# Semi Supervised Image Classification

**Urwa Muaz (um367)**
New York University

**Shivam Kumar Pathak (skp454)**
New York University

## Abstract

Over-reliance on supervised image tasks leaves us with a dependence on large volumes of human-labeled data, a luxury that is not available across all domains. In contrast, it is comparatively easy to accumulate a large dataset of unlabeled images. Consequently, to use the unlabeled datasets, it is vital to learn meaningful representation through unsupervised training. In this study, we evaluate unsupervised training methodologies by analyzing their efficacy in a downstream image classification task with limited labeled data. We find that RotNet based pretraining outperformed the fully supervised benchmark on our labeled data, which emphasizes the utility of representations that unsupervised training can learn. We also perform an elementary exploration of variational autoencoders and a siamese regularized variational autoencoders, a variant introduced by us.

## 1 Introduction

Image classification is a process which intakes an input image and classifies it into its corresponding class. In the past decade, convolutional neural networks (CNN)[15] have dominated this domain and made tremendous progress. Most of this progress was made possible by manual labeling of large datasets like Imagenet[26]. CNN based supervised learning have established state of the art on a plethora of vision tasks including object detection[25], semantic segmentation[17], and image captioning[11].

Although supervised learning produces promising results, it requires manual data annotation, which is labor intensive and expensive. This necessity can be prohibitive in cases where massive datasets are involved, or annotation requires special skills. Furthermore, the fact that humans develop semantic visual capabilities with minimum supervision also serves as a motivation to explore unsupervised learning. Recently, some unsupervised approaches have been closing gap with their supervised counterparts in image classification.[8][2][30]. Although the performance of self-supervised representations are not at par with that of supervised representations, yet they prove to be promising alternatives in low resource settings for a variety of vision tasks such as object recognition, detection, and segmentation [31][14][32][13][3][20][21][22][4].

In this study, we investigate the efficacy of unsupervised learning from unlabelled data to build useful representations that can be used to develop classification models on low resource labeled data sets. We also investigate how the performance changes with the size of the labeled data. Additionally, we perform an elementary empirical analysis of three unsupervised pretraining approaches. For that, we pick two approaches from the literature, variational auto encoder[12] and RotNet[8], and in the third approach we propose a novel modification of VAE by introducing siamese triplet loss[27] to the latent space. Based on initial comparative performance, we choose RotNet for a more thorough experimentation and analysis. We find that unsupervised pretraining is useful and surpasses the performance of entirely supervised setup in our experiment.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

## 2 Related Work

In recent years, many promising unsupervised techniques have emerged that define an learning objective to mimic a supervised setting and learn useful representations by training on it. The design of the self-supervision task is such that it forces the network to learn structure, localization, and transformation invariance. There are three broad categories of unsupervised approaches:

**Discriminate approaches** train CNN to learn an arbitrary classification task. Examples of this type of tasks are learning the relative positions of image patches [3][20], colorizing greyscale images[31][14], or learning the geometric transformations applied on images[8]. We use rotation classification as one of the pretraining approaches in this study.

**Generative methods** include the use autoencoders to learn a latent space representation [1][18][19]. Recently, variational autoencoders [12][23] have shown a lot of promising results in learning deep visual representations. [7]. Other examples are learning generative probabilistic models[9][5][24]. From this category we choose variational autoencoders for our experiments.

**Similarity-based approaches** aim to bring semantically similar images closer in the representation space. Examples are clustering based approaches [6][16][29]. Another approach is to use siamese networks[27], which in turn use triplet loss to minimize the Euclidean distance between semantically similar images. Wang et al.[28], exploits the temporal continuity of visual world to learn similar representations for subsequent frames of videos using siamese triplet loss. In this study we propose a novel siamese regularized VAE for unsupervised pre-training.

## 3 Methodology

**Dataset** is composed of 128K labelled examples half of which are for training and other half for validation. Furthermore, we are provided 512K of unlabelled images. Data contains 1000 classes in total.

Among a plethora of promising options, we decided to do preliminary empirical evaluation of a selected group of techniques and then develop further on the one that showed most encouraging results. Since the initial results informed our methodological decisions, we will discuss those results in this section. For preliminary evaluation, we decided to use three methods of unsupervised pre-training on unlabelled data:

### 3.1 Rotation classification

For each image, we produce four rotated copies of the image, and train modified AlexNet to predict the rotation. The four rotations which serve as a supervisory signal are 0,90,180 and 270 degrees.

