Semi-Supervised Image Classification

Sarthak Agarwal
sa5154@nyu.edu
Center for Data Science
New York University

Raghav Jajodia
rj1408@nyu.edu
Center for Data Science
New York University

Ieshan Vaidya
iav225@nyu.edu
Center for Data Science
New York University

Abstract

Semi-supervised learning provides a framework to incorporate unlabelled data into supervised learning tasks. Given the abundance of naturally available unlabelled data, it is an enticing prospect to utilize this data for supervised learning. Deep neural networks have achieved state of the art results on a variety of supervised learning tasks. However, the performance stagnates in the absence of large data sets. This motivates the use of semi-supervised learning techniques which incorporate unlabelled data in the learning procedure. In this work, we report our methodology in tackling the problem of image classification in semi-supervised setting. We discuss a self-supervised model that efficiently learns representations of objects from the unsupervised data which are then used to solve the supervised task.

1 Introduction

Deep convolutional neural networks have shown remarkable results on many visual tasks, one of which is image classification. Training such networks usually requires large amount of labelled data which might not be readily available. For example, the training data for the ILSVRC challenge [1] comprises of 1.2 million hand annotated images. Obtaining such large labelled data sets is not only difficult but expensive as well. On the other hand, we have access to billions of natural images and leveraging their information content can improve performance on such tasks. Semi-supervised learning is a framework that incorporates both labelled and unlabelled data in its learning. An abundance of unlabelled data, not just limited to visual tasks, would lead to wide-scale applications.

In this report, we discuss our approach and results on the competition of semi-supervised image classification. The data for the competition comprises of square color images of size 96×96 with the following structure:

- 512k unlabelled images,
- 64k labelled training images (64 examples, 1k classes),
- 64l labelled validation images (64 examples, 1k classes).

A test set of 512k labelled images is used to evaluate the performance of models. Accuracy@1 and accuracy@5 are used to determine the position on the leader-board.

2 Literature Review

2.1 Generative Adversarial Networks

Generative adversarial networks (GANs) [2] can learn reusable feature representations from unlabelled data and thus provide a straightforward extension to the semi-supervised setting. Once the GAN is trained on the unlabelled data, the labelled data can then be used to fine-tune the discriminator by adding a classifier on top. Another approach is to use these learned features to cluster the unlabelled data (along with the labelled data) into 1000 clusters. Our experiments with fine-tuning the discriminator of GANs resulted in sub-par performances in comparison to other models which are discussed in subsequent sections.

An improvement suggested by Salimans et. al. [3] for the semi-supervised setting replaces the standard real/fake classifier at top with an additional k classes (1000 in this case). The discriminator thus serves two roles: to discriminate between real and fake images and to classify real images. This requires adding an additional supervised loss term to the optimization procedure. Considering the notoriety of training GANs, our attempts to replicate the results with a standard Deep Convolutional GAN [4] architecture failed to achieve accuracy@5 beyond 25%. Given that this model shows impressive performances for small-class data sets, we believe that the discriminator is not strong enough to learn to classify 1000 classes. A possible remedy to this is to use a more powerful discriminator, particularly one which that is known to be a good classifier such as residual networks. However, based on the results we obtained and the difficulty in training GANs, we did not pursue this further.

2.2 Distillation Based Learning

Distillation refers to a class of models in which a compressed model (student), is trained to mimic the output of the larger network (teacher). It has been seen that distillation based networks perform well on semi supervised tasks.[5,6]. Our architecture for the teacher/student model is based on the implementation in [5]. Initially a large teacher network is trained over the supervised data. For every class, the predictions of this teacher model are used to rank the unlabelled images and top-K images are chosen to form a new data set. A smaller student model is then trained over the new constructed data set. Finally, the student model is fine-tuned over the labelled data. We observed an accuracy@5 of 33.5% on the validation set which is relatively poor. We believe that the lack of data was the primary reason for the sub-par performance. The teacher model performs poorly which results in poor accuracy of the student model.

