of 4.07%, 4.87%, and 4.13% for miniImageNet [5], CIFAR-FS [1], and FC100 [3] datasets respectively (section 4.2 of main text). Finally, joint optimization for both invariance and equivariance achieves the best performance and provides minimal but consistent improvement of 0.1% over enforcing only equivariance. However, joint optimization provides a much larger improvement on FSL tasks (see section 4.2 of the main text). From these experiments, we conclude that, although enforcing both invariance and equivariance is beneficial for supervised classification, injecting these inductive biases is more crucial for FSL tasks since the inductive inference for FSL tasks is more challenging (inference on unseen/novel classes).

Method	Error Rate (%)
Supervised Baseline	18.78
Ours with only Invariance	18.56
Ours with only Equivariance	16.95
Ours with Equi and Invar (W/O KD)	16.84

Table 4. Results with invariance and equivariance for supervised classification on **CIFAR-100** dataset.

References

[1] Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. 1, 2, 3

[2] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). 3

[3] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731, 2018. 3

[4] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020. 2

[5] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3630–3638. Curran Associates, Inc., 2016. 3

[6] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016. 3