Prototypical Regularization in Deep Neural Networks

Cole DeLude[†],Cusuh Ham*, James Smith*
*CS, [†]ECE

May 2019

1 Summary

As deep learning continues to match or exceed human-like performance in simple tasks such as supervised object classification, many data scientists elect to explore new and challenging problems where few or limited labeled data is available. The field of semi-supervised learning (SSL) [7, 5, 2, 6] attempts to remedy this by utilizing both labeled and unlabeled data in the training process. State-of-the-art SSL methods such as virtual adversarial training (VAT) [5] and Π-Model [4] generally rely on adding a regularization term in the label space. Our project sought to investigate the addition of regularization at layers prior to the label space, in particular the penultimate layer. Our regularization, *cluster loss* (CL), is designed to penalize each unlabeled example by the distance between its feature representation and the feature representation of its nearest labeled example. This in turn encourages the unlabeled data to cluster around the labeled data in the latent feature space. Experiments were conducted to determine the effect of cluster loss when added to models trained using no regularization, VAT, and VAT + Entropy Minimization (VAT+EM). Results show that in the case of 400 out of 5000 data points are labeled, cluster loss added to VAT boosted accuracy by 9.33% and 10.64% over VAT and VAT+EM, respectively. Smaller improvements in performance were demonstrated for the case of 4000 labeled data points. A more qualitative analysis of the labeled and unlabeled feature representations confirmed that CL achieves tighter clusters in the feature space than other methods examined.

2 Details

2.1 Introduction

Deep learning has long been celebrated for its ability to perform tasks such as supervised object classification at a level comparable or exceeding previous state-of-the-art methods. Within the framework of supervised learning, the performance of such systems are limited by the availability of suitably labeled data. In some applications this necessity for labeled data may not pose too much of a hindrance but for a large class of interesting learning problems, the task of acquiring these labels may be intractable. We then turn our attention to the field of semi-supervised learning [7, 5, 2, 6] which trains models using both labeled and unlabeled data. Current state-of-the-art-methods such as VAT [5] reconcile the relationship between the labeled and unlabeled examples through the addition of regularization in the output layer (i.e. label space). Our project sought to investigate the addition of a regularization term that is designed to encourage clustering within the penultimate layer of the network. The proposed regularization, *cluster loss*, penalizes the distance between each unlabeled feature representation and its nearest labeled feature representation. In a sense, we seek to constrain our model to learn a feature space in which the labeled examples form prototypical feature vectors and unlabeled examples are clustered around these vectors.

2.2 Method

2.2.1 Motivation

Cluster loss hinges on the assumption that the feature representations of the labeled data has some structure. Intuitively, we would like the unlabeled feature representations to not deviate too far from this learned structure since it is our "best guess" of the optimal classifier. At a high level, we are taking the unlabeled examples to be perturbed inputs and seek to minimize the degree to which these perturbations effect our model.

2.2.2 Formulation

Consider a dataset $\mathcal{D}$ containing M labeled data points denoted by $\mathcal{D}_l : \{x^l_m, y_m\}_{m=1}^M$ and N unlabeled data points denoted by $\mathcal{D}_{ul} : \{x^{ul}_n\}_{n=1}^N$. Using this data we seek to train a convolutional neural network (CNN) consisting of L layers for a classification task. Let $x^l_{i,m}$ for $m = 1, \dots, M$ denote the labeled feature representation and $x^{ul}_{i,n}$ for $n = 1, \dots, N$ denote the unlabeled feature representation at layer i. Since cluster loss penalizes the distance between each $x^{ul}_{i,n}$ and its nearest $x^l_{i,m}$ we formulate the regularization term to be the following:

$$\mathcal{L}_{CL}(i) = \sum_{n=1}^{N} \left\|x^{ul}_{i,n} - x^{l}_{i,j}\right\| \quad \text{where} \quad j = \underset{m}{\text{argmin}} \left\|x^{ul}_{i,n} - x^{l}_{i,m}\right\| \tag{1}$$

We then apply $\mathcal{L}_{CL}$, summarized in Algorithm 1, to the $i = N - 1$ (i.e. penultimate) layer of the network.

Algorithm 1 Cluster Loss Regularization

```
for each training epoch do
    calculate layer $i$ feature representations of labeled data: $x^l_{i,m}$
    for each minibatch update do
        calculate standard cross-entropy loss with labeled data: $\mathcal{L}_{CE}$
        calculate cluster loss by (1): $\mathcal{L}_{CL}$
        update network with regularized loss $\mathcal{L}_{CE} + \beta \cdot \mathcal{L}_{CL}$
```

After initial training with the labeled examples, we apply this regularization in two steps: first, labeled data points in $\mathcal{D}_l$ are used to generate prototype feature vectors within a designated layer of interest as demonstrated in Figure 1; second, the regularization loss within these layers is calculated. The prototype feature vectors are recalculated after each training epoch. The training with unlabeled data then seeks to minimize this additional loss term, which in turn encourages clustering within the feature space.

