Unsupervised Embedding Adaptation via Early-Stage Feature Reconstruction for Few-Shot Classification

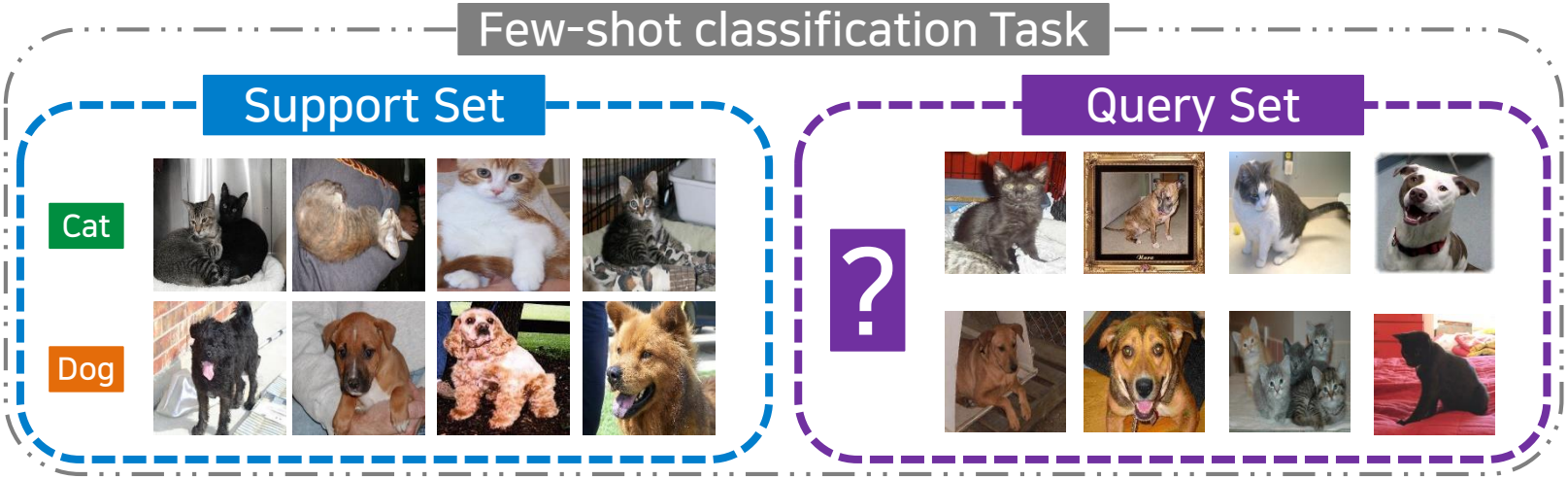
Dong Hoon Lee Sae-Young Chung

Korea Advanced Institute of Science and Technology (KAIST)

ICML 2021

Background. Few-shot classification problem

- Few-shot image classification problem
  - A small labeled support set (S) and unlabeled query set (Q)
  - Goal: classify query samples by few examples in the support set.



+ Transductive setting
  - Allow to utilize all the unlabeled query samples together to make an inference.

