Improving Self-Supervised Representation Learning by Rotation Feature Decoupling & NT-Xent Loss

Zeyu Feng et.al. (Main Author)
UBTECH Sydney AI Centre,
School of Computer Science, FEIT,
University of Sydney, Darlington, NSW 2008, Australia
zfen2406@uni.sydney.edu.au

Tushar Sangam (Improviser)
Masters Student CS dept,
University of Central Florida, Orlando, USA
tusharsangam@knights.ucf.edu

Abstract

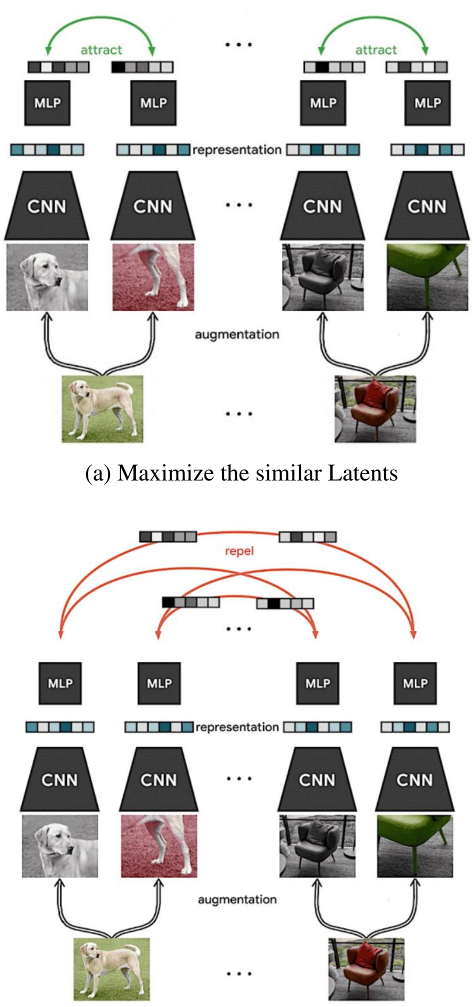
I tried to approximately implement Self-Supervised Representation Learning by Rotation Feature Decoupling by Zeyu Feng et.al. [FXT19] & tried to improve the baseline by borrowing the inspiration from the state of the art in contrastive methods (SimCLR v1 [Che+20a] & v2 [Che+20b]).
The base paper [FXT19] proposes to improve the self-supervised learning for AlexNet by learning rotation related & unrelated features.
I improved the results through two main novelties

- **Different Fine-Tuning method** (using only 20% of labeled data) results in a **26.27% relative increase** to base.
- Addition of non-linearity in projection heads & use of contrastive loss **(NT-Xent)** instead of NCE for instance prediction which results in a further **3.8% increase**.

Combined the above two improvements, base model can be improved by **31.3% relative increase** to base accuracy.
Experimentally the supremacy of proposed improvements can be established.
Since Contrastive learning has given SOTA results for self-supervised we further try to evaluate the feasibility of applying fully contrastive learning for smaller datasets to the given smaller network AlexNet

1 Introduction

Due to superior representation learning & automatic optimization deep learning has started dominating the ML world in the recent decade, yet the supervised deep learning methods require huge data & annotations. The rise of self-supervised learning has given the new direction to learn from unlabelled data which can reduce the cost of labeling.
The base paper is a step in the same direction, which is extended upon the Rotation Prediction by Gidaris et.al. [GSK18], Non-parametric soft max classification by Zhirong Wu et.al. [Wu+18] & the Positive Unlabeled learning [YLY17]
Recently, SimCLR v1 [Che+20a] has given state of the art results for self supervised learning using contrastive framework & many other methods like SWAV [Car+20] are improving on contrastive methodology.



