AN ANALYSIS OF SELF-SUPERVISED CLASSIFICATION NETWORK

TECHNICAL REPORT

Andrea Camilloni
acam@ug.kth.se

Chetan Reddy Narayanaswamy
crna@kth.se

Raghav Singhal
raghavsi@kth.se

ABSTRACT

In this work, we re-implement "Self-Classifier": the self-supervised end-to-end classification learning algorithm as proposed in [1]. Self-Classifier optimizes for same-class prediction of two augmentations of the same sample, learning representations and labels simultaneously in one stage, from beginning to finish. The method obtains results equivalent to the state-of-the-art in unsupervised representation learning and establishes a new state-of-the-art for ImageNet classification. We implement the method on the CIFAR-10 dataset and evaluate the performance. Further, we experiment with the model by changing the backbone model. Most notably, we improve performance on CIFAR-10 by changing the architecture of the MLP head. We evaluate the model's image classification performance after adding a linear classifier. We attempt to extend the algorithm for interesting applications, specifically the image segmentation task.

1 Introduction

In recent years, interest in self-supervised visual representation learning has grown. The fundamental goal is to specify and complete a pretext job that enables learning semantically meaningful representations without needing human-annotated labels. Later on, the learned representations are applied to subsequent tasks, for example, by fine-tuning on a subset of the dataset. In contrast, the paper provides a classification-based pretext job in this study whose objective is exactly linked with the final goal. Knowing the total number of classes C, they develop a self-classifying algorithm (Self-Classifier) that allows us to classify two alternative augmentations of the same picture identically.

The major contributions of the paper were:

- Developing a straightforward but efficient single-stage end-to-end self-supervised classification and representation learning method.
- Despite being straightforward, the method delivers performance on par with state-of-the-art unsupervised representation learning and sets a new benchmark for unsupervised classification on ImageNet with 41.1% top-1 accuracy.
- The work indicates a notable (up to 3.4% AMI) improvement in the new metric over the prior state of the art on the ImageNet.

We implemented "Self-Classifier" on the CIFAR-10 dataset to learn the features. We tested different backbone models, namely ResNet-18 and AlexNet, and evaluated the performance. We tested different architectures of the MLP head and observed improvements as well. Notably, we improved performance on CIFAR-10 by changing the architecture of the MLP head. We tested its image classification capabilities by adding a linear classification layer to the learned representations. We were able to obtain an impressive 87.89% top-1 accuracy on CIFAR-10. We also tried to extend the learned features for interesting tasks, specifically image segmentation but were unable to do so due to reasons outlined later in the report

Code Repository: All the code and results mentioned in this report can be found in this link.

2 Related Work

Self-supervised classification is a type of machine learning technique that uses unlabeled data to train a model. It uses techniques such as clustering, self-organizing maps, and deep learning algorithms to discover patterns in the data and then use those patterns to make predictions. Unlike supervised learning, self-supervised classification does not require a labeled dataset or human intervention to train the model. Instead, the model learns from the data itself. This type of classification can be used for various tasks, including image processing, natural language processing, and anomaly detection.

In recent years, several approaches have been proposed to improve the performance of self-supervised methods. One of the most widely used approaches is contrastive learning. In this approach, a representation of an image is learned by comparing it to other images. For example, the authors used a contrastive loss to learn representations for images without labels in the SimCLR method proposed by Chen et al. (2020) [2]. By training on large datasets of unlabeled images, the model could learn rich representations that could be used for image classification.

Another approach to self-supervised learning is to use pretext tasks. In this approach, a model is trained to solve a task that can be used to learn features from the data. For example, in the work by Doersch et al. (2015) [3], the authors used a jigsaw puzzle task to train a model for image classification. The model was then used to classify images without labels.

In addition to these approaches, some works use generative models for self-supervised learning. For example, in the work by Oord et al. (2016) [4], the authors used a generative adversarial network (GAN) to learn representations of images. The representations were then used in a downstream image classification task.

