Self-Supervised Learning with SimCLR and RotNet

Anna Dai
Tigran Harutyunyan
Deekshita Saikia

Duke
PRATT SCHOOL of ENGINEERING

Abstract

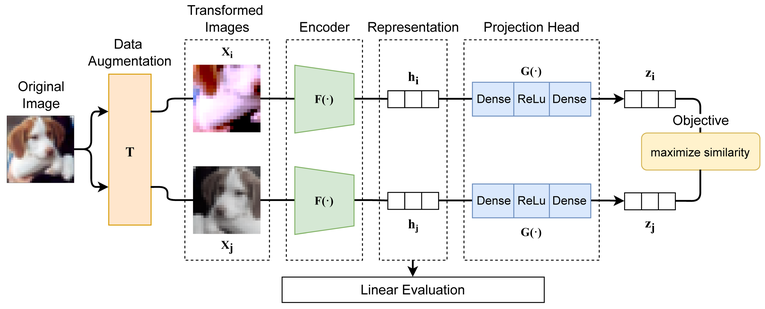
Convolutional neural networks (“ConvNets”) have unparalleled capacity to learn high level semantic image features, but usually require large amounts to labeled data to train. Labeling data comes at a high cost and is not easily scalable. The field of self-supervised learning explores how supervisory signals can be learned from the data itself, often leveraging its underlying structure, thereby reducing the need for labels for predictive tasks like classification. We explore two approaches both leveraging ResNet-20:

1. Contrastive Learning with SimCLR, where networks learn to identify similarities between augmented versions of the same image. We perform linear evaluations on our model. It performs better than RotNet linear evaluations but does not achieve performance comparable to [1].
2. Train ConvNets to learn geometric transformations of input images with RotNet. Even with poor performance in linear evaluation, we find that even with a smaller network, RotNet can achieve comparable performance as [2] in the semi-supervised setting.

Contrastive Learning with SimCLR

Framework

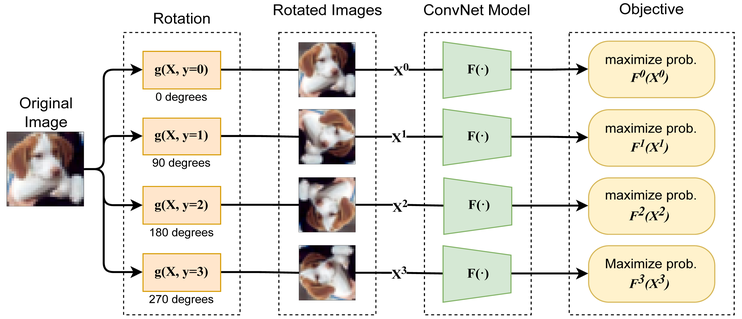
- Stochastic data augmentation modules to transform an input image into two correlated images, as a positive pair
- ConvNet base encoder (ResNet-20) to extract feature representations
- 2-Layer multi-layer perceptron (MLP) projection head to project representations to a latent space
- Contrastive loss function (*NT-Xent*) to maximize similarity between a positive pair



Learning Geometric Transformations with RotNet

Framework

- ConvNets to learn representations of geometrically transformed (rotations by 0°, 90°, 180°, and 270°) input images
- ConvNets must first learn to recognize and localize objects and their semantics in images, like type and orientation.
- It then relates object position with the dominant orientation that each object type is usually depicted within the available input images to predict rotations effectively.



Evaluations

- Pre-training is unsupervised (training encoder w/o labels) on CIFAR-10
- Semi-supervised learning carried out using 1% and 10% labeled data
- Linear evaluation protocol used to evaluate learned representations by training a linear classifier on top of the frozen encoder
- RotNet also evaluated in its semi-supervised setting – retrain last block

Model	1% Labels	10% Labels
Supervised Baseline	50.54%	81.94%
SimCLR Linear Evaluation	64.15%	69.30%
RotNet Semi-supervised	70.71%	83.64%
RotNet Linear Evaluation	47.31%	55.59%

Table : Model Performance with 1% and 10% Labeled Data.

Of the three self-supervised models, RotNet in its semi-supervised setting performed best, beating both SimCLR and Supervised Baseline (ResNet20). Performance variation could be due to limited fine-tuning.

Feature Maps of Conv Layers

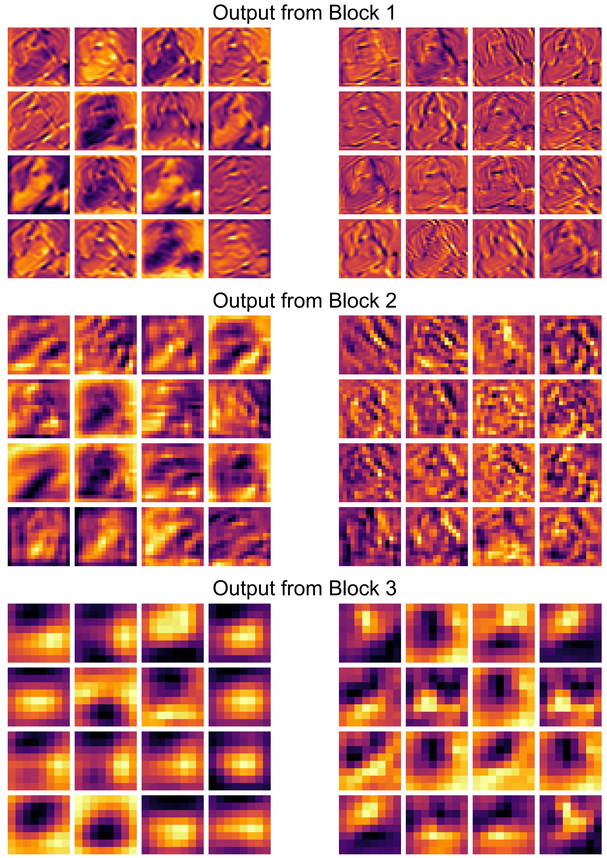


Fig.: Demonstration of features learned by the ResNet-20 encoder in the SimCLR (left) and RotNet (right) frameworks. Both architectures learn different aspects of the input images.

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations.” arXiv, Jun. 30, 2020. Accessed: Dec. 14, 2022. [Online]. Available: http://arxiv.org/abs/2002.05709

[2] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised Representation Learning by Predicting Image Rotations.” arXiv, Mar. 20, 2018. Accessed: Dec. 14, 2022. [Online]. Available: http://arxiv.org/abs/1803.07728

[3] H. Ren, SimCLR, 2020, GitHub repository, https://github.com/leftthomas/SimCLR

[4] S. Manna, simclr_pytorch, 2021, GitHub repository, https://github.com/sadimanna/simclr_pytorch

[5] Rath, S. R. (2020, August 20). Visualizing filters and feature maps in convolutional neural networks using pytorch. DebuggerCafe. Retrieved December 15, 2022, from https://debuggercafe.com/visualizing-filters-and-feature-maps-in-convolutional-neural-networks-using-pytorch/