This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Iterative Self-Learning: Semi-Supervised Improvement to Dataset Volumes and Model Accuracy

Robert Dupre, Jiri Fajtl, Vasileios Argyriou, Paolo Remagnino
Kingston University, Penrhyn Rd
Vasileios.Argyriou@kingston.ac.uk

Abstract

A novel semi-supervised learning technique is introduced based on a simple iterative learning cycle together with learned thresholding techniques and an ensemble decision support system. State-of-the-art model performance and increased training data volume are demonstrated, through the use of unlabelled data when training deeply learned classification models. Evaluation of the proposed approach is performed on commonly used datasets when evaluating semi-supervised learning techniques as well as a number of more challenging image classification datasets (CIFAR-100 and a 200 class subset of ImageNet).

1. Introduction

Semi-supervised learning has become one of the most prevalent topics within image processing and computer vision research in recent years. With the ever increasing availability of high powered GPU hardware and the success of deep learning on applications such as computer vision [10, 13], speech recognition [8], and natural language processing [19, 7] the need for large scale datasets to support these methods becomes a higher priority, as well as a bottleneck to performance improvement. Semi-supervised processes have been applied successfully in many areas such as image classification and segmentation [11], natural language processing and artificial intelligence [3]. Typically, semi-supervised deep learning uses novel model architectures, regularization methods or loss functions combining outputs from known labels with unknown labels to provide more accurate outputs, [9, 4]. Laine [14] utilises an architecture based on ensemble predictions, acquired during the training of a network at different epochs or under different regularization and input conditions. Tarvainen et al. [20] take the concept of temporal ensembling and extend it to the model weights. French et al. [6] extend by introducing class balancing and confidence thresholding. Miyato et al. [15] consider a novel regularization method to semi-supervised learning.

The following method is an extension of [5]. The method iteratively reclassifies a dataset such that the model being trained is only ever exposed to what it considers fully labelled data. Firstly, a simple and easily implemented semi-supervised learning framework, independent from model architecture or loss functions making it applicable to a wide range of classification tasks. Secondly, novel learned thresholding techniques and metrics to supervise the dataset growth, ensuring only confidently classed samples are added to a training dataset.

2. Methodology

The core assumption in this work is that generalization error always decreases with more training samples as shown by [1] and recently [2]. To address this problem the Iterative Learning-Ensemble (IL-E) approach is presented. The iterative nature of the IL-E is given by the train, classify, analyse and finally update cycle. Firstly, a model (θ) is trained on a cleanly labelled dataset, $\mathbb{D}_l$, and validated on the $\mathbb{D}_v$ dataset. The training of the model is performed in a relevant way to the application and task, neither the architecture nor the loss functions are changed in any way. Secondly, the unlabelled samples are classified and the process of updating the training set is run.

Let $x \in \mathbf{R}^d$ represent an input variable in d dimensions and $y \in \mathbf{L}^C$ represent the label associated with that sample, where C represents the number of possible class labels. In this work, x_i represents an image and y_i^c the label from C-classes. From the pool of cleanly labeled and unlabelled data, three datasets are constructed: Labelled ($\mathbb{D}_l = x_n^l, y_n^l | n = 1, \ldots, N^l$), derived from a portion of the cleanly labelled data. Unlabelled ($\mathbb{D}_u = x_m^u, y_m^u | m = 1, \ldots, M^u$), indexed from only unlabelled data and validation ($\mathbb{D}_v = x_o^v, y_o^v | o = 1, \ldots, O^v$), derived from the remaining subset of the cleanly labelled data.

The primary issue when adding newly labelled samples to the training dataset is ensuring the model is confident that the additions are labelled correctly. This confidence is achieved in two ways: firstly, well established ensembling techniques are utilised to produce better predictions

from a trained model [14, 17, 20] and secondly, a novel set of confidence metrics have been devised based solely on the posterior probabilities produced from model θ. Importantly, there are no additional clustering or preprocessing steps applied to the unlabelled data of any kind, the only assumption made within this work is that the data is of a similar quality, context and application as that within the cleanly labelled.