(a) Maximize the similar Latents

(b) Repel Dissimilar Latents

Figure 1: Contrastive Loss Illustration, taken from SimCLR

Contrastive loss is very similar to non-parametric softmax loss [Wu+18]. Non-Parametric Softmax tries to classify the latent vector using self-correlation & contrasting with negative examples. Built on similar logic contrastive loss tries to maximize similar latent vectors while reducing the agreement between dissimilar latent vectors. Figure 1a & Figure 1b illustrates the working of contrastive loss. Slowly this learning groups similar objects in clusters resulting in superior classification performance. However, if pretraining is followed by supervised finetuning using fraction of data the classification performance is further improved as seen in SimCLRv2 [Che+20b].
We combined these above two insights to get better performance than our base model.
To solve this I first had to make few approximations & assumptions that would scale down to the low resource environment. These are explained in the section 1.1. Since the base paper is combination of three different ideas it is essential to go through a brief overview of the base paper.
To demonstrate the effectiveness of the proposed method, number of experiments are conducted with linear classification & non-linear classification. To test the generalizing capabilities we perform transfer learning experiments on unseen datasets such as MiniPlaces & CIFAR10. Although further extending modifications to mimic fully contrastive framework failed, it gave more insights into the weaknesses of the current gen contrastive learning methodologies.

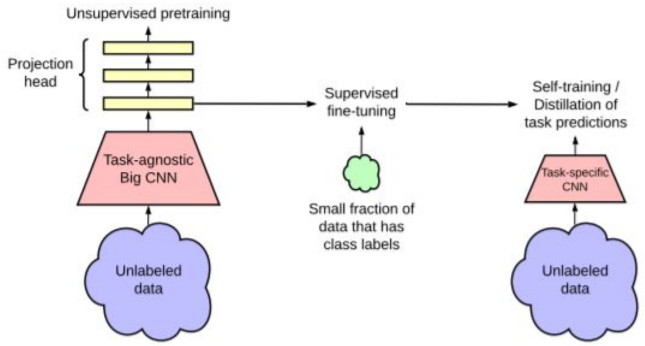


Figure 2: Process Diagram of SimCLRv2 [Che+20b]

1.1 Assumptions & Approximations

To fit the massive scale in less time & resources we use the following approximations

1. **Use Tiny-Image-Net instead of Image-Net** - Since training with Image-Net dataset 2012 (130GB) is very resource & time consuming we train with Tiny-Image-Net with only 200 classes with 80-20 split where 80,000 are training images, 10,000 are validation & 20,000 are test images
2. Use Mini Places instead of Places 205 - For the Places dataset, I use MiniPlaces with 100 classes & 10,000 training images
3. Lowering epochs from 245 to 200 for faster pretraining - I train only for 200 epochs which still takes around 12hrs
4. **Ignore PU probs** - The base paper uses PU probabilities (probability indicating whether each image is a positive unrotated image) associated with Image-net 2012 dataset obtained from pretrained RotNet(Gidaris et.al.)[GSK18]. These are applied from the epochs higher than 215. Since we are using the Tiny-Image-Net PU probabilities needs to be newly estimated as Image-net had a different number of images than Tiny-Image-Net. PU probability for each image should indicate whether that image is actual unrotated one or the rotated one (out of distribution detection), thus we need to retrain the RotNet for Tiny-Image-Net data for that reason we can't use author availed PU probs & so I ignore it in my experiments. Even when I tried to calculate PUprobs for Tiny-Image-Net from frozen RotNet, it doesn't work properly, you can see the code for non-increasing accuracy.
5. No Multiple Runs- I don't run the pretraining or downstream task multiple times to provide averaged results, but with the availability of multiple GPUs multiple training instances can be run in a mutually exclusive & parallel way.

1.2 Summary of Base Paper

Base method works by splitting AlexNet features in two equal parts to learn rotation related & unrelated features respectively.
First part is fed to a classifier who will predict the rotation class applied to the data from 0°, 90°, 180°& 270°classes.
While the other part will try to learn rotation unrelated features by minimizing the distance between each rotated feature & mean feature compounded with the instance prediction using Noise Contrastive Estimation,
A figure 3 taken from base paper illustrates the box diagram of base method.
The base paper combines below three losses in its pretraining task & divides the final feature space of 4096 into two equal splits of 2048.