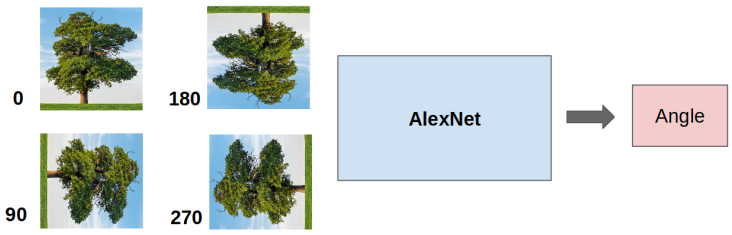


Figure 1: Rotation Classification

### 3.2 Variational Autoencoder

We use a convolutional version of VAE with the encoder architecture based on AlexNet. The network is trained using a traditional combination of KL loss and reconstruction loss[12].

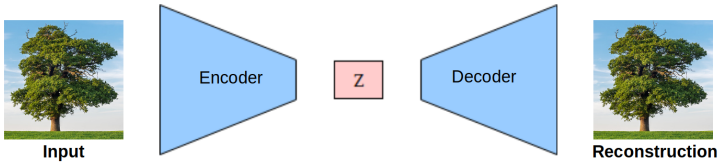


Figure 2: VAE

### 3.3 Siamese Regulated VAE

We introduce a Variational autoencoder with a siamese triplet loss on the latent space representation. KL loss and reconstruction loss, which are the usual loss terms used to train VAE inspired by the use of Siamese triplet loss for unsupervised learning from videos[28] we decided to experiment this loss term with VAE. Triplet loss needs an anchor, a positive example and a negative example[27] and tries to bring positive closer to the anchor than negative in terms of Euclidean distance in latent space. Since we did not have paired examples, we performed a random transformation (random crop, random flip, and random jittering) on the anchor image to get a positive example. The idea of maximizing the similarity between random transformations of the image for unsupervised learning is present in the literature[10].

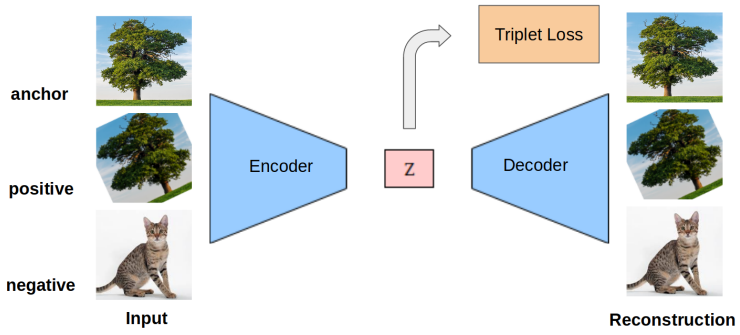


Figure 3: Siamese Regularized VAE

For simplicity and meaningful comparison, we use AlexNet based architecture for all above mentioned three tasks. We trained all the three models for 20 epochs only, which means that this preliminary analysis is not a comprehensive comparison of these approaches. For image classification, we add a neural net classifier composed of two fully connected layers and an output layer, on top of the output of fourth convolutional layer from pre-trained AlexNet which acts as a feature extractor. We train the classifier for ten epochs and use top5 accuracy to compare the results.Table1

RotNet produces the best results in our exploratory experiments, and thus, we decided to develop our solution upon it. It is interesting to note that VAE performed slightly better than VAE with triplet loss on latent space. However, these preliminary results are far from conclusive, and we think it would be interesting to explore the siamese regularized VAE in more depth. Furthermore, using latent space as features might show better results for our proposed variant because it explicitly pulls similar

Table 1: Preliminary exploration results for selected approaches.

| Model | Top 5% Accuracy |
|---|---|
| RotNet | 32.23% |
| VAE | 11.75% |
| VAE + Siamese | 10.38% |

Table 2: Final Results

| | Validation Accuracy | |
|---|---|---|
| Examples per class | Top 5% | Top 1% |
| 1 | 5.09% | 1.83% |
| 2 | 9.44% | 3.60% |
| 4 | 15.70% | 6.49% |
| 8 | 23.85% | 10.50% |
| 16 | 32.60% | 15.00% |
| 32 | 40.46% | 19.95% |
| 64 | 46.24% | 23.99% |
| **64(FineTuning)** | **50.17%** | **26.83%** |

images together in latent space. Based on these results, we decided to proceed with RotNet based architecture.

**Pretraining**:

AlexNet was trained for rotation classification using extensive data augmentation for 63 epochs. We used the hyperparameters documented by Rotnet[8] in their paper.

**Classifier Training**:

Features were extracted from the fourth convolution layer, and three fully connected layers were appended to it. These layers were randomly initialized and trained with a scheduled decreasing learning rate. Early stopping was implemented through manual supervision because we were too excited and never took our eyes off the logs. We trained seven models, each using a different number of labeled training examples per class. This was done to understand how the size of the training data influences the performance of our semi-supervised setup.

**Whole Network Fine Tuning**:

Eventually, we fine-tuned the network trained on the entire labeled data. Both feature extractor and classifier, which were separately trained before, were fine-tuned together with a small learning rate for 15 epochs. This network was submitted to the competition track.

