Beyond the Final Linear Layer: Enhancing Decision Boundaries

Michael Majurski
Information Technology Lab, NIST
michael.majurski@nist.gov

Sumeet Menon
University of Maryland, Baltimore County

David Chapman
University of Maryland, Baltimore County

Abstract

SSL leverages an abundance of unlabeled data to improve deep learning based model performance under limited training data regimes. This paper presents a novel extension to any image classification architecture which improves accuracy in low-label regimes. We extend the FixMatch [1] training scheme with our novel last layers and demonstrate test accuracy improvement. The novelty consists of 2 elements: first we replace the last linear layer with a GMM trained via backprop, and second, we impose class-wise constraints on the embedding space the GMM operates on. These methods match published SOTA 250 label Cifar10 [2] results and come close to matching SOTA in the 40 label regime without the significant model complexity of methods like SimMatchV2 [3]. Our method achieves 94.8% and 94.2% accuracy with 250 and 40 Cifar10 labels respectively. Our code is available at: https://github.com/mmajurski/ssl-gmm

1. Introduction

SSL leverages an abundance of unlabeled data to improve deep learning based model performance under limited training data regimes [4, 5, 6]. Image classification has become a playground for exploring new SSL ideas. The early successes of deep learning based methods relied on large annotated datasets to enable models to learn the relevant features to perform the task, i.e. image classification build on top of ImageNet [7]. With data annotation becoming a significant bottleneck, especially in application domains outside of the standard benchmarks, another learning paradigm was needed.

There are several flavors of SSL. Contrastive learning methods leverage the intuition that similar instances should be close in the representation space, while different instances are farther apart [8, 9]. Consistency regularization borrows the intuition that modified views of the same instance should have similar representations and predictions [1, 10, 11, 12]. Pseudo-labeling methods like FixMatch [1] fall within the consistency regularization domain.

This work argues that pseudo-labeling methods can be improved with better calibration of the network logits used to filter the pseudo-labels into reliable and unreliable. Neural networks are known to be overconfident in their predictions [13], and this affects the pseudo-labeling process. Potentially allowing for the inclusion of more incorrect pseudo-labels any specific logit threshold would otherwise have. This work demonstrates the better calibrated replacements for a models final linear layer can improve the final accuracy of pseudo-labeling based SSL algorithms in very label scarce regimes.

With that goal in mind, this work proposes replacing the final linear layer of the neural network with one of a few options, all of which have explicit modeling of class cluster centroids. First, a nearest neighbor (k=1) kmeans layer. Second, an axis-aligned differentiable Gaussian Mixture Model (GMM) trained via back prop. Third, an identity covariance axis-aligned differentiable GMM trained via back prop. These three methods represent differing levels of prescription about how the final embedding space should be arranged.

Additionally, to assist these cluster based final embedding classification layers, we explore the effect of different constraints on the embedding space. We explore 1) no constraints, 2) an L2 penalty between observed cluster centers and the learned cluster centroids, 3) a mean/covariance constraint which combined the l2 penalty with an identity covariance constraint, and finally 4) a full method of moments enforcing an identity representation for the first 4 gaussian moments. These four constraints represent increasing levels of prescription about how the embedding space should be organized. The L2 constraint is adding a penalty term for any embedding which does not have an identity first gaussian moment. The mean/covariance constraint extents that penalty to enforce an identity first and second gaussian moment. The final constraint explores whether including all of the first 4 gaussian moments improves the final learned model test accuracy.

This paper contributes a simple easy to implement improvement to pseudo-labeling methods where very few annotations are available, replace the last linear layer with a kmeans layer which explicitly models class cluster centers. We demonstrate this methodology using Cifar10 and Cifar100 [2] with 40 and 400 labels respectively. Additionally, we explore and demonstrate that high level prescriptive constraints on the embedding space produce significantly worse outcomes than allowing the embedding space to take on whatever emergent structure the training process produces. Finally, because the L2 cluster centroid penalty is applied to all unlabeled data and not just the valid pseudo-labels, our method extracts some training signal from every unlabeled data point, unlike FixMatch [1] and other methods which only learn from the valid pseudo-labels.