The goal of ensembling is to find the most positive class distribution for use with the confidence metrics. To this end, a number of augmentations are applied to an unlabelled sample such that $\mathbb{D}_{aug} = \mathbf{x}_a^{(j)}, y^{(j)} | a = 1, \ldots, A^j$ represents a single sample $\mathbf{x}^{(j)}$, augmented in A different ways, each with the same label $y^{(j)}$. The augmented samples, including the original, are now passed to the model for inference and the posterior probability vectors, or class distributions, for all the augmented samples $\tilde{z}$ returned.

$$\tilde{z} = P(y|\mathbb{D}_{aug}; \theta) \tag{1}$$

The returned posterior probability vectors are then scaled by the similarity of the class distributions returned as a result of Eq. 1. The standard deviation of the posterior probabilities between class labels across the augmented samples is then subtracted from $\tilde{z}$ as form of scaling. Augmented ensembles which when evaluated differ greatly in their class distributions, result in a larger standard deviation. In turn, this would penalise the final confidence score $\mathbf{x}_a^{(j)}$ more so than when the model produces similar distributions across the ensemble. Finally, the augmented sample a with the highest posterior probability for any class label, is selected and its original unscaled class distribution used as input $\mathbf{x}^{(j)}$ for the confidence metrics highlighted in Eq. 4, 5 and 6.

$$a =_a (\tilde{z}_a - \sigma) \tag{2}$$

The confidence metrics used to further ensure unlabelled samples are correctly labelled, cover three distinct areas computed from the posterior probabilities after evaluation of the unlabeled data $\mathbb{D}_u$, (see supplementary material for visual representations of these concepts). Firstly, the single highest class activation obtained from the posterior distribution c_a (higher is better). Formally, consider an unlabeled sample $\mathbf{x^{(j)}} \in \mathbb{D}_u$

$$y_1, y_2 =_y P(\mathbf{y}|\mathbf{x^{(j)}}; \theta) \tag{3}$$

Where y_1 and y_2 are labels corresponding to the first and second highest posterior probabilities. The c_a is then

$$c_a = P(y_1|\mathbf{x^{(j)}}; \theta) \tag{4}$$

Second, the difference between the highest and second highest activation c_b (larger difference is better) is computed according to Eq. 5.

$$c_b = P(y_1|\mathbf{x^{(j)}}; \theta) - P(y_2|\mathbf{x^{(j)}}; \theta) \tag{5}$$

Lastly, c_c is calculated as the Euclidean distance between the posterior distribution for the unlabeled sample $\mathbf{x}^{(j)}$ and the average distribution $p_t(y_1)$ for the predicted class y_1 (lower score is better). $p_t(y_1)$ is computed over all training samples of class y_1. These average distributions per class are computed at the end of each model training iteration and are recorded for use in these confidence computations.

$$c_c = \|P(\mathbf{y}|\mathbf{x}^{(j)}; \theta) - p_t(y_1)\| \tag{6}$$

For each of these three metrics a value is returned, in the cases of c_a and c_b the value returned by the model should be high and for c_c the distance between the two posterior probability distributions should be low, however the c_c scores are inverted so as to have a uniform, higher is better policy. The weighted sum of these metrics scores is then used to provide a final confidence score for a specific unlabeled sample $\mathbf{x}^{(j)}$. As some metrics are more informative that others their contribution to the final confidence c should reflects this. The weighting is found experimentally but is rooted on the accuracy of the metric on a set of unlabelled samples. Importantly these values may change based on application as certain metrics may be more informative in different problems.

