Domain-Agnostic Clustering with Self-Distillation

M. Adnan, Y. Ioannou, K. Tsai, G. Taylor

UNIVERSITY of GUELPH | VECTOR INSTITUTE | UNIVERSITY OF CALGARY

Introduction

- Current state-of-the-art Self-Supervised Learning (SSL) and clustering algorithms use data augmentation either for learning contrastive representations or as a regularizer.
- For example, color jittering commonly used in SSL cannot be used on black and white x-ray images, and random cropping is not relevant to histopathology images.
- One way to achieve domain agnosticism is to remove the augmentation, incurring a decrease in accuracy.
- However, SSL methods rely heavily on data augmentation, and their performance decreases significantly after removing data augmentation.
- One explanation of the decrease in the accuracy after removing data augmentations could be attributed to the generalization property of data augmentations.
- We introduce a new self-distillation based algorithm for domain-agnostic clustering. Our method builds upon the existing deep clustering frameworks and requires no separate student model.

Deep Cluster

- Caron et al. proposed Deep Clustering [3] (DeepCluster) to jointly learn the neural network and cluster the resulting features by iteratively applying k-means.
- DeepCluster can learn generalizable features in an unsupervised manner using clustering as a pretext task
- DeepCluster uses k-means to obtain pseudo-labels for the output of CNN and then uses the psuedo-labels for backpropagation.
- This process is repeated iteratively until the model converges. After training, the classifier is discarded, and the CNN can be used for downstream learning tasks.

Knowledge Distillation

- Knowledge Distillation (KD) is a model compression method in which a smaller 'student' model is trained to mimic the behavior of a large 'teacher' model [4].
- The smaller model is trained by minimizing the loss on the output class probabilities (soft labels) of the large model.
- It has been found that the smaller model achieves similar, or often better performance than the original model
- Soft labels provide much more information about the semantic information present in the image.
- For example, given a dog image from CIFAR-10, the class probability of the image being a cat will be much higher than the class probability of the image being a car.
- Thus, the softmax values give additional hints to the network that images of dogs and cats contain similar semantic information.
- Knowledge distillation also improves the loss landscape and helps find flat minima, which in turn improves generalization. Distillation has been shown to amplify regularization in the Hilbert space, and thus, it improves generalization [5].

Be Your Own Teacher

- Zhang et al. modified the ResNet architecture to have an additional three bottleneck branches [6]. Secondly, an auxiliary classifier is also added on top of bottleneck branches
- During the training phase, all three bottleneck classifiers along with the original classifier are utilized.
- Bottleneck classifiers are trained as student models via distillation from the deepest classifier, which acts as the teacher model.
- Three different losses are introduced:
  i. **Cross-Entropy**: All bottleneck classifiers (student models) have cross entropy loss from the pseudo-labels obtained from k-means clustering
  ii. **KL Divergence** is used to train the bottleneck classifiers using soft labels from the deepest classifier.
  iii. **L2 loss** between the bottleneck feature map and the deepest layer features is added to provide implicit knowledge or hints to the bottleneck classifiers

