Self-Supervised and Representation Learning Using a SimCLR and RotNet Model

Samuel Obeng, Hakeem Shitta-Bey, and Tony Nunn

Duke PRATT SCHOOL of ENGINEERING

Introduction and Motivation

Self-supervised learning (SSL) is an emerging innovation in the machine learning space where visual representations can be learned without the need for fully labeled data. Contrastive learning frameworks, like SimCLR, have demonstrated competitive performance with that of supervised learning models when put in scenarios with limited labels available. This capability is especially applicable in the real-world where labeling large datasets can be too expensive or impractical.

- Developed a full SimCLR pipeline from scratch and trained it on CIFAR-10 data
- Evaluated performance of the SimCLR under limited data scenarios and compared it to a “supervised baseline”
- Implemented the RotNet framework and replicated the above procedures for direct comparison
- Conducted a comparative study between semi-supervised RotNet and SimCLRmodels
- Proposed and tested a hybrid SSL model that integrates SimCLR with RotNet techniques

Methodology

SimCLR (Figure A) learns meaningful patterns from unlabeled images by maximizing agreement between augmented views. We used a ResNet-50 backbone, an MLP projection head, and the NT-Xent (Normalized Temperature-scaled Cross Entropy) loss.

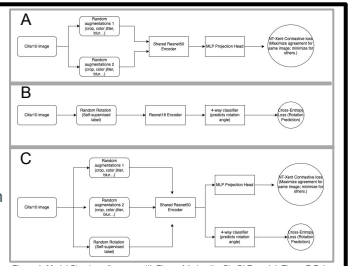

Figure 1: Model Structure diagrams with Figure A being the SimCLR model, Figure B Being the Rotnet Model, and Figure 3 being the hybrid model

RotNet (Figure B) learns representations by classifying input images rotated by (0°, 90°, 180°, or 270°). A ResNet-18 encoder extracts features, followed by a linear head trained with cross-entropy loss over the four rotation classes.

The SimCLR+RotNet model uses a shared encoder to extract features from input images, which are passed to separate SimCLR and RotNet heads. Each computes its own task-specific loss, and a weighted sum of both is used to update the model. To limit RotNet’s influence after early convergence, its head is frozen later in training.

Experimental Evaluations

Applying the Linear Classifier to the Models for a Comparative Study

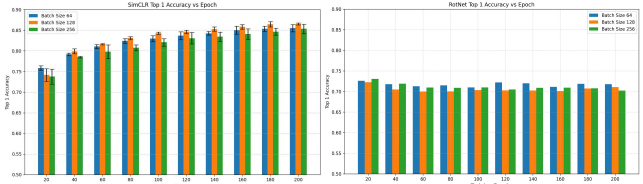

Figure 2a and 2b: Test Accuracy Across 200 Epochs for SimCLR (Figure 2a, Left) and RotNet (Figure 2b, Right) for Batch Sizes 64, 128, and 256

The above figures show accuracy from a Broyden–Fletcher–Goldfarb–Shanno (BFGS) linear evaluation as a function of epoch number and batch size. For SimCLR (Figure 2a) each bar represents an average between three different training pipelines with learning rates equal to 0.0125, 0.025, and 0.05, and the error bars represent the standard deviation. Consistent with current academic evaluations, longer pre-training and larger batch sizes leads to stronger feature representations, which results in a better model performance. RotNet (Figure 2b) does not reflect this pattern, and plateaus or even degrades at higher epochs. Its framework is more suited for learning general representations, so later training epochs could actually result in overfitting. The difference between SimCLR’s feature-rich semantic learning and RotNet’s basic instance-level classification explains their large discrepancy in performance accuracy.

Comparing Performance After Semi-Supervised Learning

We compared SimCLR, RotNet, and supervised baselines (ResNet50 and ResNet18) using only 1%, 10%, and 50% of labeled CIFAR-10 data. As shown in Figure 3, SimCLR consistently outperformed all models at each data fraction, achieving 78.6% accuracy with just 1% of labeled data. At this low-data setting, RotNet (51.6%) also outperformed both supervised models, with ResNet18 and ResNet50 respectively. However, ResNet18 ultimately surpasses all models at 50% data (89.8%), while RotNet plateaus at 69.0%. These results highlight the practical value of self-supervised learning in scenarios where labeled data is scarce, while supervised data improves as labeled data grows.

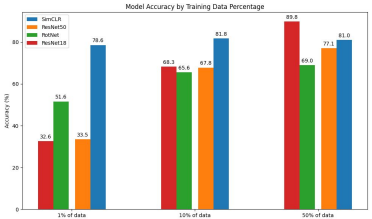

Figure 3: Comparing the Performance of SimCLR, RotNet, and Hybrid Models After Semi-Supervised Learning

Discussion

Exploring a Hybrid Model by Combining SimCLR and RotNet Techniques

We combined contrastive learning with rotation prediction through a shared encoder and loss function. Hypothetically, this leverages both instance-level discriminative features and global semantic understanding. The performance per epoch (Figure 4) shows a decrease in performance compared to using either approach in isolation. This is due to the models’ conflicting objectives: contrastive learning promotes invariance to augmentations, whereas RotNet requires a level of sensitivity to rotations. These competing signals can create representational tension, which can deprecate overall feature quality during learning.

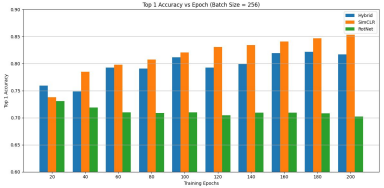

Figure 4: Performance Comparison of SimCLR, RotNet, and SimCLR+RotNet

Conclusion

Using linear evaluation on CIFAR-10, we compared SimCLR, RotNet, and a supervised ResNet baseline. When evaluated on limited data, SimCLR outperformed all other models: learning richer features, while RotNet converged early and focused on low-level patterns. SimCLR surpassed the supervised baseline at lower training percentage data, due to its more transferable representations.

The SimCLR+RotNet model nearly matched SimCLR’s performance but required stronger augmentations, and RotNet’s limited impact suggests the hybrid approach offers little benefit.

Overall, self-supervised learning proves competitive when compared to fully supervised models and excels when labeled data is limited.

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, arXiv preprint arXiv:2002.05709, 2020. [Online]. Available: https://arxiv.org/pdf/2002.05709.pdf

[2] S. Gidaris, P. Singh, and N. Komodakis, Unsupervised representation learning by predicting image rotations, arXiv preprint arXiv:1803.07728, 2018. [Online]. Available: https://arxiv.org/abs/1803.07728