




Supplementary Material for Deep Image Clustering with Model-Agnostic Meta-Learning

Kim Bjerge¹^a, Paul Bodesheim²^b and Henrik Karstoft¹^c

¹*Department of Electrical and Computer Engineering, Aarhus University, Finlandsgade 22, 8200 Aarhus N, Denmark*

²*Computer Vision Group, Friedrich Schiller University, Ernst-Abbe-Platz 2, 07743 Jena, Germany*
{kbe,hka}@ece.au.dk, paul.bodesheim@uni-jena.de

Keywords: Deep clustering, Episodic training, Few-Shot Learning, Multivariate loss, Semi-supervised learning

Abstract: Deep clustering has proven successful in analyzing complex, high-dimensional real-world data. Typically, features are extracted from a deep neural network and then clustered. However, training the network to extract features that can be clustered efficiently in a semantically meaningful way is particularly challenging when data is sparse. In this paper, we present a semi-supervised method to fine-tune a deep learning network using Model-Agnostic Meta-Learning, commonly employed in Few-Shot Learning. We apply episodic training with a novel multivariate scatter loss, designed to enhance inter-class feature separation while minimizing intra-class variance, thereby improving overall clustering performance. Our approach works with state-of-the-art deep learning models, spanning convolutional neural networks and vision transformers, as well as different clustering algorithms like K-means and Spectral clustering. The effectiveness of our method is tested on several commonly used Few-Shot Learning datasets, where episodic fine-tuning with our multivariate scatter loss and a ConvNeXt backbone outperforms other models, achieving adjusted rand index scores of 89.7% on the EU moths dataset and 86.9% on the Caltech birds dataset, respectively. Hence, our proposed method can be applied across various practical domains, such as clustering images of animal species in biology. This document contains supplementary material for the paper "Deep Image Clustering with Model-Agnostic Meta-Learning".

SUPPLEMENTARY MATERIAL

This document presents tables detailing the results of training deep learning models, including ResNet50v2, EfficientNetB3, ConvNeXt-B, and ViT-B/16, on various datasets such as EU Moths, Caltech Birds (CUB), tiered-ImageNet, and mini-ImageNet, as outlined in the accompanying paper.


Tables with Results


Tables 1 to 3 presents the CA, NMI, AMI, and ARI metrics for K-means and Spectral clustering of features extracted from fine-tuned models using both classic and episodic training of the EU moths, CUB and tiered-ImageNet datasets. Metrics in all tables are computed across 5 random runs of clustering fea-

ture vectors from fine-tuned DL models, with average (AVG) and standard deviations (SD). The best metric results for each clustering method in each table is highlighted with bold.

Table 4 presents the clustering metrics for features extracted from fine-tuned models using both classic and episodic training of the mini-ImageNet. The best results were obtained using K-means clustering on features from an episodically trained model with multivariate scatter loss ($\alpha = 0.5$). K-means achieved an average ARI of 0.977 and AMI of 0.984 across 5 runs, nearly identical to Scatter clustering, which produced an ARI of 0.971 and AMI of 0.984.

It is observed that an $\alpha > 0.0$ consistently contributes to the best model performance during episodic training. In two cases (ResNet50v2 and ViT-B/16), the highest scores were achieved with Spectral clustering at $\alpha = 1.0$. Episodic training consistently outperformed classic training, with ARI scores increasing by 1-4%. However, in one instance — EfficientNetB3 with Spectral clustering — classic training slightly outperformed episodic training.

^a <https://orcid.org/0000-0001-6742-9504>

^b <https://orcid.org/0000-0002-3564-6528>


^c <https://orcid.org/0000-0003-3739-8983>

Table 1: Shows the metrics (CA, NMI, AMI, ARI) for K-means and spectral clustering of features from fine-tuned models with classic and episodic training of the EU moths dataset, tested with clustering of 50 classes. A best α value above zero indicates that the multivariate scatter loss improved episodic training in finding the best model.