$$c = c_a w_a + c_b w_b + \frac{1}{c_c} w_c \tag{7}$$

Using a defined threshold T_c, samples can now be approved for inclusion in the labeled dataset $\mathbb{D}_l$, updated for use in the next training iteration. The threshold T_c could be defined manually, allowing for policies where only very confidently analyzed samples are added or, through the use of a lower threshold, a more "quantity over quality" policy can be adopted. In this work the threshold value T_c is learned. A process is run to find a threshold T_c, which when applied would add samples to $\mathbb{D}_l$ with a defined accuracy T_a, i.e defining a threshold T_c whereby 99% of samples added to $\mathbb{D}_l$ are correctly labelled. The process is run using only cleanly labelled data. The function $acc()$ calculates the percentage of correctly labeled samples q in a dataset $\mathbf{X_u}$ classified by model θ against their ground truth labels $\mathbf{y}$, given the threshold T_c.

$$q = acc(\mathbf{X_u}, \mathbf{y}, T_c, \theta) \tag{8}$$

Therefore given the required addition accuracy T_a the max T_c can be calculated,

$$T_c = \max t_c \quad \text{subject to} \quad acc(\mathbf{X}, \mathbf{y}, t_c, \theta) > T_a \tag{9}$$

Accuracy T_a was set to $> 99\%$. This process is run once on training data, as the model will be most confident on samples it has already seen and, as a result of this, impose a higher threshold than one defined using the validation set.

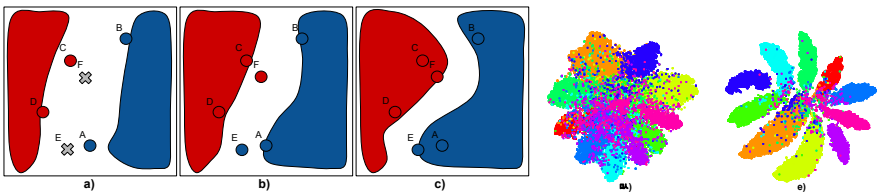

Figure 1. (Left) A-F $\in \mathbb{D}_u$, red and blue areas are two class manifolds $\in \theta$, a) Iter 1: A,B,C and D are classified as red or blue using proximity to the respective manifolds, while E and F remain unclassified (confidence is not high enough). b) Iter 2: after retraining the model with new samples A-D, confident becomes sufficient to classify E and F and add them to the new training set. c) Iter 3: the manifolds are updated with the samples E and F. (Right) t-SNE of the last fully connected layer (1024 neurons) of JFNet2 when evaluating the SVHN validation dataset. d) Clusters with the model trained on the initial 1000 samples. e) Clusters after IL-E has been run for 75 iterations, increasing the training dataset size and improving classification accuracy.

Table 1. IL-E results across the three datasets (SVHN, CIFAR-100, TinyImageNet), with full and subset benchmark training results.

	Error Rate% (σ)	Error Rate % (Improvement)	Error Rate% (σ)	Added Samples (Acc. %)
		SVHN		
Model	1k Benchmark	1k Samples	Full Benchmark	
GAN [18]	N/A	8.11%	N/A	-
$\prod$ model [14]	N/A	4.82%	2.54% ($\pm$0.04)	-
Temporal E. [14]	N/A	4.42%	2.74% ($\pm$0.06)	-
VAT+EntMin [15]	N/A	3.86%	N/A	-
ResNet-18 (IL-E)	19.74% ($\pm$0.32)	**4.29% (-15.45)**	2.98% ($\pm$0.04)	71,068 (94.89%)
LeNet-5 (IL-E)	25.24% ($\pm$1.55)	11.11% (-14.13)	7.16% ($\pm$0.09)	42,999 (96.86%)
JFNet (IL-E)	20.18% ($\pm$0.50)	5.64% (-14.54)	3.84% ($\pm$0.05)	66,421 (96.13%)
		CIFAR-100		
	5k Benchmark	5k Samples	Full Benchmark	
Temporal E. [14]	N/A	38.65% (10k Samples)	N/A	-
ResNet-18 (IL-E)	32.49% ($\pm$0.45)	**28.09% (-4.4)**	17.53% ($\pm$0.09)	42,526 (75.1%)
LeNet-5 (IL-E)	89.21% ($\pm$0.22)	87.47% (-1.74)	65.55% ($\pm$0.38)	375 (72.53%)
JFNet (IL-E)	39.66% ($\pm$0.22)	66.49% (-1.36)	39.66% ($\pm$0.22)	4,786 (73.21%)
		Tiny ImageNet		
	10k Benchmark	10k Samples	Full Benchmark	
ResNet-18 (IL-E)	37.47% ($\pm$0.46)	**33.68% (-3.79)**	27.38% ($\pm$0.15)	56,619 (81.37%)
LeNet-5 (IL-E)	95.48% ($\pm$0.43)	94.43% (-1.05)	81.58% ($\pm$0.27)	69 (43.49%)
JFNet (IL-E)	83.40% ($\pm$0.12)	81.61% (-1.79)	60.98% ($\pm$0.25)	684 (83.19%)

