CSCI-GA 2572 Deep Learning Sping 2021
Semi-supervised Learning Competition Report-Team 02

Yang Zhou[1] **Qingfu Wan**[2] **Wenjie Li**[2]

1. Introduction

We present an approach for image classification with only 0.5% labeled training data as our final project to Deep Learning course in Spring 2021 at New York University. Our approach utilizes unsupervised learning, pseudo-label iterations, semi-supervised learning and active learning methods. Our final result achieves 55.80% accuracy on the test dataset with 0.5% labeled data and 57.19% accuracy on the test dataset with requested additional 0.25% labeled data. Sec. 2 reviews related literature, Sec. 3 describes our methods in details, and finally Sec. 4 presents an analysis of the model's performance.

2. Related Works

2.1. Unsupervised Learning Methods

Unsupervised learning exploits the information in unlabeled data to regularize downstream supervised tasks. One recent trend is to use discriminative approaches based on contrastive learning, where embeddings from related images are optimized to be closer than embeddings from unrelated images. SimCLR (Chen et al., 2020a) maximizes the agreement of representations under different data transformations. MoCo (He et al., 2020) maintains a ”momentum” network to output negative examples. To reduce the requirement of large batch size, MoCo v2 (Chen et al., 2020c) integrates two design principles in SimCLR into MoCo. SimCLRv2 (Chen et al., 2020b) borrows the memory mechanism from MoCo, and exploits larger ResNets and deeper projection head. Contrastive methods often require myriads of negative examples, and hence large batch size (Chen et al., 2020a) or extra memory (He et al., 2020). This raises the question as to whether negative pairs are indispensable.

One way to learn representations without explicit negative examples is to use clustering. SwAV (Caron et al., 2020) replaces previous expensive feature comparisons with clustering consistency among different views of the same image. The underlying assumption is that the representation of an augmented view should be predictive of the representation of another augmented view of the same image. Under this, a concurrent work BYOL (Grill et al., 2020) distills the knowledge from the target network to teach the online network.

The above paradigms $i.e.$ contrastive learning, clustering and distillation optimize a similarity maximization objective to avoid trivial solutions. As opposed, a very recent work Barlow Twins (Zbontar et al., 2021) makes another attempt at learning a redundancy reduction objective. Barlow Twins constructs a cross-correlation matrix from the representations of two different views of the same image while encouraging this matrix to be diagonal. We utilize Barlow Twins as our unsupervised feature embedding method.

2.2. Semi-supervised Learning Methods

Semi-supervised methods consider both an unsupervised loss on a vast amount of data and a classification loss over a few labeled data. The one that is closely related to ours is Pseudo-labeling (Lee et al., 2013) where the class from the maximum predicted probability is treated as the true label. In Sec. 3.2 we detail the adaptation of this method.

Regarding large-scale semi-supervised training, (Yalniz et al., 2019) employs a ranking strategy to teach the student network with unlabeled images trained on the labeled dataset. Temporal Ensembling (Laine & Aila, 2016) stores an exponential moving average (EMA) of label predictions on each training sample, which we find helpful to smooth the model at the last stage of training. Instead of averaging label predictions, Mean teacher (Tarvainen & Valpola, 2017) takes the average of model weights.

To summarize, we adopt the pseudo-label selection protocol from (Yalniz et al., 2019) and utilize FixMatch as our semi-supervised learning method. In Sec. 3.3, we discuss these methods in detail.

2.3. Active Learning

The key to active learning is to design precise query strategies that select samples to maximize the classification accuracy at a given point. One direction Heterogeneity-Based

[1] Department of Electrical and Computer Engineering, New York University, Brooklyn, USA [2] Courant Institute of Mathematical Sciences, New York University, New York, USA.

Models learn the regions that show the greatest heterogeneity $e.g.$ in terms of the uncertainty of classification (Sensoy et al., 2018). In this regime, we use margin of confidence as the uncertainty sampling method for active learning.