Overall, self-supervised learning can potentially reduce the cost of labeling by learning meaningful representations from large datasets. Numerous approaches have been proposed to improve the performance of self-supervised classification, including contrastive learning, pretext tasks, and generative models.

3 Methods

Without going into too many details, we describe the fundamental logic behind the paper. The paper introduces an unsupervised classifier (Self-Classifier) such that two alternative augmentations of the same picture are categorized similarly while just knowing the number of classes C. Generally, a task like this is prone to degenerate solutions, where every sample is put in the same class. The paper asserts a uniform prior on the common cross-entropy loss function to prevent them, making an answer that evenly divides the data the best option. The architecture is shown in 1. We have avoided going into too many details about the paper, but interested readers can refer [1] for more details.

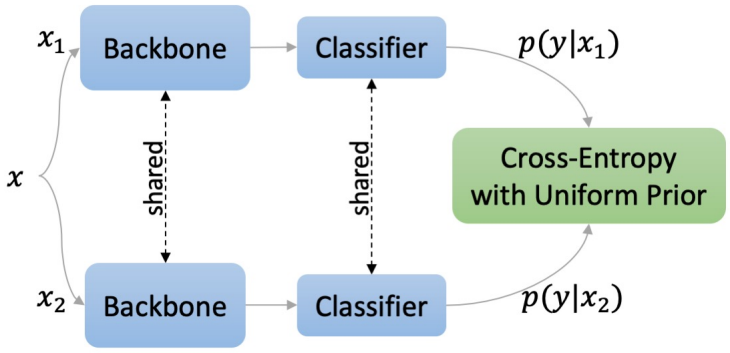


Figure 1: Self-Classifier architecture as depicted in [1]

4 Data

The CIFAR-10 dataset consists of 60,000 color images that are divided into 10 classes. Each class consists of 6,000 images which are of size 32x32 pixels. Examples of CIFAR-10 images can be seen in Fig. 2. The classes are *Airplanes,*

Cars, Birds, Cats, Deer, Dogs, Frogs, Horses, Ships, and Trucks. The images are in RGB format, which means that each pixel is represented by three 8-bit integers ranging from 0 to 255. The dataset is divided into five training batches and one test batch, each with 10,000 images. The training batches contain the data used for learning, while the test batch contains the data used to evaluate the performance of the model created. The data is further divided into 50,000 training images and 10,000 test images[5].

The CIFAR-10 dataset is well-known and widely used for image recognition tasks. It is a popular benchmark for computer vision and machine learning tasks. It is also used for benchmarking different deep-learning models.

Overall, the CIFAR-10 dataset is an excellent resource for unsupervised learning, with many images of different classes, providing a broad range of visual information. It is also an excellent resource for research in computer vision and machine learning, as it is well-labeled and easy to use.

Figure 2: Examples of CIFAR-10 images, containing the following classes: *airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck*.

State-of-the-art unsupervised classification on CIFAR-10: The model RUC (Robust Learning for Unsupervised Clustering) described in [6] uses unsupervised image clustering with robust learning to obtain 90.3% unsupervised classification accuracy on CIFAR-10. SCAN (Semantic Clustering by Adopting Nearest neighbors) [7] uses a combination of representation and end-to-end learning to achieve 88.3% accuracy on the same.

5 Experiments

All experiments were performed on an Nvidia Tesla T4 GPU provided by the Google Cloud Engine.

5.1 Unsupervised Image Classification: ResNet-18

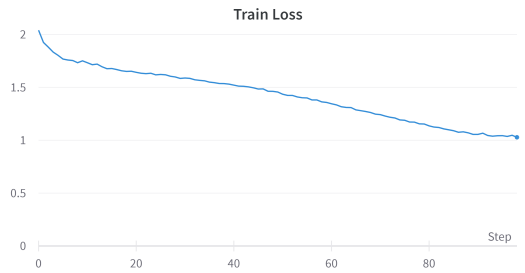
The original paper used a ResNet-50 backbone, followed by an MPL head with two hidden layers (of sizes 4096 and 128) with Batch Normalization, leaky-ReLU activations, and $l2$ normalization after the last layer. This experiment configuration consists of a ResNet-18 randomly initialized model followed by an MLP head composed of one hidden layer with input size 512 and output size 128. We trained using the original MLP head architecture initially. Since we were exploring different architectures, we guessed that a smaller MLP head might give better results on smaller datasets like CIFAR-10, which was true. Thus, we report results on the MLP head with one hidden layer, which improved performance compared to the model described in the original paper when trained on CIFAR-10.