Q. Can we leverage (deep) unsupervised learning for few-shot classification?

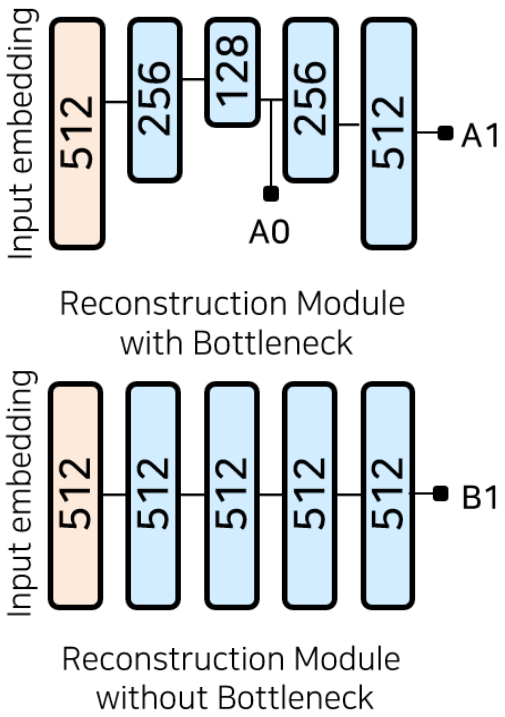
Method – 1. Feature reconstruction

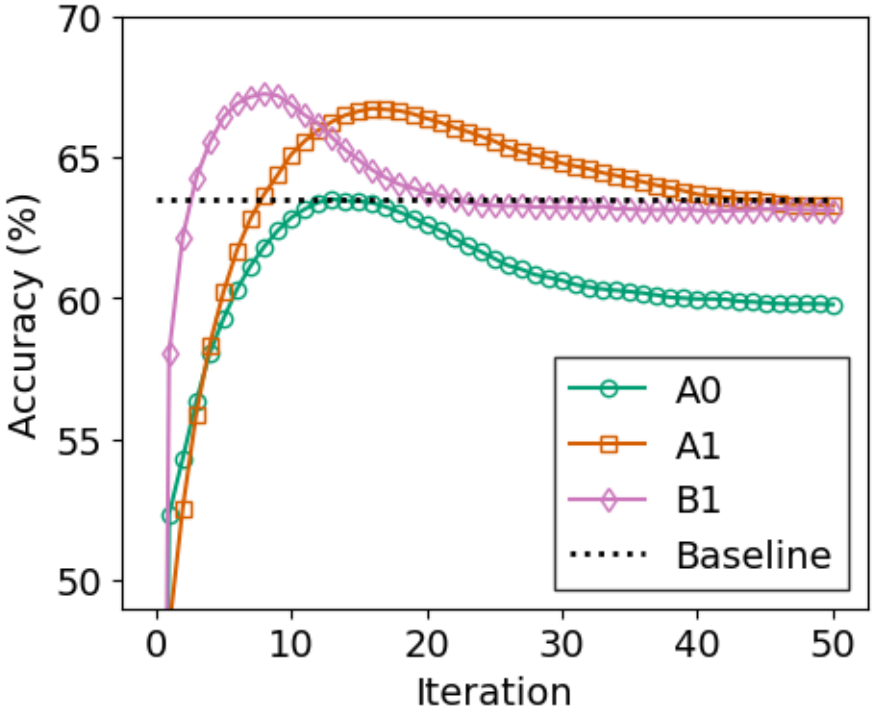
- We study how unsupervised learning can contribute to few-shot classification.
- Unsupervised learning: Feature reconstruction

$$\mathcal{L}_{\mathrm{FR}} = \frac{1}{|S \cup Q|} \sum_{z \in S \cup Q} d_{\cos}(z, g_\phi(z))$$

g_ϕ : is a reconstruction module (4-layer fully connected NN)

- We specifically focus on "embedding adaptation"





Method – 1. Feature reconstruction

- We study how unsupervised learning can contribute to few-shot classification.
- Unsupervised learning: Feature reconstruction

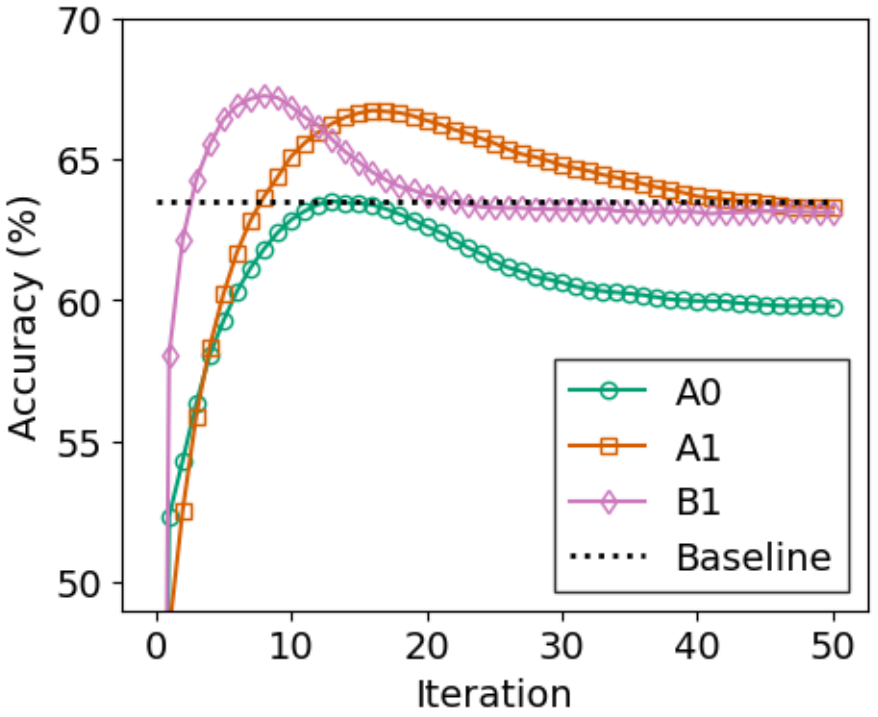
$$\mathcal{L}_{\mathrm{FR}} = \frac{1}{|S \cup Q|} \sum_{z \in S \cup Q} d_{\cos}(z, g_\phi(z))$$