3. Methods

In this section, we introduce our method which combines techniques from unsupervised learning, pseudo-label iterations, semi-supervised learning, and active learning. Sec. 3.1 explains how to obtain feature embeddings from unsupervised learning. Sec. 3.2 introduces a balanced pseudo-label iteration approach which generates pseudo-label from unlabeled data. Sec. 3.3 describes a method to further improve the performance by consistency regularization with semi-supervised learning. Sec. 3.4 shows our active learning approach which selects extra samples to be labeled in order to obtain higher accuracy. Last but not least, We discuss the effect of test-time augmentation in Sec. 3.5.

3.1. Feature Embedding from Unsupervised Learning

We adopt Barlow Twins (Zbontar et al., 2021) to obtain a feature representation that is invariant to distortions of the input samples. For each image, we create two distorted images by applying different image transformations. We train a neural network that produces feature representation Z^A, Z^B for these two samples. The goal is to encourage the empirical cross-correlation matrix between two representations from different distortions to be an identity matrix. Specifically, The loss function includes two parts: an invariance term and a redundancy term.

$$\ell_{BT} = \sum_i (1 - \ell_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \ell_{ij}^2 \tag{1}$$

$$\ell_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b (z_{b,i}^A)^2} \sqrt{\sum_b (z_{b,j}^B)^2}} \tag{2}$$

The invariance term $\sum_i (1 - \ell_{ii})^2$ encourages the diagonal element of the cross-correlation matrix ℓ to be 1, and the redundancy reduction term $\sum_i \sum_{j \neq i} \ell_{ij}^2$ encourages other elements than the diagonal entries to be zero. Positive λ controls the ratio between these two terms.

Without the need of negative samples, Barlow Twins learns a meaningful feature embedding from unlabeled data. We use the backbone of ResNet-50 with three 8192-dimensional fully connected projection layers as the neural network. We train the network for 1000 epochs with learning rate set to 0.5, $\lambda = 0.0051$, and the same data augmentation adopted in the original paper.

After we obtained the feature embedding, we attach a fully connected layer as the linear classifier behind the ResNet-50 backbone (without three projection layers), and we finetune the network on the labeled training dataset with 0.5% images. We use 0.05 for both learning rates of the backbone and the linear classifier, and train the network for 40 epochs with zero weight decay. This procedure obtains a 46.67% top-1 accuracy on the validation dataset with only unsupervised learning.

3.2. Self-training with Pseudo-label Iterations

To further improve the performance, we apply a self-training approach with pseudo-label iterations. The idea introduced by (Yalniz et al., 2019) is to generate label-balanced pseudo-labels which are consistent with the label-balanced training dataset. We follow the same protocol and keep iterating the process until the performance reaches a bottleneck. In our setting, we have $N = 512000$ unlabeled data and $C = 800$ classes. For each iteration, we use the current best model to predict the class distribution of each sample. We pick top-K classes for each sample to obtain $K \times N$ pairs of images and pseudo-labels. In this step, K pseudo-labels correspond to each image. We choose $P = 10$ in this work. Next, we pick top-P samples for each class to create $P \times C$ pairs of images and pseudo-labels.

Ideally, if the model is trained well and P is less than the minimum number of samples among all classes, each image should only correspond to one pseudo-label. However, due to the imperfect model and class imbalance, we should expect to see some images correspond to multiple pseudo-labels. To keep the quality of the pseudo-label, we set $P = 200, 300, 400, 500$ for the total four iterations. The more we train the model with pseudo-labels, the higher the quality of the pseudo-labels generated by the model will be.

For the self-training of the pseudo-label iterations, we finetune the current best model with selected pseudo-labels and the labeled training dataset first, and we further finetune the model with only the labeled training dataset. For both finetune processes, we train the linear classifier from random weights. The self-training with pseudo-label iterations gives us a 5% boost on the top-1 accuracy which adds to a total of 52.38% top-1 accuracy on the validation dataset.