2.3 Self Supervised Learning

Among the techniques to learn high level feature representations, a prominent approach is self supervised learning that defines a separate task, using the unlabelled data, in order to provide a signal for the classification task. We used the task of predicting image rotations [7] to extract useful features from the unlabelled data set.

In the following section we explain our self supervised methodology followed by our results. Finally we provide conclusions and future work.

3 Rotation Learning

3.1 Overview

The goal of this project was to extract useful features from the unlabelled images to finally help in the classification task. To achieve this, we train a model $F(.)$ to perform classification on a dataset generated using transformations. We define a set of K transformations $G = \{g(.|y)\}_{y=1}^{K}$ where $g(.|y)$ represents the transformation function which when applied to image X produces a transformed image $X^y = g(x|y)$ with label y. The model $F(.)$ receives X^{y^*} as an input and is trained to predict the label y^*. The model generates a probability distribution over the K transformations:

$$F(X^{y^*}|\theta) = \{F^y(X^{y^*}|\theta)\}_{y=1}^{K}$$

where $F^y(X^{y^*}|\theta)$ represents the probability that the model assigns to the yth class and θ represents the model parameters.

The model $F(.)$ is trained using the unsupervised dataset $\{X_i\}_{i=1}^{N}$. The following cost function is optimized:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \left[-\frac{1}{K} \sum_{y=1}^{K} log(F^y(g(X_i, y)|\theta)) \right]$$

The geometric transformations G should be such that the features captured help in the final classification task. We choose the set G as all the image rotations by multiple of $90°$ as in [7].

The main intuition behind using rotations as the transformation is that, to correctly predict the angle of rotation, the model first needs to recognize and detect classes of objects in the image.

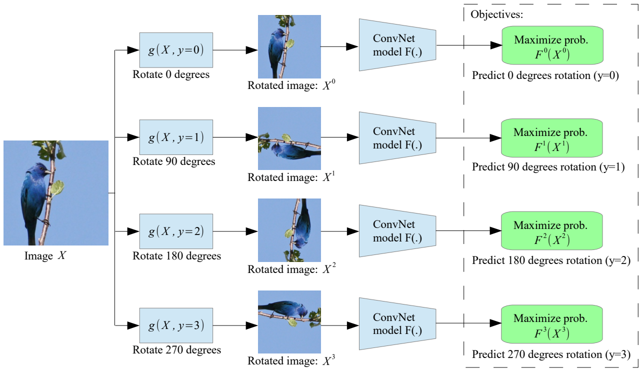


Figure 1: Rotation Learning (adapted from [7])

3.2 Methodology

The 512k images from the unsupervised dataset were randomly rotated by $0°, 90°, 180°$ or $270°$ and labelled as $0, 1, 2$ and 3 respectively. We replaced the last layer of a traditional ResNet-34 architecture for 4-way classification. This model was used to fit the data using cross entropy loss.

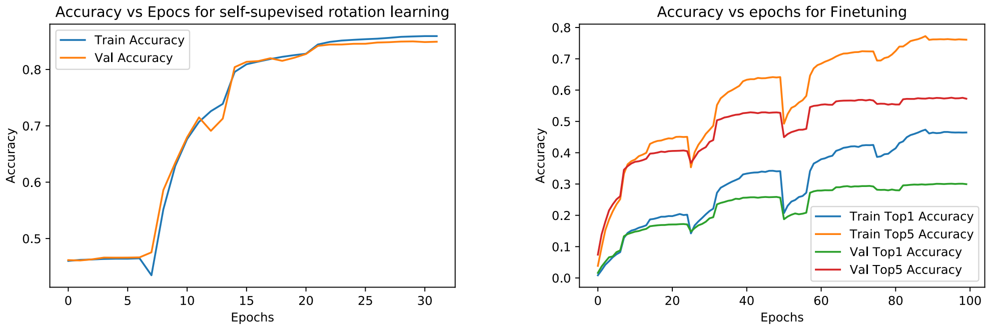
For fine-tuning, we again replaced the last layer for 1000-class classification. The model is fine-tuned using the supervised data.