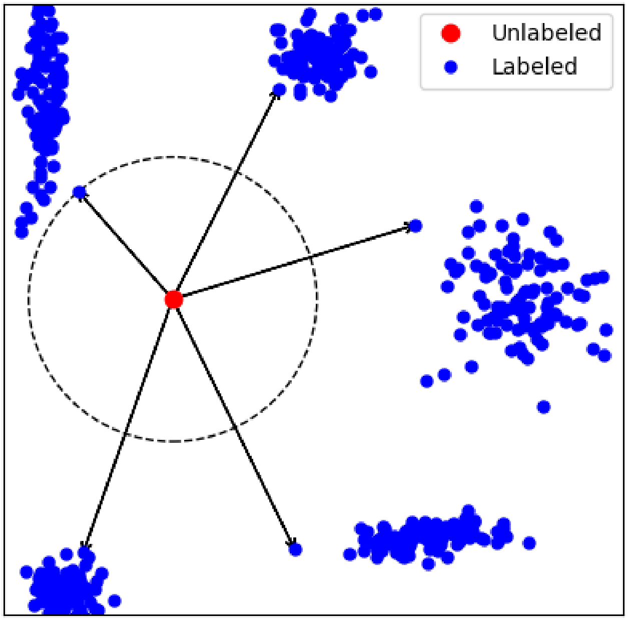


Figure 1: Visualization of cluster loss applied to a single unlabeled feature representation (red) given labeled feature representations (blue). During training, the cluster loss is applied to all unlabeled features.

2.3 Experimental Design

Empirical analysis of CL regularization was demonstrated with semi-supervised learning experiments on the CIFAR-10 dataset [3]. We compare CL to state-of-the-art VAT regularization by using the 9-layer CNN published in [5] (Figure 2). We specifically compare CL to a baseline model with no regularization, VAT, and VAT + Entropy Minimization (EM) [1].

We considered two scenarios of labeled data availability: 4000 labeled examples and 400 labeled examples. We reused the training schedule (ADAM optimization; 128 examples per batch; 123 training epochs; learning rate of $1e-3$ which linearly decays to 0 for the last 40 epochs) and VAT hyperparameters ($\epsilon = 10$) from [5]. For CL, we tuned two hyperparameters: a loss weighting term, β, and the number of centroids (which consist of labeled examples), K. Both of these hyperparameters were tuned using a log-scale parameter sweep and 1000 validation data points. We found the optimal values of $\beta = 5e-3$ and $K = 400$. The models were evaluated for classification accuracy using the full CIFAR-10 testing dataset and both a mean and standard deviation of across five trials per scenario was recorded.

2.4 Results and Analysis

The analysis of cluster loss was done in two parts. The first part centers on determining if the addition of cluster loss offered any performance benefits over existing methods of regularization. The second part concentrates on examining if the desired clustering within the latent feature space is achieved.

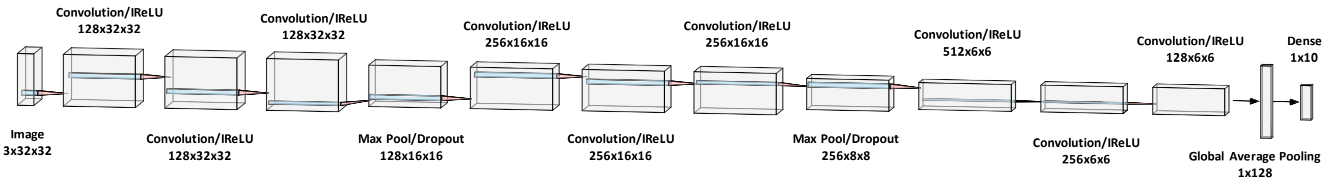


Figure 2: 9-Layer CNN Architecture used in experiments.

Table 1: Accuracy results for 400 and 4000 labeled examples.

400 labeled	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Accuracy
Baseline	0.4706	0.4636	0.4767	0.466	0.4703	0.4649 ± 0.0050
VAT	0.663	0.6608	0.6666	0.6636	0.6673	0.6642 ± 0.0026
VAT+EM	0.6556	0.6474	0.655	0.6518	0.6457	0.6511 ± 0.0044
CL	0.5132	0.5131	0.5105	0.5058	0.5053	0.5096 ± 0.0038
VAT+CL	0.7498	0.7720	0.7600	0.7514	0.7542	**0.7575** ± **0.0090**
VAT+EM+CL	0.6858	0.6707	0.6581	0.6578	0.6678	0.6680 ± 0.0115