3.3. Consistency Regularization of Semi-supervised Learning

We apply FixMatch (Sohn et al., 2020), a semi-supervised learning method to further improve the robustness of the network. FixMatch generates label distributions from both a weakly-augmented version and a strongly-augmented version of the same image. It uses the class with the highest prediction probability from the weakly-augmented version as the pseudo-label and minimizes the cross entropy between the pseudo-label and the prediction from the strongly-augmented version. During the training of FixMatch, we take samples both from the labeled and unlabeled dataset.

We use $\alpha(\cdot), \mathcal{A}(\cdot)$ as the weakly and strongly transformation. For the samples from the labeled training dataset, we use a standard cross-entropy loss ℓ_s. p_b is the ground-truth label, u_b is the sampled image, and $p_m(y|\alpha(x_b))$ is the prediction of weakly-augmented version of the image.

$$\ell_s = \frac{1}{B}\sum_{b=1}^{B} H(p_b, p_m(y|\alpha(x_b))) \tag{3}$$

$$\ell_u = \frac{1}{\mu B}\sum_{b=1}^{\mu B} 1(\max(q_b) \geq r) H(\hat{q}_b, p_m(y|\mathcal{A}(u_b))) \tag{4}$$

For the unlabeled sample, we use the cross-entropy loss ℓ_u. $q_b = p_{\mathrm{m}}(y|\alpha(u_b))$ is the predicted class distribution of the weakly-augmented version of the unlabeled image. We use $\hat{q}_b = \arg\max(q_b)$ as the pseudo-label from the weakly-augmented image.

We combine ℓ_s, ℓ_u into one loss $\ell_s + \lambda_u \ell_u$ as the final loss where λ_u is the positive hyper-parameter ratio between these two losses. We train the network from the current best model for 40 epochs with learning rate set to 0.01 and exponential moving average with decay 0.999. Neesterov Momentum optimizer with momentum 0.9 and cosine annealing learning rate scheduler is applied. In addition, we use RandAugment with a random magnitude as the strong augmentation and random horizontal flip as the weak augmentation. The labeled dataset consists of the labeled training dataset and $(K = 10, P = 16)$ pseudo-labels from Sec. 3.2; the unlabeled dataset includes all $N - (P \times C)$ unlabeled images. We obtain 54.46% top-1 accuracy on the validation dataset after training with the semi-supervising approach FixMatch.

3.4. Active Learning

In order to compile a request for additional (0.25%) labels, we combine uncertainty sampling and diversity sampling. Our diversity sampling method makes use of cluster-based sampling. We use K-Means clustering with cosine similarity. We select the clustering number as 200 because from our observation, there are in total of 800 categories in the dataset, many of which are similar to each other. Although some categories do not have repetitive categories, categories that belong to a higher level category should also be grouped. We use the embedding trained in the unsupervised learning process as the feature of each sample and apply the K-means algorithm. After grouping 512,000 samples into 200 clusters, we select 64 samples based on the uncertainty sampling method: Margin of Confidence sampling, an approach that utilizes the difference between the two most confident predictions, is chosen as the uncertainty sampling method. We use FixMatch to finetune the current best model with labeled training dataset and the additional extra label (0.75% data) to obtain 56.07% top-1 accuracy on the validation dataset.

Table 1. Effect of multi-scale test-time augmentation on the validation dataset.

LABELED DATA	TOP-1 W/O. TTA	TOP-1 W. TTA
0.5% DATA	54.56%	56.02%
0.75% DATA	56.07%	57.54%

Table 2. Ablation Study on the validation dataset

MODEL	TOP-1
UNSUPERVISED LEARNING	46.57%
PSEUDO-LABEL ITERATIONS	52.38%
SEMI-SUPERVISED LEARNING	54.46%
ACTIVE LEARNING	56.07%
TEST-TIME AUGMENTATION	57.54%

3.5. Test-time Augmentation

As the last step, we use multi-scale inference test-time augmentation to boost the performance of the network. We average the prediction distribution of the three scales of model inference, (96×96), (144×144), and (192×192), to obtain the final prediction. The effect of the test-time augmentation is analyzed in Tab. 1.