The MLP head utilizes $l2$ normalization after the last layer and is accompanied by a classification head into 10 classes, a simple linear layer without an additive bias term. The row-softmax temperature τ_{row} was set to 0.1 and the column-softmax temperature τ_{col} to 0.05 for the loss.

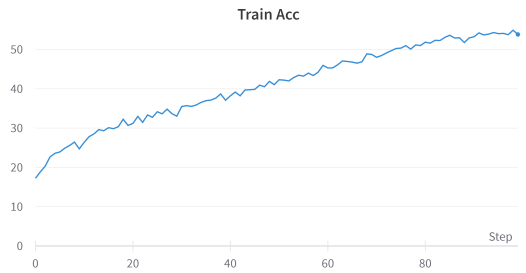
The unsupervised classification is done strictly using the 10-classes classification head. The model was trained using the CIFAR-10 training set with a batch size of 64, using the SGD optimization with Cosine scheduler and LARS.

The optimizer configuration for this experiment was chosen according to [1], as it provides optimal performance for the model.

Figure 3 shows the history of the model's training. The figure shows that the model's accuracy increases during the whole training, and the loss decreases as well.



(a) Unsupervised image classification training loss

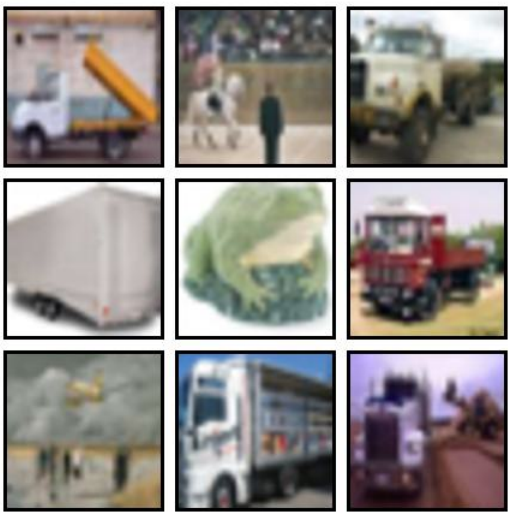


(b) Unsupervised image classification training accuracy

Figure 3: Unsupervised image classification training loss and accuracy curves on CIFAR-10 with ResNet-18 backbone

Figure 4 shows the results of the inference made with the unsupervised classification model. The results show that the model was able to group accurately similar images into the same groups. Further results of the model can be found in section 5.3 of the paper, which provides a detailed analysis of the model's performance.

(a) An example class (horse)



(b) An example class (truck)

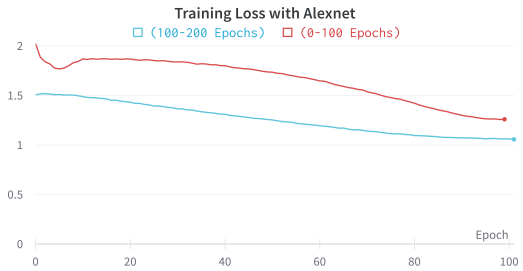
Figure 4: Example classes predicted on CIFAR-10 validation set with ResNet-18 backbone

5.2 Unsupervised Image Classification: AlexNet

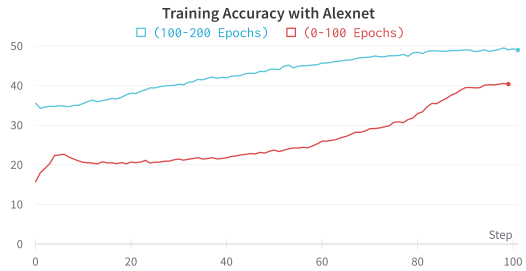
This experiment uses AlexNet which is a different backbone architecture with randomly initialised weights. The MLP head was implemented according to [1] followed by a classification head of 10 classes. The other training configurations are exactly the same as experiment 5.1 which uses ResNet-18 architecture.