g_ϕ : is a reconstruction module (4-layer fully connected NN)

  - We specifically focus on "embedding adaptation"

1. The figure shows an interesting behavior that the accuracies with new embeddings initially increase then decrease.

2. The peak accuracy of B1 exceeds the baseline accuracy of the original embedding.



Method – 2. LID based early stopping

- Cause of the behavior?
  - Can be explained with the property of DNN training[1]

    "DNNs learn to generalize before memorizing"

  => Early retained generalizable features are more likely to be task-relevant in classification.

- Local Intrinsic Dimensionality (LID) based early stopping
  - Based on our hypothesis and prior works[2], we propose to use LID as the early stopping criteria of our method.

$$\widehat{\mathrm{LID}}(\phi) = -\sum_{z \in S \cup Q} \left[\frac{1}{m} \sum_{i=1}^{m} \ln \frac{r_i(g_\phi^{L-2}(z))}{r_m(g_\phi^{L-2}(z))} \right]^{-1}$$

g_ϕ^{L-2}: is the hidden representation of the second-to-last layer of g_ϕ

$r_i(g_\phi^{L-2}(z))$: distance between $g_\phi^{L-2}(z)$ and its i-th nearest neighbor

[1]Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A. C., Bengio, Y., and Lacoste-Julien, S. A closer look at memorization in deep networks. In ICML, 2017.
[2]Ma, X., Wang, Y., Houle, M. E., Zhou, S., Erfani, S. M., Xia, S., Wijewickrema, S. N. R., and Bailey, J. Dimensionality-driven learning with noisy labels. In ICML, 2018.

Method – 2. LID based early stopping

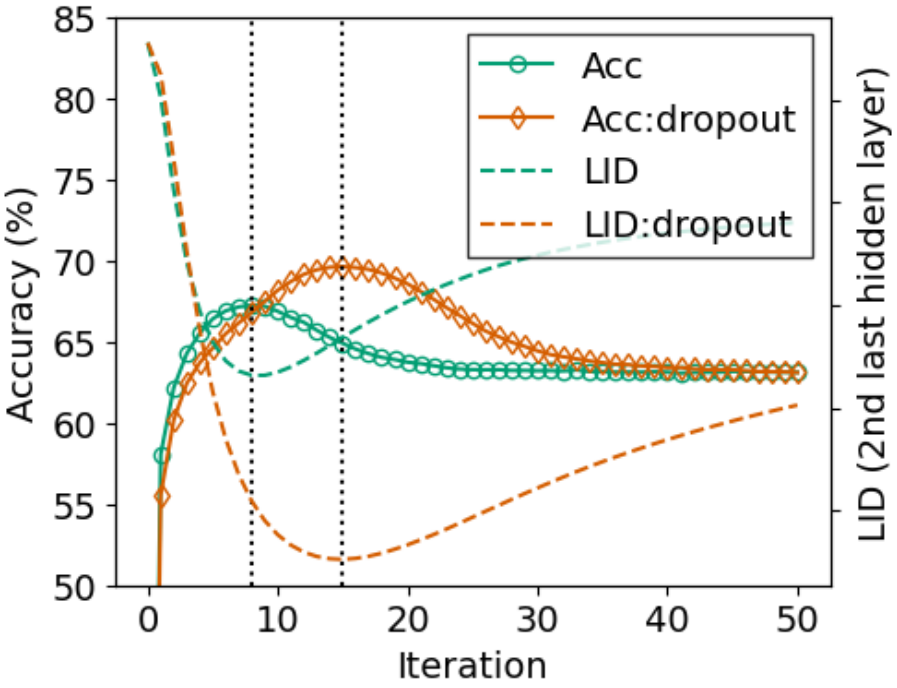
- Local Intrinsic Dimensionality (LID) based early stopping
  - Based on our hypothesis and prior works, we propose to use LID as the early stopping criteria of our method.