1. **Rotation Prediction with PU probabilities** - the paper identifies that the rotation prediction is a good starting point to learn meaningful representations from images as proved by Gidaris et.al.[GSK18]. But the problem lies in learning from rotation agnostic images which adds to the confusion in learning as shown in Supplementary Material of the base paper [FXT19]. Some images may look the same from all angles & it's redundant to predict their rotations. Authors suggest that we identify such images & apply weights in Cross Entropy loss of rotation prediction to reduce the contribution of such rotation ambiguous images. Positive Unlabeled Learning PU learning [YLY17] states that if we treat the labeled data points as positive & unlabeled data points as negatives, a binary classification between positive & negative will not only tell whether a given point is positive or negative but also the probabilities assigned to negative data points will tell how much that datapoint is closer to known positive points. Author extend upon this methodology & treats the Image-Net datapoints as Positive Labelled Examples while their rotations as Negative Unlabeled examples & obtain probabilities of being an original or rotated image for each of the datapoint & their 4 rotated views, these probabilities are used as weights in rotation prediction cross-entropy loss as each rotated view that is closer to original view will be weighed less reducing the confusion. However even if its a good idea author only applies it after 215th epoch in total of 245 epochs. Due to dataset mismatch & failure to generate PU probs on Tiny-Image-Net we ignore this loss weighting
2. **Rotation Invariance** - The author understands that not all features can be captured through rotation prediction & therefore instills the constraint of rotation invariance where each of the features of 4 views are brought closer to mean features of these four views using MSE or Euclidean loss, forcing the network to capture rotation unrelated features.
3. **Instance Prediction** - This idea extends upon the Non-Parametric classification by Zhirong Wu et.al.[Wu+18] where the mean feature obtained in step 2 are mapped into 128-dimensional latent space vector of unit length (normalized) & these are used to predict a class from 0 to num-training-images, mapping each image to its index class. If data order is preserved by this type of mapping, similar data will be mapped closer. This is implemented through NCE (Noise Contrastive Estimation) whose code has been borrowed by the author from original author Wu et.al.

The overall Objective to minimize is

$$\zeta_c = \min_{\theta} \frac{1}{NK} \sum_{i=1}^{N} \sum_{y=1}^{K} l(F(X_{i,y};\theta), y)$$

$$\zeta_r = \frac{1}{NK} \sum_{i=1}^{N} \sum_{y=1}^{K} d(f_{i,y}^{(2)}, \bar{f}_i)$$

$$\zeta_n = -\sum_{i=1}^{N} log(P(i|\hat{f}_i))$$

$$\min_{\theta_f, \theta_c} \lambda_c \zeta_c + \lambda_r \zeta_r + \lambda_n \zeta_n$$

2 Proposed Improvements

1. **Use of Contrastive Loss (NT-Xent) instead of NCE & rotation invariance (MSE)** - Since the NCE method tries to map latent vector to data index, similar in that fashion a contrastive loss uses latent vectors of positive & negative examples to maximize the agreement between positive pairs & minimize the agreement between dissimilar pair.Contrastive loss has been successful in learning meaningful representations in recent developments. In SimCLR v1 [Che+20a], different views of the same image are brought closer, to apply similar logic, in this case, I maximize the agreement between latent vector of mean feature & latent vector of individual features of rotated views, enforcing the same rotation invariance constraint through contrastive loss. Thus each of the rotated feature needs to be normalized