2. Related Work

2.1. Pseudo-Labeling

Self-supervised learning was among the initial approaches employed in the context of semi-supervised learning to annotate unlabeled images. This technique involves the initial training of a classifier with a limited set of labeled samples and incorporates pseudo-labels into the gradient descent process, exceeding a predefined threshold [14, 15, 16, 17, 18, 19]. A closely related method to self-training is co-training, where a given dataset is represented as two distinct feature sets. These independent sample sets are subsequently trained separately using two distinct models, and the sample predictions surpassing predetermined thresholds are utilized in the final model training process [20, 21]. A notably advanced approach to pseudo-labeling is the Mean Teacher algorithm [22], which leverages exponential moving averages of model parameters to acquire a notably more stable target prediction. This refinement has a substantial impact on enhancing the convergence of the algorithm.

2.2. Consistency Regularization

Consistency regularization operates on the premise that when augmenting an unlabeled sample, its label should remain consistent. This approach implicitly enforces a smoothness assumption, promoting coherence between unlabeled samples and their basic augmentations. In other words, the model should be able to predict the unlabeled sample x exactly the same way it predicts the class for Augmented(x) [1, 23, 24, 25]. In addition to evaluating image-wise augmentations, recent research has demonstrated that incorporating class-wise and instance-based consistencies yields superior performance outcomes [9, 26].

2.3. Contrastive Learning

3. Methodology

3.1. KMeans

3.2. Axis Aligned Differentiable Gaussian Mixture Model

3.3. Embedding Constraints

4. Experiments

4.1. Ablation Study

5. Conclusions

References

[1] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 1, 2

[2] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 1, 2

[3] Mingkai Zheng, Shan You, Lang Huang, Chen Luo, Fei Wang, Chen Qian, and Chang Xu. Simmatchv2: Semi-supervised learning with graph consistency. *arXiv preprint arXiv:2308.06692*, 2023. 1

[4] Xiaojin Zhu and Andrew B Goldberg. *Introduction to semi-supervised learning*. Springer Nature, 2022. 1

[5] Yu-Feng Li and De-Ming Liang. Safe semi-supervised learning: a brief introduction. *Frontiers of Computer Science*, 13:669–676, 2019. 1

[6] Mohamed Farouk Abdel Hady and Friedhelm Schwenker. Semi-supervised learning. *Handbook on Neural Information Processing*, pages 215–239, 2013. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[8] Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14421–14430, 2022. 1

[9] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9475–9484, 2021. 1, 2

[10] Doyup Lee, Sungwoong Kim, Ildoo Kim, Yeongjae Cheon, Minsu Cho, and Wook-Shin Han. Contrastive regularization for semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3911–3920, 2022. 1

[11] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 1

[12] Jiwon Kim, Youngjo Min, Daehwan Kim, Gyuseong Lee, Junyoung Seo, Kwangrok Ryoo, and Seungryong Kim. Conmatch: Semi-supervised learning with confidence-guided consistency regularization. In *European Conference on Computer Vision*, pages 674–690. Springer, 2022. 1

[13] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, pages 23631–23644. PMLR, 2022. 1

[14] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995. 2

[15] David McClosky, Eugene Charniak, and Mark Johnson. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 337–344, 2006. 2

[16] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. 2

[17] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019. 2

[18] Ioannis E Livieris, Konstantina Drakopoulou, Vassilis T Tampakas, Tassos A Mikropoulos, and Panagiotis Pintelas. Predicting secondary school students' performance utilizing a semi-supervised learning approach. *Journal of educational computing research*, 57(2):448–470, 2019. 2

[19] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, volume 1, pages 29–36, 2005. 2

[20] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998. 2

[21] V Jothi Prakash and Dr LM Nithya. A survey on semi-supervised learning techniques. *arXiv preprint arXiv:1402.4645*, 2014. 2

[22] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2

[23] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[24] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 2

[25] Aamir Mustafa and Rafał K Mantiuk. Transformation consistency regularization–a semi-supervised paradigm for image-to-image translation. In *European Conference on Computer Vision*, pages 599–615. Springer, 2020. 2

[26] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching, 2022. 2