Deep Learning Final Project

Kuixian Zhu
Center for Data Science
New York University
New York, NY 10012
kz1041@nyu.edu

Marvin Mananghaya
Center for Urban Science and Progress
New York University
New York, NY 11201
msm796@nyu.edu

Daniel Rivera Ruiz
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012
drr342@nyu.edu

Abstract

During the last years, semi-supervised learning has emerged as an exciting new direction in machine learning research. It is closely related to profound issues of how to do inference from data and to the situation where relatively few labeled training points are available, but a large number of unlabeled points are given. In this project, we explore several semi-supervised learning techniques in the context of image classification.

1 Introduction

Traditionally, there have been two fundamentally different types of tasks in machine learning: unsupervised and supervised learning. The goal of unsupervised learning is to find interesting structure in unlabeled data X, while in supervised learning the goal is to learn a mapping from X to Y, given a training set made of pairs (x_i, y_i). Here, the $y_i \in Y$ are called the labels or targets of the examples x_i [1].

Semi-supervised learning is halfway between supervised and unsupervised learning. In addition to unlabeled data, the algorithm is provided with some supervision information – but not necessarily for all examples. Often, this information will be the targets associated with some of the examples but other forms of partial supervision are possible, e.g. there may be constraints such as "these points have (or do not have) the same target".

Within the field of semi-supervised learning, we can distinguish four main classes of algorithms:

1. Generative Models. Inference using a generative model involves the estimation of the conditional density $p(x|y)$. In this setting, any additional information on $p(x)$ is useful. A strength of the generative approach is that knowledge of the structure of the problem or the data can naturally be incorporated by modeling it.
2. Low-density Separation. These algorithms aim to push the decision boundary away from the unlabeled points. The most common approach to achieving this goal is to use a maximum margin algorithm such as support vector machines. The method of maximizing the margin for unlabeled as well as labeled points is called the transductive SVM (TSVM). However, the corresponding problem is nonconvex and thus difficult to optimize.
3. Graph-based Methods. The common denominator of these methods is that the data are represented by the nodes of a graph, the edges of which are labeled with the pairwise distances of the incident nodes (and a missing edge corresponds to infinite distance).

4. Change of Representation. These algorithms are not intrinsically semi-supervised, but instead perform two-step learning. First they perform an unsupervised step on all data ignoring the available labels. This can, for instance, be a change of representation, or the construction of a new metric or a new kernel. Then they ignore the unlabeled data and perform plain supervised learning using the new distance, representation, or kernel.

In this project we explored three semi-supervised techniques to address the image classification problem: (1) Semi Supervised Learning - Deep Convolutional Generative Adversarial Networks (SSL-DCGANs), which can be considered as a generative model, (2) the Mean Teacher method, which can be considered as belonging to the low-density separation category, (3) and the Deep Clustering algorithm, which falls into the change of representation category.

2 Experiments

2.1 Baseline

The baseline experiment is to see the performance of image classification without taking advantage of unlabeled data to inform the representation learning of images. For this experiment, we made use of ResNet152 as the underlying architecture. It ran for 90 epochs, with a batch size of 32 and with default hyperparameter settings specified in the original paper [2]. Since ResNet152 required that images to be of size 224x224, the images were resized to match this specification. All these resulted to producing a top-1 and top-5 accuracy in the validation set of 30.45% and 55.20%, respectively.

2.2 SSL-DCGAN

Deep Convolutional Generative Adversarial Networks, particularly it's discriminator, have shown properties capable of unsupervised learning as described in [3]. [4] makes this case, if that the discriminator is made to classify $K + 1$ classes, where K is total number of classes of real images and the other being a fake image, then it should like any image classifier. This allows it's discriminator network to act like a standard classifier of images while learning representations from unlabeled images.

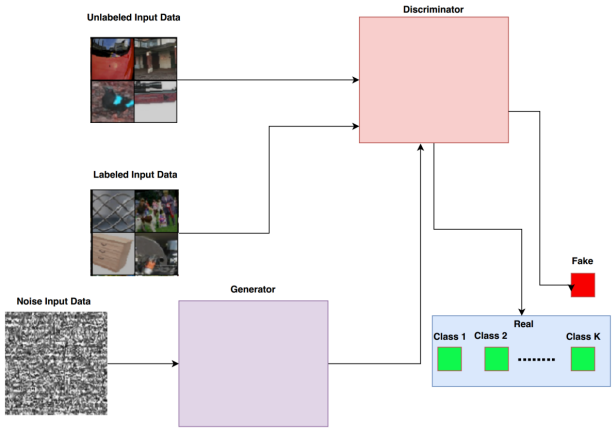

Figure 1: Architecture of the SSL-DCGAN

For this experiment, The code used for this experiment was adapted from [5]. It ran for 20 epochs as the the model quickly suffered non-convergence. It ran with a batch size of 32 and with default hyper-parameter settings specified by the authors. Data augmentation transformations is also applied to the input images such as image rotation, normalization,image flipped and color jittered. All these resulted to producing a top-1 and top-5 accuracy in the validation set of 10.10% and 27.04%, respectively.