$$\widehat{\mathrm{LID}}(\phi) = -\sum_{z \in S \cup Q} \left[\frac{1}{m} \sum_{i=1}^{m} \ln \frac{r_i(g_\phi^{L-2}(z))}{r_m(g_\phi^{L-2}(z))} \right]^{-1}$$

g_ϕ^{L-2}: is the hidden representation of the second-to-last layer of g_ϕ

$r_i(g_\phi^{L-2}(z))$: distance between $g_\phi^{L-2}(z)$ and its i-th nearest neighbor

- We experimented the relationship between the LID and accuracy during reconstruction training.
- We find that LID can be used to find the early stopping time of the best possible new embeddings.
  - Early stop when LID started to raise.



Method. ESFR

- We propose Early-Stage Feature Reconstruction (ESFR) method that finds task-adapted embeddings.
  - Use the observed behavior that *"Early retained features are more generalizable."*
  - Consists of (1) Feature reconstruction training + (2) LID based early stopping

Algorithm 1 ESFR

Input: embedding support set S_f, embedding query set Q_f, and few-shot classifier Alg : $S_f, Q_f \rightarrow \widehat{Y}_Q$
Initialize: $\phi^{i=1:N_e}$
for $i = 1$ **to** N_e **do**
  prev_lid $= \widehat{\text{LID}}(\phi_0^i)$
  Initialize: optimizer
  for $j = 0$ **to** MAX_ITERATION **do**
    $\phi_{j+1}^i \leftarrow \phi_j^i - \nabla_{\phi_j^i}\mathcal{L}(\phi_j^i)$ from equation 5 or 7
    lid $= \widehat{\text{LID}}(\phi_{j+1}^i)$
    if lid $>$ prev_lid **then**
      $\phi_*^i = \phi_{j+1}^i$
      break
    end if
    prev_lid $=$ lid
  end for
end for
$S^{\text{ESFR}} = \{(z', y) | z' = \frac{1}{N_e}\sum_{i=1}^{N_e} g_{\phi_*^i}(z), (z, y) \in S_f\}$
$Q^{\text{ESFR}} = \{z' | z' = \frac{1}{N_e}\sum_{i=1}^{N_e} g_{\phi_*^i}(z), z \in Q_f\}$
Output: $\widehat{Y}_Q = \text{Alg}(S^{\text{ESFR}}, Q^{\text{ESFR}})$

1 Dropout perturbation
: based on our hypothesis

$$\mathcal{L}_{\text{FR}}(\phi) = \frac{1}{|S \cup Q|}\sum_{z \in S \cup Q}\mathbb{E}[d_{\cos}(z, g_\phi(z \odot \mu))]$$

2 Embedding ensemble
: to reduce the variance by random initialization

To make our method solid: 1 2

ESFR is used as a plug and play module

Experiment. Improvement by ESFR

- ESFR consistently improves baseline few-shot classification methods in all settings
  - Methods (Linear, NN, BD-CSPN†), various datasets (mini-/tiered-ImageNet, and CUB), backbones (ResNet18/WidResNet/Conv), settings (1- and 5-shot)
  - ESFR can offer a complementary improvement to semi-supervised approaches.

		mini-ImageNet		*tiered*-ImageNet	
Backbone	**Method**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
ResNet-18	Linear	62.45	79.32	68.49	83.77
	+ **ESFR**	**70.38**+7.93	**81.6**+2.28	**76.98**+8.49	**86.09**+2.32
	NN	64.04	79.71	71.60	84.62
	+ **ESFR**	**70.94**+6.9	**81.61**+1.9	**77.44**+5.84	**85.84**+1.22
	BD-CSPN†	70.00	**82.36**	77.28	**86.55**
	+ **ESFR**	**73.98**+3.98	82.32-0.04	**80.13**+2.85	86.34-0.21
	+ **ESFR-Semi**		**82.89**+0.53		**86.83**+0.28
WRN-28-10	Linear	64.53	80.81	69.78	84.91
	+ **ESFR**	**73.33**+8.8	**83.65**+2.84	**78.57**+8.79	**87.37**+2.46
	NN	66.73	81.85	72.97	85.74
	+ **ESFR**	**74.01**+7.28	**83.58**+1.73	**79.13**+6.16	**87.08**+1.34
	BD-CSPN†	72.74	84.14	78.89	**87.72**
	+ **ESFR**	**76.84**+4.10	**84.36**+0.22	**81.77**+2.88	87.61-0.11
	+ **ESFR-Semi**		**84.97**+0.83		**88.10**+0.38

ESFR-Semi:
Add additional support classification loss during reconstruction training.