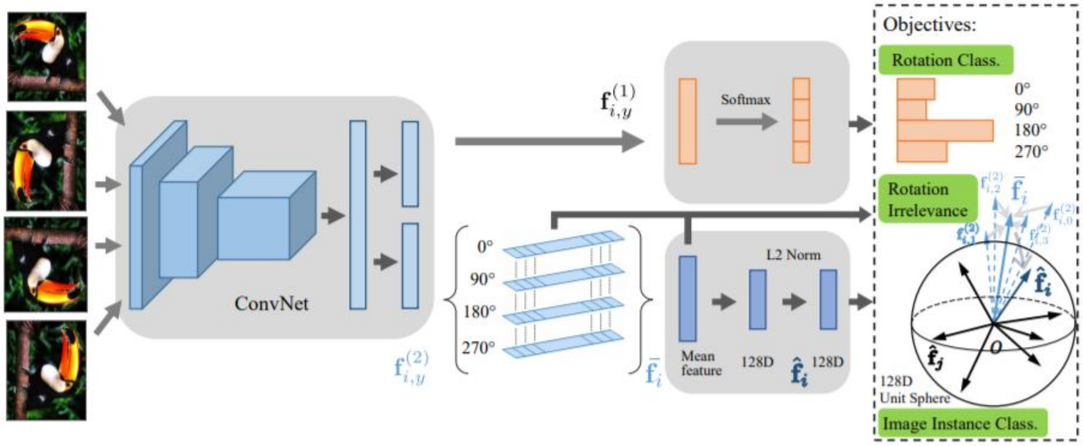


Figure 3: Block Diagram of Base Method

& mapped into a 128D vector.But unlike NCE, the contrastive loss doesn't require any memory bank to store latent vectors. We remove the MSE loss from the optimization target.

2. **Addition of non-linear layer between projection head & rotation classifier** - SimCLR [Che+20a] proves that the non-linear projection head improves representation quality for the penultimate layer, thus I introduce a non-linearity in the projection layer which converts the feature to 128D latent vector & in the rotation classifier to learn more complex non-linear relationships. We combine these two methods & call it the Contrastive method which results in a 3.8% relative increase

3. **Different Fine-Tuning method with 20% labeled data**- following SimCLRv2 [Che+20b] self-supervised models are strong semi-supervised learners, which proves that self-supervised representations are improved by a supervised fine-tuning schedule on a fraction of labeled data. We extend upon the same logic. Although in SimCLR the fine-tuning is done from last convolutional layer, where a new classifier is attached atop convolutional layer & trained using cross-entropy loss, provided feature extractor is not frozen. We employ similar technique but we differ in the placement of our classifier network, we place a linear classifier on top of MLP feature extractor in AlexNet, i.e. we don't fine-tune from the last convolution layer rather from the MLP layers. While trying to fine-tune from the last convolutional layer I observed slower convergence. Thus I decided to exploit an additional non-linear relationship between convolution & linear features learned in fc-block of AlexNet.

The new objective to minimize is

$$\zeta_c = \min_{\theta} \frac{1}{NK} \sum_{i=1}^{N} \sum_{y=1}^{K} l(F(X_{i,y};\theta), y)$$

$$l_{i,j}(contrastive) = -log \frac{exp(sim(z_i, z_j/\tau))}{\sum_{k=1}^{N} 1_{k \neq i}\, exp(sim(z_i, z_k/\tau))}$$

$$\bar{f}_i = \frac{1}{K} \sum_{y=1}^{K} f_{i,y}$$

$$\hat{f}_{i,y} = ProjectTo128D(L2Norm(f_{i,y}))$$

$$\hat{f}_i = ProjectTo128D(L2Norm(f_i))$$

$$\zeta_{contrastive} = \frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{K} l_(contrastive)(\hat{f}_{i,y}, \hat{f}_i)$$

$$\min_{\theta_f,\theta_c} \zeta_c + \zeta_{contrastive}$$

3 Experiments

- **Comparison between top-1 accuracies on Tiny-Image-Net under Linear Classification for base vs contrastive method & fine-tuned base vs fine-tuned contrastive method** - To check the effectiveness of proposed improvements we compare basemethod with new method (we call that contrastive) via LinearClassification of Tiny-Image-Net.
  Experiment can be divided in three procedures