GANs are known to be difficult to train, as explored further by [6]. The model used in this experiment suffered non-convergence at around 20 epochs, where accuracy kept oscillating at around $17.24\% -$

17.89%. In previous experiments, we've encountered mode collapse or one network overpowering another due to slight changes. It maybe worth while to explore various architectures, regularization or loss functions. However, it might also be better to explore simpler frameworks in conducting semi-supervised learning considering these difficulties.

2.3 Deep Clustering

Clustering is a class of unsupervised learning methods that has been extensively applied and studied in computer vision. However, little work has been done to adapt it to the end-to-end training of visual features on large scale datasets. On the other hand, pre-trained convolutional neural networks have become the building blocks in most computer vision applications: they produce excellent general-purpose features that can be used to improve the generalization of models learned on a limited amount of data.

Deep Clustering is a novel approach for the large scale end-to-end training of convnets that obtains useful general-purpose visual features within a clustering framework. This approach, summarized in Figure 2, consists in alternating between clustering of the image descriptors and updating the weights of the convnet by predicting the cluster assignments.

For this project we used the training code provided by [7], using all the unlabeled data (512k images) as our training set. We used AlexNet as the underlying architecture of our convnet, 1,000 clusters for the k-means algorithm, normalization of the input images and a mini-batch size of 256. We trained the convnet for 30 epochs and the classifier (linear layers) on top of it for another 30 epochs. All other hyperparameters were set to the default values suggested in the original paper. With this setup, our Deep Clustering model yielded top-1 and top-5 accuracies in the validation set of 13.48% and 28.60%, respectively.

We believe that given more time to train the model over a larger number of epochs, these results could improve substantially (as reference, the original paper suggests training the convnet for 200 epochs and the classifier for 90 epochs).

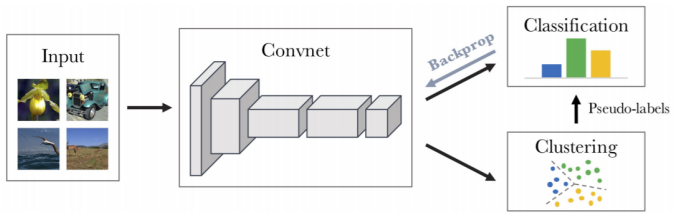

Figure 2: Illustration of the Deep Clustering Method

2.4 Mean-Teacher

Mean Teacher [8] is a simple framework for semi-supervised learning. It consists of the following steps:

1. Take any supervised model, call it the **student**.
2. Make a copy of the student model, call it the **teacher**. Let the teacher's weights to be the exponential moving average (EMA) of the student's weights.
3. While training, use the same images as inputs to both models, but add random noise (augmentation) to them separately.
4. Update the student by the classification loss as usual, yet add an additional *consistency loss* between the student and teacher outputs.

2.4.1 Training

Data Augmentation As described in [9], we randomly augmented images using a 10 degree rotation, a crop with aspect ratio between 3/4 and 4/3, a random horizontal flip and a color jitter. We also re-sized the images to 224x224 and normalized the inputs to have channel-wise zero mean and unit variance over the training set.

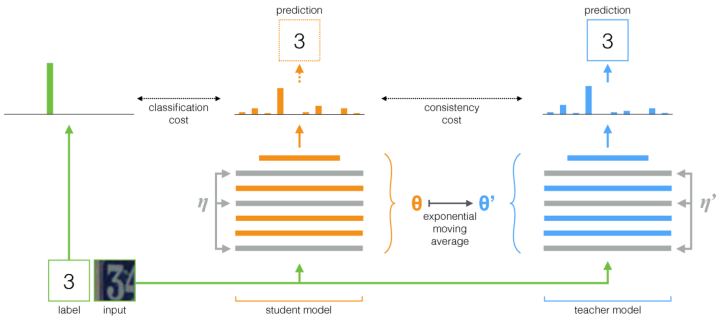

Figure 3: Mean Teacher Framework

Hyper-parameters For the Mean Teacher framework, We used Resnet152 as the underlying architecture. For the final submitted model, We trained the network using stochastic gradient descent (SGD) with initial learning rate 0.02 and Nesterov momentum 0.9, with a L2 weight decay with coefficient 1e-4. We trained the network for 90 epochs (due to limited time and computing resources), decaying the learning rate with cosine annealing [10]. We used the dual output trick described by the original paper with MSE cost between logits with coefficient 0.01. We used the KL consistency cost, and a EMA decay value of 0.999.