Also, a separate ResNet-34 model was trained using only the supervised dataset. To avoid overfitting, data augmentation techniques like adding Gaussian noise to the input layer, random resized crop and random horizontal flip were used.

Finally an equal weight ensemble of supervised ResNet-34 model and fine tuned rotation learning model was created to improve the validation accuracy.

4 Results and Summary

Figure 2a shows the accuracy for the 4-way classification task in rotation learning. The performance over both, the training and validation set saturates around 90% accuracy. This model is then fine-tuned for the original image classification task. Figure 2b shows the training and validation accuracies (top-1 and top-5) for the fine-tuning task.



(a) Rotation learning

(b) Fine-tuning

Figure 2: Rotation learning accuracies

To evaluate the dependence of the ensemble model on size of supervised data, we report the validation accuracies for training with 1, 2, 4, 8, 16, 32 and 64 labelled samples per class. As expected, having more labelled samples per class elevates the performance. Table 1 shows the overall results we obtained for all the models.

Samples	Supervised		Rotation		Ensemble	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
1	0.30	1.80	0.80	3.10	0.35	1.57
2	0.65	2.20	1.55	5.85	0.53	2.64
4	0.67	3.10	1.47	6.68	0.68	3.31
8	1.03	4.81	3.30	11.60	2.19	8.56
16	3.21	10.53	6.95	21.22	5.14	18.05
32	6.34	20.92	11.64	31.50	9.68	28.66
64	29.04	53.69	29.96	57.50	32.38	59.25

Table 1: Accuracies (%) for all models

4.1 Conclusions and Future Work

We achieved a top-1 accuracy of 32.38% and a top-5 accuracy of 59.25% using the ensemble method. Looking at the performance of the rotation task, there seems to be a margin of improvement. This can be done by designing the transformations sophistically. Joint training on self-supervised tasks like jigsaw training and colorization can also improve performance.

Acknowledgments

We would like to thank Professor Yann LeCun for his invaluable insight and guidance throughout the course. We would also like to thank Alfredo for his witty and informative teaching. We thank Katrina along with all our fellow students for their support and discussion.

References

[1] Russakovsky, Olga and Deng, Jia and Su, Hao and Krause, Jonathan and Satheesh, Sanjeev and Ma, Sean and Huang, Zhiheng and Karpathy, Andrej and Khosla, Aditya and Bernstein, Michael and Berg, Alexander C. and Fei-Fei,Li (2015) ImageNet Large Scale Visual Recognition Challenge. In *International Journal of Computer Vision (IJCV)*

[2] Goodfellow, Ian and Pouget-Abadie, Jean and Mirza, Mehdi and Xu, Bing and Warde-Farley, David and Ozair, Sherjil and Courville, Aaron and Bengio, Yoshua (2014) Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc.

[3] Salimans, Tim and Goodfellow, Ian and Zaremba, Wojciech and Cheung, Vicki and Radford, Alec and Chen, Xi and Chen, Xi (2016) Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems 29*, pp. 2234–2242. Curran Associates, Inc.

[4] Radford, Alec and Metz, Luke and Chintala, Soumith (2015) Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *ICLR*.

[5] Yalniz, I. Zeki and Jégou, Hervé and Chen, Kan and Paluri, Manohar and Mahajan, Dhruv (2019) Billion-scale semi-supervised learning for image classification.

[6] Tarvainen, Antti and Valpola, Harri (2017) Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30*, pp. 1195–1204. Curran Associates, Inc.

[7] Gidaris, Spyros and Singh, Praveer and Komodakis, Nikos (2018) Unsupervised Representation Learning by Predicting Image Rotations. In *ICLR*.