  – **Pre train Task** - we first pre-train both the base method & contrastive method using the Tiny-Image-Net dataset. Both the models are trained for 200 epochs & rest of the optimization hyper parameters are the same as the author provided for pretraining experiment

  – **Fine-tune Task** - A new shorter dataset of 20% data points that of original Tiny-Image-Net is created for the fine-tuning task. Fine-tuning is performed for 65 epochs with the same optimization parameters as provided by the author for linear evaluation tasks but we use 10x higher learning rate as compared to the author. In fine-tune task pretrained model is updated & thus we store the new version of the pretrained model called fine-tuned model. Also, features are extracted from the fc-block of AlexNet.

  – **Down Stream Task (Linear Classifier)**- a downstream task of fitting Tiny-Image-Net labels using linear classifier atop frozen feature extraction network is performed for 65 epochs for both pretrained models. Here the features are taken from last convolutional layer in AlexNet. While we keep the same optimization parameters as what the author provided for the Linear Classification experiment.
  The results show that **contrastive method performs 3.8% relatively better** than the base method for linear evaluation without finetuning, While **fine-tuning** the base method alone gives **perf boost of 26.2% relative to base method**.
  While **Contrastive fine-tuned** performs **31.3% higher than base method.** Refer Table 1.

- **Comparison With Supervised Baseline** - To obtain the supervised baseline we train the network using 20% of labeled data for 65 epochs in a supervised fashion.
  This gives us a comparison of how much extra boost is due to the self-supervised pretraining. After training the AlexNet for 65 epochs & it matches the accuracy of 18% without converging and would catch up to the pre-trained models with few more epochs which hints that the 200 epochs pretraining isn't enough. Also the dataset size should be large enough to capture data variations. For instance, pretraining on Tiny-Image-Net & downstreaming on CIFAR10 boosts the results. See Table 2 for Tranfer Learning results.

- **Non-Linear Classifier on Tiny-Image-Net** - In most of the transfer learning practices feature extractor is frozen & a non-linear MLP or other architecture is trained for the desired task with a new dataset, meaning that the transfer relationship between the pretraining dataset & new similar dataset can be aptly mapped by non-linear layers. Thus it is essential to assess the Non-Linear Classification performance. Table 1 shows the non-linear classification performance for the base, contrastive & their respective fine-tuned models. The overall procedure is similar to Linear Classification experiment.
  As expected the result for contrastive method is always better than the base.

- **Transfer Learning Experiment on Mini Places Dataset** - A Tiny-Image-Net pretrained model is fit onto unseen places dataset to simulate transfer learning scenarios using a Linear Classifier & frozen feature extraction similar to experiment in Linear Classification.
  A base model, a contrastive model & both of their fine-tuned (on Tiny-Image-Net) models are tested on an unseen dataset of Mini Places.
  Their results are shown in the Table 2.
  Contrastive method only performs 0.8% higher than the base, while the fine-tuned versions perform only 3.5% better.
  Results prove miniplaces to be equally challenging as Tiny-Image-Net.

Table 1: Tiny-Image-Net top-1 accuracies, both of our improvements perform better

Model	Linear Classifier	Non Linear Classifier
Supervised AlexNet (65 epochs)	18.42	N/A
BaseModel	26.58	37.22
ConstrastiveModel(improved)	**27.54**	**38.19**
Fine-tuned Base	**33.56**	**40.25**
Fine-tuned Contrastive	**34.90**	**40.99**

Table 2: top-1 accuracies for transfer learning on different datasets, performance is better on smaller dataset

Model	Mini Places	CIFAR10
BaseModel	26.31	65.60
ConstrastiveModel(improved)	**26.54**	**66.46**
Fine-tuned Base	**27.25**	**67.5**5
Fine-tuned Contrastive	**27.62**	**67.7**9

- **Transfer Learning Experiment on CIFAR 10 Dataset** - To test the effectiveness of transferring to a smaller dataset we perform transfer-learning on CIFAR10.
  All four models are fit on CIFAR10 using a linear classifier & frozen layer protocol.
  All the pretrained & fine-tuned models fit with high accuracy on CIFAR10.
  Although these results show us that fine-tuned models doesn't produce any significant accuracy boost as fine-tuning might be over fitting the model on the fine-tuning dataset.
  Thus the fine-tuning should be avoided or limited if the downstream use case is for transfer learning.