Results Table 1 includes the results for our Mean Teacher experiments. Our best model (Mean Teacher Resnet152 trained with 64 samples per class) achieves a top-1 accuracy of 34.38%, which beats the performance of vanilla Resnet152 (30.45%). Thus, we argue that the Mean Teacher framework can help when labels are scarce by using the consistency loss.

2.4.2 Analysis of Training Curves

As shown in Figure 4, the Mean Teacher Resnet152 model learns faster than the baseline Resnet152 model, and eventually converges to a better result. This result can show the power of the unsupervised consistency loss regularization.

Also, the EMA-weighted model (teacher model) gives more accurate predictions than the student model after an initial period. This is consistent with the claim in the original paper that using the EMA-weighted model as the teacher improves the student via the consistency cost..

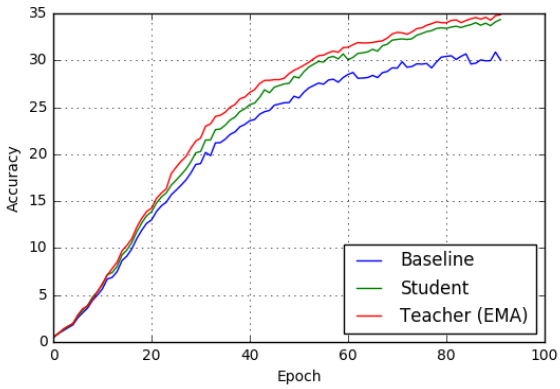

Figure 4: Mean Teacher Training Curves

2.4.3 Strength and Weakness

The major strength of the Mean Teacher framework is that the approach scales to large datasets and on-line learning. Also, this framework is not only simple and straightforward, but also extremely flexible. Actually, it is convenient for us to replace the resnet152 in our model by any other models, ranged from simple linear models to deep neural networks.

On the other hand, one limitation of the model is that it still requires the teacher model to at least 'learn' to some extent, or there will be the hazard of confirmation bias. As we can observe in Table 1, the Mean Teacher method does not work when the number of labeled data per class is very small (1 or 2). One possible explanation can be that the consistency loss can hardly work if there is not enough data for the teacher model to even learn anything, and thus it cannot really 'teach' the student model (or even worse, teach in the wrong direction).

Table 1: Top-1 accuracies measured on the validation set for the baseline ResNet152 model and our three SSL models: DCGAN, Mean Teacher and Deep Clustering. For every experiment, we present each model only with a subset of the training data: we start at 1 labeled example per class and duplicate this number until we fully utilize the whole training set (64 examples per class). For the experiments that utilize the whole training set we also show top-5 accuracies. Empty entries in the table mean that we did not run that experiment due to resource limitations.

	Labeled examples per class during training							
	1	2	4	8	16	32	64	64
Model							acc@1	acc@5
Baseline (Resnet152)	–	–	–	–	–	–	30.45	55.20
SSL-DCGAN	–	–	–	–	–	–	10.10	27.04
Mean Teacher (Resnet152)	1.71	1.65	–	–	15.21	20.59	**34.38**	**59.81**
Deep Clustering	1.40	2.38	3.38	5.60	7.71	9.39	13.48	28.61

3 Conclusion and Future Work

In this project, we experimented with SSL-DCGAN, Mean Teacher, and Deep Clustering as semi-supervised learning methods. From our experiment results, we have seen the power and limitations of semi-supervised learning. However, unfortunately, due to the limitation of time and computing resources, we were not able to really push any of these methods to the limit, which is definitely a work that can be done in the future.

References

[1] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-Supervised Learning*. The MIT Press, 1st ed., 2010.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *ICLR*, 2016.

[4] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *CoRR*, vol. abs/1606.03498, 2016.

[5] B. Lecouat, C. S. Foo, H. Zenati, and V. R. Chandrasekhar, "Semi-supervised learning with gans: Revisiting manifold regularization," 2018.

[6] N. Kodali, J. D. Abernethy, J. Hays, and Z. Kira, "How to train your DRAGAN," *CoRR*, vol. abs/1705.07215, 2017.

[7] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," *CoRR*, vol. abs/1807.05520, 2018.

[8] A. Tarvainen and H. Valpola, "Weight-averaged consistency targets improve semi-supervised deep learning results," *CoRR*, vol. abs/1703.01780, 2017.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *CoRR*, vol. abs/1709.01507, 2017.

[10] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with restarts," *CoRR*, vol. abs/1608.03983, 2016.