4. Evaluation

4.1. Ablation Study

Tab. 2 shows the Top-1 accuracy of each step. The largest improvement comes from self-training with pseudo-label iterations. Since the active learning task is based on a sub-optimal model with only 18% accuracy in order to meet the label request deadline, it does not improve the model by a large margin. We also show the Top-1 accuracy of our method on the test dataset in Tab. 3 which outperforms the runner-up team by 6.5%.

4.2. Effect of the Amount of Pseudo-label

In order to choose a suitable set of pseudo-labels, we apply the following procedure: our best model generates the top $K = 10$ most probable labels for each sample. For each label, P number of samples with highest probabilities are selected from the pool of $10 \times 512,000$ samples generated from the above-mentioned method. We argue that the less duplicates labels there are $i.e.$ an image being sampled to more than one label, the more confident the classification is. We conduct this experiments with the best model. When $P = 300$, there are 22 cases of duplicate labels, making up less than 0.01% of all selected samples; when $P = 400, 500$, and 625, the number of duplicate labels rise to $845(00.27\%)$, $8,043(2.02\%)$, and $44,570(8.91\%)$, respectively. Hence we conclude that $P \in [300, 400]$ is a

Table 3. Top-1 accuracy on the test dataset.

MODEL	TOP-1 (%)
0.5% DATA	55.08%
0.75% DATA	57.19%

Table 4. Ablation Study of Pseudo-label Iterations on validation.

ITERATION	P	TOP-1
1	200	47.91%
2	300	47.98%
3	400	51.53%
4	500	52.38%

good borderline for choosing pseudo-labels. In the early training stage with pseudo-labels, we use a smaller P number since the model is not confident, and we increase P to include more pseudo-labels incrementally at each training iteration. The result of training at each iteration with incremental P value is shown in Tab. 4

4.3. Learning Rate Searching

Considering that the linear classifier is prone to overfitting, we test the performances of different combinations of backbone and classifier learning rates besides the chosen $(0.05, 0.05)$ during the training process. The result is shown in Tab. 5. Increasing the classifier learning rate will lead to longer convergence time whereas a lower classifier learning rate leads to local convergence at an early stage of training. We conclude that the original choice is optimal. The use of weight decay does not show a significant improvement.

4.4. Feature Map Visualization

Given the input dog image, we visualize feature maps of the last layer within each residual block of ResNet-50 across different models in Fig. 1. As the model becomes more advanced, the feature maps are sparser implying more regularization. This is especially true when the model flows to higher layers, where global features are captured are captured from the perspective of semantics. We observe that feature maps of advanced models are more focused on key parts of the dog $e.g.$ smooth hair, small head, and legs, whereas feature maps from Barlow Twins have activations all over the entire image region. This indicates that the inclusion of pseudo-labels and FixMatch is conducive to attaining clearer and more distinctive features, which the baseline model has difficulty achieving.

Table 5. Top-1 accuracy with combinations of different backbone and classifier learning rates, weight decay = 1.5 when applicable.

LR-BACKBONE	LR-CLASSIFIER	TOP-1 W/O. WD	TOP-1 W. WD
0.03	0.05	48.59%	48.63%
0.01	0.03	48.46%	48.60%
0.01	0.05	48.11%	48.04%
0.005	0.01	48.27%	48.73%
0.005	0.03	47.76%	47.89%
0.005	0.05	47.36%	47.26%
0.003	0.005	47.46%	47.25%
0.003	0.01	47.95%	47.86%
0.003	0.03	47.29%	47.37%
0.003	0.05	46.74%	46.48%
0.001	0.003	44.55%	44.98%
0.001	0.005	45.98%	46.09%
0.001	0.01	46.78%	46.71%
0.001	0.03	46.16%	46.40%
0.001	0.05	45.61%	45.59%