The motivation to use AlexNet is the fact it can be trained quickly due to a simpler architecture compared to ResNet-18. We also wanted to check the performance of a simpler backbone.

Figure 5 shows the history of the model's training using AlexNet backbone. It was observed that the loss was on a downward trend even after 100 epochs and the training was resumed until 200 epochs. The graphs show that the expected trends of the loss decreasing and the accuracy increasing as the training is carried out.



(a) Training loss



(b) Training accuracy

Figure 5: Unsupervised image classification training loss and accuracy on CIFAR-10 with AlexNet backbone

(a) An example class (truck) predicted after 100 epochs

(b) An example class (truck) predicted after 200 epochs

Figure 6: Example classes predicted on CIFAR-10 validation set with AlexNet backbone

5.3 Metrics for Unsupervised Classification

We evaluate the metrics below (as defined in [1]) for the different backbones and the numbers of epochs. All the metrics were computed on unseen test data. As can be seen, the ResNet-18 backbone with one hidden layer performs the best.

Backbone	Epochs	NMI	AMI	ARI	ACC
ResNet-18 (1 HL)	**100**	**61.1**	**61.0**	**50.5**	**68.7**
ResNet-18 (2 HL)	100	60.3	60.2	49.9	67.9
AlexNet (2 HL)	100	41.9	41.8	28.4	46.5
AlexNet (2 HL)	200	48.8	48.7	34.7	52.0

Table 1: **CIFAR-10 unsupervised classification using different backbones.** NMI: Normalized Mutual Information, AMI: Adjusted Normalized Mutual Information, ARI: Adjusted Rand-Index, ACC: Clustering accuracy

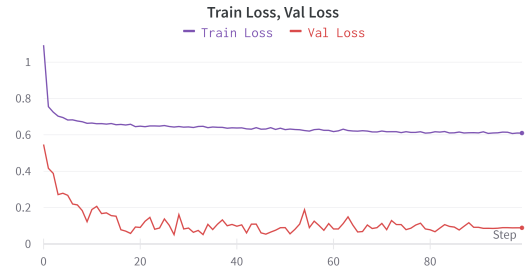
5.4 Image Classification with Linear Models

After unsupervised training, we used the trained model to further train a linear classifier. We used the Resnet-18 backbone since it performed better. We freeze the features of the self-supervised model (from experiment 5.1) and train on top of it a supervised linear classifier (a single fully-connected layer).

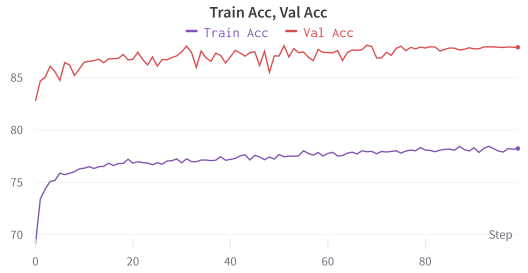
The model was trained for 100 epochs using the Stochastic Gradient Descent (SGD) optimizer and LARS. This optimizer was chosen according to [1]. The Cosine Scheduler adjusted the learning rate throughout the training process. The training batch size of 64, and a validation batch size of 256. This allowed the model to quickly learn features from the training data while still being able to validate the performance of the model.

The training history in Figure 7 exhibits an unusual behavior of the validation and training losses due to the complexity of the images in the validation set being lower than those in the training set. This discrepancy between training and validation is due to a combination of augmentation and batch normalization in the training pipeline.

The top-1 accuracy on the validation set was $\boxed{87.89\%}$.