Model	Cluster method	Training	Best α	CA AVG (SD)	NMI AVG (SD)	AMI AVG (SD)	ARI AVG (SD)
ResNet50v2	K-means	Classic	-	0.575 (0.029)	0.741 (0.013)	0.574 (0.020)	0.385 (0.026)
EfficientNetB3	K-means	Classic	-	0.494 (0.025)	0.674 (0.011)	0.470 (0.016)	0.280 (0.017)
ConvNeXt-B	K-means	Classic	-	0.773 (0.022)	0.872 (0.009)	0.785 (0.014)	0.653 (0.026)
ViT-B/16	K-means	Classic	-	0.769 (0.020)	0.868 (0.010)	0.778 (0.017)	0.649 (0.024)
ResNet50v2	K-means	Episodic	0.1	0.731 (0.024)	0.837 (0.014)	0.727 (0.024)	0.584 (0.030)
EfficientNetB3	K-means	Episodic	0.0	0.712 (0.019)	0.830 (0.008)	0.715 (0.014)	0.562 (0.014)
ConvNeXt-B	K-means	Episodic	0.1	0.837 (0.029)	0.912 (0.005)	0.851 (0.008)	0.748 (0.009)
ViT-B/16	K-means	Episodic	0.1	0.799 (0.030)	0.890 (0.009)	0.815 (0.014)	0.686 (0.028)
ResNet50v2	Spectral	Classic	-	0.690 (0.003)	0.797 (0.003)	0.657 (0.005)	0.484 (0.018)
EfficientNetB3	Spectral	Classic	-	0.649 (0.008)	0.751 (0.005)	0.588 (0.008)	0.330 (0.021)
ConvNeXt-B	Spectral	Classic	-	0.874 (0.009)	0.922 (0.003)	0.868 (0.006)	0.787 (0.009)
ViT-B/16	Spectral	Classic	-	0.866 (0.005)	0.914 (0.001)	0.853 (0.002)	0.774 (0.006)
ResNet50v2	Spectral	Episodic	0.5	0.846 (0.008)	0.893 (0.004)	0.818 (0.006)	0.727 (0.012)
EfficientNetB3	Spectral	Episodic	0.1	0.840 (0.008)	0.889 (0.004)	0.811 (0.007)	0.711 (0.011)
ConvNeXt-B	Spectral	Episodic	0.1	0.940 (0.011)	0.962 (0.004)	0.935 (0.007)	0.897 (0.012)
ViT-B/16	Spectral	Episodic	0.2	0.921 (0.005)	0.950 (0.002)	0.914 (0.003)	0.859 (0.005)

Table 2: Shows the metrics (CA, NMI, AMI, ARI) for K-means and spectral clustering of features from fine-tuned models with classic and episodic training of the CUB dataset, tested with clustering of 40 classes. A best α value above zero indicates that the multivariate scatter loss improved episodic training in finding the best model.

Model	Cluster method	Training	Best α	CA AVG (SD)	NMI AVG (SD)	AMI AVG (SD)	ARI AVG (SD)
ResNet50v2	K-means	Classic	-	0.705 (0.025)	0.771 (0.015)	0.729 (0.018)	0.558 (0.025)
EfficientNetB3	K-means	Classic	-	0.612 (0.009)	0.715 (0.007)	0.664 (0.009)	0.470 (0.011)
ConvNeXt-B	K-means	Classic	-	0.844 (0.019)	0.885 (0.010)	0.864 (0.011)	0.759 (0.026)
ViT-B/16	K-means	Classic	-	0.813 (0.034)	0.857 (0.018)	0.831 (0.022)	0.713 (0.035)
ResNet50v2	K-means	Episodic	0.1	0.747 (0.016)	0.813 (0.007)	0.779 (0.008)	0.627 (0.014)
EfficientNetB3	K-means	Episodic	0.0	0.676 (0.013)	0.765 (0.005)	0.722 (0.006)	0.547 (0.013)
ConvNeXt-B	K-means	Episodic	0.1	0.834 (0.025)	0.885 (0.005)	0.864 (0.006)	0.745 (0.038)
ViT-B/16	K-means	Episodic	0.3	0.839 (0.009)	0.883 (0.004)	0.862 (0.005)	0.755 (0.008)
ResNet50v2	Spectral	Classic	-	0.801 (0.001)	0.823 (0.001)	0.791 (0.001)	0.657 (0.003)
EfficientNetB3	Spectral	Classic	-	0.712 (0.004)	0.763 (0.005)	0.720 (0.006)	0.543 (0.010)
ConvNeXt-B	Spectral	Classic	-	0.936 (0.000)	0.932 (0.000)	0.919 (0.000)	0.881 (0.000)
ViT-B/16	Spectral	Classic	-	0.894 (0.001)	0.897 (0.001)	0.878 (0.001)	0.809 (0.002)
ResNet50v2	Spectral	Episodic	0.0	0.829 (0.001)	0.847 (0.001)	0.818 (0.001)	0.706 (0.002)
EfficientNetB3	Spectral	Episodic	0.2	0.768 (0.001)	0.801 (0.003)	0.765 (0.004)	0.606 (0.003)
ConvNeXt-B	Spectral	Episodic	0.0	0.928 (0.002)	0.930 (0.002)	0.917 (0.002)	0.869 (0.004)
ViT-B/16	Spectral	Episodic	0.4	0.921 (0.001)	0.926 (0.000)	0.913 (0.000)	0.863 (0.002)

Table 3: Shows the metrics (CA, NMI, AMI, ARI) for K-means clustering of features from fine-tuned models with classic and episodic training of the tiered-ImageNet dataset, tested with clustering of 160 classes. Spectral clustering is only measured for ConvNeXt-B with limited number of α values. A best α value above zero indicates that the multivariate scatter loss improved episodic training in finding the best model.