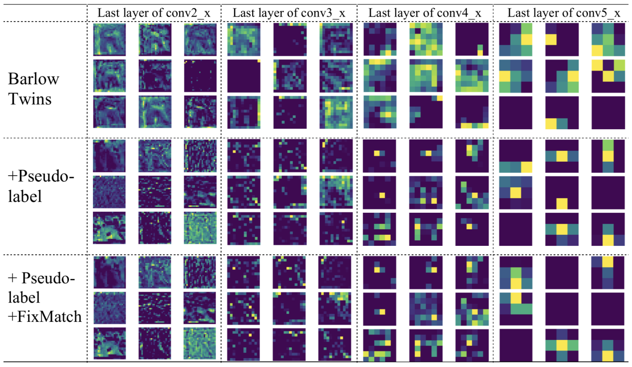

Figure 1. **Feature map visualization.** A randomly chosen subset of features from different models (*Row* 1-3) are selected from the last layer within each residual block $\{2, 3, 4, 5\}$ of ResNet-50. (*Col* 2-5) . Notice the feature grid becomes sparser as the model becomes more advanced compared to the Barlow Twins baseline.

4.5. Analysis of Successful and Failed classifications

We observe the top 625 images with highest probabilities by our model for all 800 classes and manually assign accuracy scores based on human impression. In Fig. 2, we demonstrate the classification results of two classes with the highest accuracy and two with the lowest. Classes of category that are rich in the dataset, such as all the classes with specific dog breeds, share high accuracy. On the other hand, classes whose key features occupy a small proportion of pixels in the samples often have lower accuracy. We also notice low accuracy caused by class imbalance in the dataset, which is evident by a drastic increase in error rate among images with lower probability ratings. For example in the sewing machine class, the model achieves high accuracy in the best 100 predictions but its performance quickly decreases after the top 400 marks.

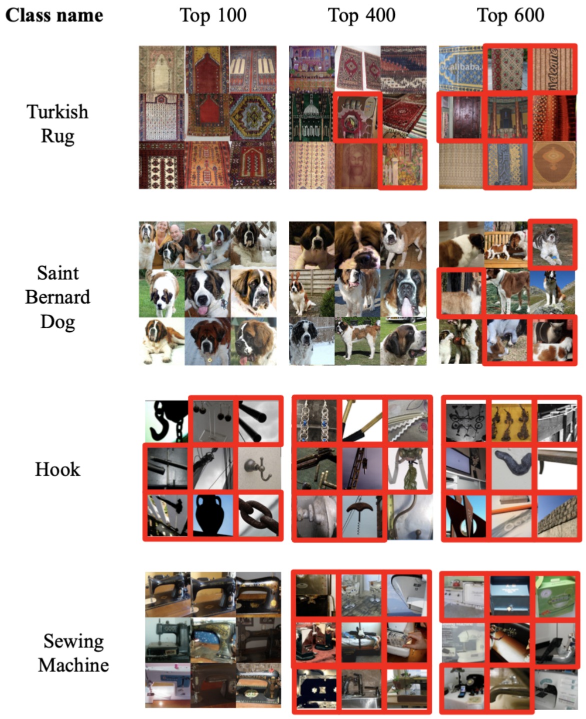

Figure 2. Examples of images collected by our procedure to demonstrate correct and incorrect labeling

5. Conclusion

In this report, we present a semi-supervised learning method that won the final competition of the Deep Learning Course SP21 edition at NYU. Our approach utilizes unsupervised learning, pseudo-label iterations, semi-supervised learning, and active learning jointly, achieving a 55.80% accuracy when trained with 0.5% labeled data and 57.19% with additional 0.25% labels. This promising result validates the effectiveness of the SOTA methods from different machine learning research areas and shows the importance to combine diverse methods in order to solve practical problems. For future work, the method we propose can be simplified and combined into a unified semi-supervised architecture by introducing novel loss functions and self-training iterations.

Acknowledgements

Thanks to NYU HPC for providing 400 hours of GPU time.

References

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *CoRR*, abs/2006.09882, 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020b.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. *arXiv preprint arXiv:1806.01768*, 2018.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., and Mahajan, D. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.