†BD-CSPN, Liu, J., Song, L., and Qin, Y. Prototype rectification for few-shot learning. In ECCV, 2020.

Experiment. Comparison to prior works

Table 2. Comparison with state-of-the-art methods of 5-way 1- and 5-shot accuracy (in %) on *mini*-ImageNet, *tiered*-ImageNet and CUB. The best results are reported in **bold**.

		***mini*-ImageNet**		***tiered*-ImageNet**		**CUB**	
Method	**Backbone**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
MAML (Finn et al., 2017)	ResNet-18	49.61	65.72	-	-	68.42	83.47
Chen (Chen et al., 2019)	ResNet-18	51.87	75.68	-	-	67.02	83.58
ProtoNet (Snell et al., 2017)	ResNet-18	54.16	73.68	-	-	72.99	86.64
TPN (Liu et al., 2019)	ResNet-12	59.46	75.65	-	-	-	-
TEAM (Qiao et al., 2019)	ResNet-18	60.07	75.90	-	-	80.16	87.17
SimpleShot (Wang et al., 2019)	ResNet-18	63.10	79.92	69.68	84.56	70.28	86.37
CTM (Li et al., 2019)	ResNet-18	64.12	78.64	68.41	84.28	-	-
FEAT (Ye et al., 2020)	ResNet-18	66.78	82.05	70.80	84.79	-	-
BD-CSPN (Liu et al., 2020)	ResNet-18	70.00	82.36	77.28	86.55	78.89	88.70
LaplacianShot (Ziko et al., 2020)	ResNet-18	72.11	82.31	78.98	86.39	80.96	88.68
BD-CSPN + ESFR (Ours)	ResNet-18	**73.98**	82.32	**80.13**	86.34	**82.68**	88.65
BD-CSPN + ESFR-Semi (Ours)	ResNet-18	-	**82.89**	-	**86.83**	-	**89.10**
LEO (Rusu et al., 2019)	WRN	61.76	77.59	66.33	81.44	-	-
wDAE-GNN (Gidaris & Komodakis, 2019)	WRN	62.96	78.85	68.18	83.09	-	-
FEAT (Ye et al., 2020)	WRN	65.10	81.11	70.41	84.38	-	-
Tran. Baseline (Dhillon et al., 2020)	WRN	65.73	78.40	73.34	85.50	-	-
SimpleShot (Wang et al., 2019)	WRN	65.87	82.09	70.90	85.76	-	-
SIB (Hu et al., 2020)	WRN	70.0	79.2	-	-	-	-
BD-CSPN (Liu et al., 2020)	WRN	72.74	84.14	78.89	87.72	-	-
LaplacianShot (Ziko et al., 2020)	WRN	74.86	84.13	80.18	87.56	-	-
BD-CSPN + ESFR (Ours)	WRN	**76.84**	**84.36**	**81.77**	87.61	-	-
BD-CSPN + ESFR-Semi (Ours)	WRN	-	**84.97**	-	**88.10**	-	-

- State-of-the-art performance on all mini-/tiered-ImageNet and CUB datasets.
- For 1-shot, 1.2%~2.0% improvements in accuracy over the previous best performing.

Summary

- In this work.
  - We propose unsupervised embedding adaptation method: ESFR.
  - Experiments show that our method consistently improves the baseline methods and achieves the new state-of-the-art.
  - We show that deep unsupervised learning can offer complementary and comparable improvement to previous few-shot classification methods.
    - We hope that our work will become a starting point for future unsupervised learning studies on few-shot classification.

Thanks for listening!

Speaker: Dong Hoon Lee

donghoonlee [at] kaist.ac.kr

https://github.com/movinghoon/ESFR