(a) Training and validation loss curves



(b) Training and validation accuracy curves

Figure 7: Training & validation loss & accuracy curves for image classification with a linear model with ResNet-18 backbone

5.5 Image Segmentation

We attempted to fine-tune our pre-trained model end-to-end in the target datasets using the public codebase from MoCo [8] and further evaluate the quality of our unsupervised features by transferring them to other tasks - object detection and image segmentation. However, the overall results were very poor due to low computational resources and the model being trained in a low-resolution dataset. The major issue is the features learned from CIFAR-10 are not good enough to extend to such applications. Thus, we did not report the results of this experiment.

6 Challenges

The major challenge we faced during the implementation of the paper was the lack of computational resources. The original paper distributed the training on ImageNet across 64 NVIDIA V100 GPUs. We did not have such resources. Thus, we trained the model on a smaller dataset, CIFAR-10. Due to the same lack of resources, we could not perform all the experiments till the model converged. Thus, we could have obtained better performance in them. This was especially detrimental to the image segmentation task.

7 Conclusion

We achieved good performance when we implemented the paper on CIFAR-10 (metrics mentioned above). Upon addition of a linear classifier to the learned representations, we obtained an impressive 87.89% top-1 accuracy. We found that using 1 hidden layer in the MLP head for CIFAR-10, instead of 2 hidden layers as proposed in the paper for ImageNet, gave better results. Thus, we believe we contributed positively to the architecture decisions when training the algorithm on smaller datasets like CIFAR-10.

The lack of computational resources is the biggest bottleneck while reimplementing the paper. One has to be creative to tackle this problem. Future applications of this work would be to use the learned representations, trained on bigger datasets like ImageNet, for interesting applications such as image segmentation, object detection, etc. But we are most excited about using the algorithm's capabilities to leverage the vast amount of unlabeled medical data available for the segmentation of cancer cells and so on. We believe this could potentially be a very high-impact application of this work, beneficial to large parts of society.

8 Ethical Consideration

Goal 9 of the UN Sustainable Development Goals (SDG) states, "**Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation**." We believe that this paper contributes directly to this goal through innovation. It is a novel method to reduce human costs associated with labeling large datasets and leverage scenarios wherein obtaining labeled data is highly problematic. One such application could be the field of medical diagnosis, wherein large amounts of unlabeled data are available. In our opinion, this is a very high-impact domain in which this work can be highly beneficial.

9 Self Assessment

Code Repository: All the code and results mentioned in this report can be found in this link.

Self Assessment: We believe that we deserve an **Excellent (A)** grade as per the grading guidelines. The reasons are as follows:

- We implemented the proposed method on a different dataset and achieved coherent results with the original paper.
- We achieved better results on CIFAR-10 using one hidden layer instead of two hidden layers, as was done in the paper. Thus, we show that lowering the number of hidden layers gives us benefits while working on smaller datasets, which is something that was not explored in the original paper. Thus, we contributed positively to the architecture decisions when training the algorithm on smaller datasets like CIFAR-10.
- We experimented with different backbone models, reporting the reasons for the results.
- We attempted to extend the paper to perform image segmentation, which was not very effective due to the reasons outlined in the paper.

References

[1] E. Amrani, L. Karlinsky, and A. Bronstein, "Self-supervised classification network," *arXiv preprint arXiv:2103.10994*, 2021.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, *A simple framework for contrastive learning of visual representations*, 2020. DOI: 10.48550/ARXIV.2002.05709. [Online]. Available: https://arxiv.org/abs/2002.05709.

[3] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1422–1430. DOI: 10.1109/ICCV.2015.167.

[4] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, *Pixel recurrent neural networks*, 2016. DOI: 10.48550/ARXIV.1601.06759. [Online]. Available: https://arxiv.org/abs/1601.06759.

[5] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)," [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html.

[6] S. Park, S. Han, S. Kim, *et al.*, "Improving unsupervised image clustering with robust learning," *CoRR*, vol. abs/2012.11150, 2020. arXiv: 2012.11150. [Online]. Available: https://arxiv.org/abs/2012.11150.

[7] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, "Scan: Learning to classify images without labels," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 268–285, ISBN: 978-3-030-58607-2.

[8] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.