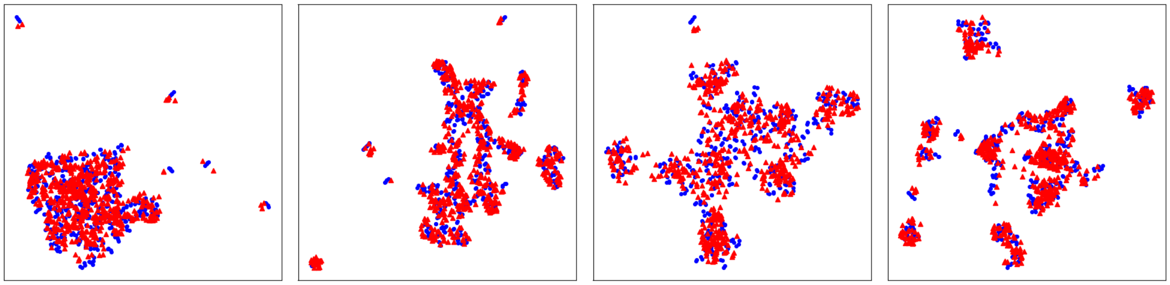
4000 labeled	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Accuracy
Baseline	0.7591	0.7611	0.7424	0.7546	0.7591	0.7553 ± 0.0076
VAT	0.8489	0.8523	0.8437	0.8462	0.8442	0.8468 ± 0.0038
VAT+EM	0.8594	0.8532	0.8448	0.8447	0.8448	0.8494 ± 0.0067
CL	0.7787	0.7574	0.769	0.7677	0.763	0.7672 ± 0.0079
VAT+CL	0.8523	0.8490	0.8533	0.8487	0.8487	0.8504 ± 0.0022
VAT+EM+CL	0.8668	0.8507	0.8522	0.8533	0.849	**0.8544** ± **0.0071**

2.4.1 Model Accuracy

Due to the stochastic nature of the training process, each trial yields slightly varying test accuracies. Therefore the test accuracy of each trial was calculated and the resulting means and standard deviations of performance were examined. A collection of these results and statistics are given in Table 1. These results are encouraging, since in both the 400 and 4000 labeled example scenarios the addition of cluster loss resulted in increased accuracy. In particular, for the 400 labeled example scenario the VAT+CL model gave an accuracy boost of 9.33% and 10.64% over VAT and VAT+EM, respectively. In the case of 4000 labeled examples, the boost using the VAT+EM+CL was 0.78% and 0.5% when compared to the same models.

2.4.2 Clustering in Feature Space

For the given CNN architecture feature representations at the penultimate layer are of dimension 128, which makes direct analysis inherently difficult. We settle for a more qualitative analysis



(a) (b) (c) (d)

Figure 3: Comparison of low dimensional representations of labeled (blue) and unlabeled (red) data for models trained using (a) baseline (b) CL (c) VAT + ENT (d) VAT + ENT + CL.

utilizing low-dimensional embeddings of the feature representations. We used the UMAP embedding technique as it was found to offer more interpretable results than general PCA. Further, it offers the ability to apply a parametric transform to new data which other popular methods, such as t-SNE, lack. The feature vectors formed by the labeled training data were used to train the UMAP model. The unlabeled examples were then transformed to the learned low-dimensional space. The intuition behind this is that if proper clustering has been achieved this low dimensional model formed using the labeled data should also well represent the unlabeled data. Therefore the unlabeled data should cluster around the labeled data points in this low dimensional space. The results of this embedding of labeled and unlabeled data for the models trained using 400 labeled examples is shown in Figure 3. Clearly, the addition of cluster loss to the baseline and VAT+ENT training models encourages the desired clustering of unlabeled feature representations around labeled representations. Therefore the overarching goal of constraining our model through addition of regularization in the feature space was achieved.

2.5 Conclusion

In this project, we introduced a novel regularization term, *cluster loss*, for semi-supervised learning algorithms. This loss term penalizes the distances between feature representations of unlabeled examples and their nearest labeled example feature representation. We demonstrated qualitative and quantitative improvements using our approach; qualitatively, CL encourages the clustering of unlabeled features around the learned prototypical features and quantitatively, CL outperformed VAT and VAT+ENT on CIFAR-10 classification. While the improvements were significant in the case of 400 labeled examples, the performance for 4000 examples is indicative that this method demands further investigation. Our method was simple, utilizing L_2 penalties as a metric. In the future, we aim to explore various distance metrics and show additional evaluations on the Street-View House Numbers dataset.

2.6 Contributions

James led the majority of the coding for the experiments, Cusuh led the experimental trials, and Cole led the analysis of the results.

References

[1] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 2004. MIT Press.

[2] Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models, 2014.

[3] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

[4] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning, 2016.

[5] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[6] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. 2018.

[7] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks, 2015.