3.1 A failed SimCLR experiment

We further tried to extend the contrastive learning in order to avoid feature splitting.

Maximizing agreement between full view of rotated images overfits the data & no meaningful insights are learned, Refer augmentation table by SimCLR 4 which shows rotation to be poor choice of augmentation for contrastive loss.

SimCLR touts its success based on maximizing the two different views of the same image with the the color augmentations which forces the network to learn contextual relations in the data, for instance pretext task becomes similar to in painting [Pat+16] or identifying missing puzzle pieces.

Following the same method, we removed the rotation classifier & produced data in two random crops of the same image with color jitter. Altered the design of the AlexNet feature extractor & combined projection head with feature extraction to project the image onto 128D latent vector & maximized the agreement between positive pairs using contrastive loss.

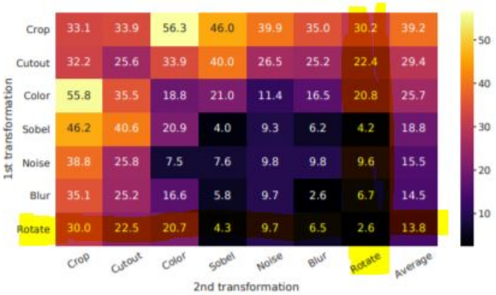
Pretrained for the same 200 epochs yet the performance of SimCLR pretrained model lagged in Primary LinearClassification task as compared to the base method.

On careful inspection, it was evident that the smallest model used in contrastive methods is ResNet50 which is much higher than simple AlexNet, even on the server with 10GB of VRAM, there wasn't any space remaining for training after loading the model.

On the other hand, the contrastive methods work best with higher epochs & larger batch-sizes because it learns slowly in absence of any guidance. Thus it requires more resources, questioning the feasibility of SimCLR for smaller datasets & resource budget environments. A more viable option would be to get a self-supervised pretrained model on Image-Net & fit that onto the custom dataset.

4 Discussions

- **Improvement by Knowledge Distillation** These results could be further improved by Knowledge Distillation, following the Knowledge Distillation by Hinton.et.al.[HVD15] & Born Again Neural Networks by Furlanello et.al.[Fur+18], knowledge extracted to the



1st transformation \ 2nd transformation	Crop	Cutout	Color	Sobel	Noise	Blur	Rotate	Average
Crop	33.1	33.9	56.3	46.0	39.9	35.0	30.2	39.2
Cutout	32.2	25.6	33.9	40.0	26.5	25.2	22.4	29.4
Color	55.8	35.5	18.8	21.0	11.4	16.5	20.8	25.7
Sobel	46.2	40.6	20.9	4.0	9.3	6.2	4.2	18.8
Noise	38.8	25.8	7.5	7.6	9.8	9.8	9.6	15.5
Blur	35.1	25.2	16.6	5.8	9.7	2.6	6.7	14.5
Rotate	30.0	22.5	20.7	4.3	9.7	6.5	2.6	13.8

Figure 4: Augmentation Table by SimCLR clearly shows rotation as augmentation to be poor choice for Contrastive Learning

same-sized model can improve the accuracy further. BANs (Born Again Neural Networks) show that the third generation self distillation gives the best accuracy boost.