Model	Cluster method	Training	Best α	CA AVG (SD)	NMI AVG (SD)	AMI AVG (SD)	ARI AVG (SD)
ResNet50v2	K-means	Classic	-	0.740 (0.000)	0.847 (0.000)	0.845 (0.000)	0.645 (0.000)
EfficientNetB3	K-means	Classic	-	0.721 (0.000)	0.827 (0.000)	0.825 (0.000)	0.616 (0.000)
ConvNeXt-B	K-means	Classic	-	0.895 (0.005)	0.942 (0.001)	0.941 (0.001)	0.853 (0.005)
ViT-B/16	K-means	Classic	-	0.832 (0.000)	0.895 (0.000)	0.894 (0.000)	0.728 (0.000)
ResNet50v2	K-means	Episodic	1.0	0.746 (0.006)	0.846 (0.001)	0.845 (0.001)	0.640 (0.005)
EfficientNetB3	K-means	Episodic	1.0	0.603 (0.000)	0.761 (0.000)	0.759 (0.000)	0.476 (0.000)
ConvNeXt-B	K-means	Episodic	0.5	0.900 (0.010)	0.942 (0.001)	0.941 (0.001)	0.856 (0.009)
ViT-B/16	K-means	Episodic	0.0	0.867 (0.009)	0.923 (0.002)	0.922 (0.002)	0.811 (0.007)
ConvNeXt-B	Spectral	Classic	-	0.903 (0.003)	0.941 (0.001)	0.941 (0.001)	0.832 (0.006)
ConvNeXt-B	Spectral	Episodic	-	0.893 (0.002)	0.937 (0.000)	0.936 (0.000)	0.813 (0.001)

Table 4: Shows the metrics (CA, NMI, AMI, ARI) for K-means and spectral clustering of features from fine-tuned models with classic and episodic training of the mini-ImageNet dataset, tested with clustering of 20 classes. A best α value above zero indicates that the multivariate scatter loss improved episodic training in finding the best model.

Model	Cluster method	Training	Best α	CA AVG (SD)	NMI AVG (SD)	AMI AVG (SD)	ARI AVG (SD)
ResNet50v2	K-means	Classic	-	0.879 (0.037)	0.926 (0.013)	0.926 (0.013)	0.842 (0.044)
EfficientNetB3	K-means	Classic	-	0.858 (0.053)	0.902 (0.027)	0.902 (0.027)	0.779 (0.097)
ConvNeXt-B	K-means	Classic	-	0.975 (0.024)	0.984 (0.007)	0.984 (0.007)	0.965 (0.028)
ViT-B/16	K-means	Classic	-	0.904 (0.029)	0.934 (0.016)	0.934 (0.016)	0.870 (0.033)
ResNet50v2	K-means	Episodic	0.2	0.917 (0.023)	0.940 (0.011)	0.940 (0.011)	0.888 (0.021)
EfficientNetB3	K-means	Episodic	0.4	0.883 (0.026)	0.905 (0.011)	0.904 (0.011)	0.815 (0.041)
ConvNeXt-B	K-means	Episodic	0.5	0.984 (0.021)	0.986 (0.007)	0.986 (0.007)	0.977 (0.023)
ViT-B/16	K-means	Episodic	0.3	0.940 (0.032)	0.963 (0.012)	0.963 (0.012)	0.918 (0.039)
ResNet50v2	Spectral	Classic	-	0.937 (0.000)	0.961 (0.000)	0.961 (0.000)	0.920 (0.000)
EfficientNetB3	Spectral	Classic	-	0.936 (0.000)	0.953 (0.000)	0.953 (0.000)	0.909 (0.000)
ConvNeXt-B	Spectral	Classic	-	0.978 (0.018)	0.982 (0.007)	0.982 (0.007)	0.966 (0.023)
ViT-B/16	Spectral	Classic	-	0.940 (0.000)	0.960 (0.000)	0.960 (0.000)	0.915 (0.000)
ResNet50v2	Spectral	Episodic	1.0	0.938 (0.000)	0.962 (0.000)	0.962 (0.000)	0.923 (0.000)
EfficientNetB3	Spectral	Episodic	0.5	0.927 (0.000)	0.937 (0.000)	0.937 (0.000)	0.892 (0.000)
ConvNeXt-B	Spectral	Episodic	0.2	0.982 (0.019)	0.984 (0.007)	0.984 (0.007)	0.971 (0.024)
ViT-B/16	Spectral	Episodic	1.0	0.945 (0.000)	0.970 (0.000)	0.970 (0.000)	0.928 (0.000)