Proposed Method

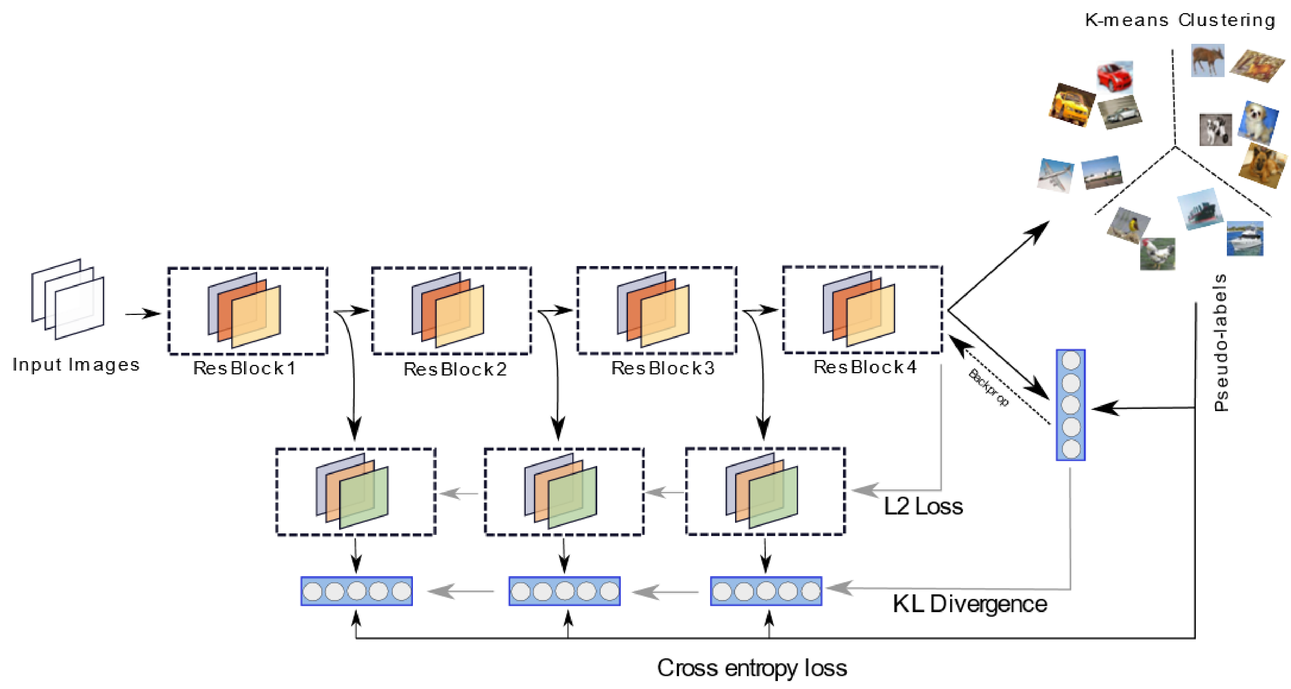

Fig 1. Schematic overview of our method. Bottleneck branches between ResBlocks act as student models and the deepest layer acts as a teacher model.

- The underlying principle behind our proposed method is that semantically similar images of different classes are often embedded close in Euclidean space and are consequently misclassified by k-means.
- However, soft labels may provide more information about semantically similar classes, and thus by using self-distillation, the model can be provided with further information to distinguish those inputs.
- Our proposed method uses the DeepCluster framework while also introducing a self-distillation loss. We use ResNet as the CNN for extracting features and introduce bottleneck classifiers.

Results

- We evaluate the performance of our approach on the CIFAR-10 dataset.
- In contrast to most existing unsupervised learning algorithms, we do not use any domain knowledge for data augmentations.
- The trained model can be used to extract general-purpose features for downstream machine learning tasks. Results are shown in Table 1.

Methods	**Accuracy**
ID [1]	18.7%
IDFD [1]	23.6%
ConCURL [2]	29.88%
DeepCluster-v2 [3]	33.27 ± 0.06 %
DeepCluster+KD (ours)	**38.00 ± 0.34%**

Table 1: Domain Agnostic Clustering on CIFAR-10. DeepCluster was trained with no data augmentation.

References

[1] Tao et al, Clustering-friendly Representation Learning via Instance Discrimination and Feature Decorrelation.

[2] Deshmukh et al., Representation Learning for Clustering via Building Consensus.

[3] Caron et al., Unsupervised Learning of Visual Features by Contrasting Cluster Assignments.

[4] Hinton et al., Distilling the knowledge in a neural network.

[5] Mobahi et al., Self-distillation amplifies regularization in hilbert space.

[6] Zhang et al, Be your own teacher: Improve the performance of convolutional neural networks via self distillation