## 4 Results

We were able to get an accuracy of 82% for pretraining on rotation classification. For classifier training, top5 accuracy saturated around the value of 46.24%, and finetuning of the entire network yielded the final figure of 50.17%. By leveraging the pretraining, we get better performance than the supervised benchmark of 40 top 5%*. As expected, the validation accuracy decreases with the decrease in labeled training data. However, the decrease in performance is not as significant as one would expect in a supervised setting. A 50% decrease in training data from 64 examples per class to 32 examples per class only results in 15% decrease in the validation accuracy. We further noticed that for 8 samples per class and below, the model overfits and yields a perfect fit for the training data. Thus, regularization of decreasing classifier complexity could have improved the performance slightly in those cases.

## 5 Conclusion

It is fairly evident that in a scenario when you do not have access to a large number of labelled images, unsupervised pretraining on unlabelled dataset can provide advantage over a fully unsupervised approach. Furthermore, the performance after unsupervised pretraining appears to be reasonably robust and does not decrease dramatically with large decreases in size of labelled data. Furthermore,our initial experimentaiton did not reveal any advantage of siamese regularized VAE over traditional VAE. But we believe that a more thorough investigation is required to investigate the usefulness of our proposed method.

## References


[1] Yoshua Bengio et al. "Greedy layer-wise training of deep networks". In: *Advances in neural information processing systems*. 2007, pp. 153–160.

[2] Mathilde Caron et al. "Deep clustering for unsupervised learning of visual features". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 132–149.

[3] Carl Doersch, Abhinav Gupta, and Alexei A Efros. "Unsupervised visual representation learning by context prediction". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1422–1430.

[4] Carl Doersch and Andrew Zisserman. "Multi-task self-supervised visual learning". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2051–2060.

[5] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. "Adversarial feature learning". In: *arXiv preprint arXiv:1605.09782* (2016).

[6] Alexey Dosovitskiy et al. "Discriminative unsupervised feature learning with convolutional neural networks". In: *Advances in neural information processing systems*. 2014, pp. 766–774.

[7] Katerina Fragkiadaki et al. "Learning to segment moving objects in videos". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4083–4090.

[8] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations". In: *arXiv preprint arXiv:1803.07728* (2018).

[9] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[10] Xu Ji, Joao F Henriques, and Andrea Vedaldi. "Invariant information distillation for unsupervised image segmentation and clustering". In: *arXiv preprint arXiv:1807.06653* (2018).

[11] Andrej Karpathy and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3128–3137.

[12] Durk P Kingma et al. "Semi-supervised learning with deep generative models". In: *Advances in neural information processing systems*. 2014, pp. 3581–3589.

[13] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. "Colorization as a proxy task for visual understanding". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6874–6883.

[14] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. "Learning representations for automatic colorization". In: *European Conference on Computer Vision*. Springer. 2016, pp. 577–593.

[15] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[16] Renjie Liao et al. "Learning deep parsimonious representations". In: *Advances in Neural Information Processing Systems*. 2016, pp. 5076–5084.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[18] Fu-Jie Huang Marc'Aurelio Ranzato, Y-Lan Boureau, and Yann LeCun. "Unsupervised learning of invariant feature hierarchies with applications to object recognition". In: *Proc. Computer Vision and Pattern Recognition Conference (CVPR'07). IEEE Press*. Vol. 127. 2007.

[19] Jonathan Masci et al. "Stacked convolutional auto-encoders for hierarchical feature extraction". In: *International Conference on Artificial Neural Networks*. Springer. 2011, pp. 52–59.

[20] Mehdi Noroozi and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles". In: *European Conference on Computer Vision*. Springer. 2016, pp. 69–84.

[21] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. "Representation learning by learning to count". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5898–5906.

[22] Deepak Pathak et al. "Learning features by watching objects move". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2701–2710.

[23] Yunchen Pu et al. "Variational autoencoder for deep learning of images, labels and captions". In: *Advances in neural information processing systems*. 2016, pp. 2352–2360.

[24] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434* (2015).

[25] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems*. 2015, pp. 91–99.

[26] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[27] Jiang Wang et al. "Learning fine-grained image similarity with deep ranking". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1386–1393.

[28] Xiaolong Wang and Abhinav Gupta. "Unsupervised learning of visual representations using videos". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2794–2802.

[29] Jianwei Yang, Devi Parikh, and Dhruv Batra. "Joint unsupervised learning of deep representations and image clusters". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5147–5156.

[30] Liheng Zhang et al. "Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data". In: *arXiv preprint arXiv:1901.04596* (2019).

[31] Richard Zhang, Phillip Isola, and Alexei A Efros. "Colorful image colorization". In: *European conference on computer vision*. Springer. 2016, pp. 649–666.

[32] Richard Zhang, Phillip Isola, and Alexei A Efros. "Split-brain autoencoders: Unsupervised learning by cross-channel prediction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1058–1067.