- **Possible Improvement by Semi-Supervised Contrastive Learning combining contrastive loss with a fraction of labels** - A recent paper Supervised Contrastive Learning by Prannay Khosla et.al.[Kho+20] guided the contrastive loss of self-supervised learning with labels to push the SOTA on supervised learning. Since we have fewer labels available in general, a semi-supervised form of the above loss with guidance from a fraction of labels & self-supervised contrastive loss for remaining unknown data points can be combined to speed up the slow contrastive learning in self-supervised.
- **Improvements by Model Widening & Deepening** - Lately, ResNet50 has become a new standard for SOTA self-supervised methods. We can also try to widen the existing AlexNet to get extra perf & later squeeze the size using knowledge extractions. Even selecting a dense high parameterized model for feature extraction should improve the results.
- **Improvements by More Training** - As one can say due to its non-converging nature a self-supervised pretraining can be run for longer epochs to give even favourable results.
- **Possible Improvement by post processing after pretraining** - A recently developed unsupervised labeling algorithm by Gansbeke et.al. [Gan+20] applies clustering on pretrained models allowing the model to learn classification or clustering boundaries. Expanding on the same intuition, pretrained representations can be further separated into clusters or class boundaries allowing to reach almost the supervised baseline.

5 Broader Impact

Increasing data needs of current-gen supervised methods have triggered the search for smarter ways like self-supervised & unsupervised methods in deep learning which would obviate the need for a high volume of labeled data.
Even though contrastive methods are state of the art, they require stronger models & higher training hrs, making them almost infeasible for resource budget users.
There is still a burning need to find smarter ways to self-supervised learning & make it more accessible.

6 Conclusion

In the base paper, we saw how learning rotation features & rotation unrelated features can be used to learn meaningful representations where we further improved the results by replacing NCE loss with Contrastive loss & maximized the agreement between the mean of features & each rotated image feature.
By fine-tuning on 20% of labeled data results can be further boosted.
Although these are not state of the art, much of ingenuity is required to expedite the self-supervised pretraining.

References

[HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML].

[Pat+16] Deepak Pathak et al. *Context Encoders: Feature Learning by Inpainting*. 2016. arXiv: 1604.07379 [cs.CV].

[YLY17] Pengyi Yang, Wei Liu, and Jean Yang. "Positive unlabeled learning via wrapper-based adaptive sampling". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 2017, pp. 3273–3279. DOI: 10.24963/ijcai.2017/457. URL: https://doi.org/10.24963/ijcai.2017/457.

[Fur+18] Tommaso Furlanello et al. *Born Again Neural Networks*. 2018. arXiv: 1805.04770 [stat.ML].

[GSK18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. *Unsupervised Representation Learning by Predicting Image Rotations*. 2018. arXiv: 1803.07728 [cs.CV].

[Wu+18] Zhirong Wu et al. *Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination*. 2018. arXiv: 1805.01978 [cs.CV].

[FXT19] Zeyu Feng, Chang Xu, and Dacheng Tao. "Self-Supervised Representation Learning by Rotation Feature Decoupling". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[Car+20] Mathilde Caron et al. "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments". In: *arXiv preprint arXiv:2006.09882* (2020).

[Che+20a] Ting Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations*. 2020. arXiv: 2002.05709 [cs.LG].

[Che+20b] Ting Chen et al. *Big Self-Supervised Models are Strong Semi-Supervised Learners*. 2020. arXiv: 2006.10029 [cs.LG].

[Gan+20] Wouter Van Gansbeke et al. *SCAN: Learning to Classify Images without Labels*. 2020. arXiv: 2005.12320 [cs.CV].

[Kho+20] Prannay Khosla et al. *Supervised Contrastive Learning*. 2020. arXiv: 2004.11362 [cs.LG].

7 Important Hyperlinks

- https://github.com/CSAILVision/miniplaces
- http://www.Image-Net.org/
- https://roywrightme.wordpress.com/2017/11/16/positive-unlabeled-learning/
- https://www.kdnuggets.com/2019/07/introduction-noise-contrastive-estimation.html
- https://github.com/mdiephuis/SimCLR/
- https://margokhokhlova.com/index.php/archives/437