During these incremental updates the model is trained using an ever growing dataset. The dataset volume increases by the addition of unlabelled samples which the model has confidently identified belong to a respective class (i.e $> T_c$). As a result the model develops its *knowledge* of specific classes and is therefore better able to identify additional samples in latter iterations. This process is symbolically shown in Figure 1 a-c, whereby a subset of new, unlabeled, samples get projected closer to the existing manifolds due to already learned characteristics of respective classes. Figure 1 d-e shows a real world example of the effect IL-E has had on the JFNet2 model's class manifolds. Additionally as the model is re-initialized at the beginning of each iteration, this method can leverage randomly initialised weights to help with the classification of unlabelled samples.

3. Results and Conclusions

SVHN [16] is used for benchmarking and to better validate the performance of this iterative approach on a more challenging task, CIFAR-100 [12], and a 200-class subset of ImageNet known as Tiny ImageNet are used. Initially benchmarks are run for each of the three models on the three datasets. Table 1 (columns 1 & 3) outlines the benchmark error rates for each of these model architectures on both a subset of the training data and the full. The training subset size is based on 50 samples per class, CIFAR-100 uses 5,000 samples and Tiny ImageNet uses 10,000 samples. As the SVHN dataset is one of the most commonly used datasets when comparing semi-supervised learning techniques, the standard 1,000 samples is used (100 samples from each of the 10 classes). Each training subset is made up of an even distribution of classes with images from each class chosen at random. Each experiment was conducted four times with the average results presented along with the standard deviation given in brackets. The inclusion of these benchmarks is vital, especially for any result that uses a customised loss function or architecture, as without it is difficult to ascertain if improvement gains can be attributed to the model architecture used or the semi-supervised method. As demonstrated the simple iterative approach to semi-supervised learning IL-E has a number of benefits. Most notable being state of the art error rates on the CIFAR-100 dataset and near state of the art on SVHN dataset, achieved with no changes to the training methods, loss functions or model architectures used. The IL-E demonstrates, through the application of novel confidence metrics, the ability for a model to leverage its own confidence scores to improve classification accuracy.

4. ACKNOWLEDGMENT

This work is co-funded by the NATO within the WITNESS project under grant agreement number G5437. The Titan X Pascal used for this research was donated by NVIDIA.

References

[1] A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

[2] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6241–6250, 2017.

[3] A. Carlson, J. Betteridge, and B. Kisiel. Toward an Architecture for Never-Ending Language Learning. *Conference on Artificial Intelligence (AAAI)*, pages 1306–1313, 2010.

[4] S. Cicek, A. Fawzi, and S. Soatto. SaaS: Speed as a Supervisor for Semi-supervised Learning. *arXiv preprint arXiv:1805.00980*, 2018.

[5] R. Dupre, J. Fajtl, V. Argyriou, and P. Remagnino. Improving Dataset Volumes and Model Accuracy with Semi-Supervised Iterative Self-Learning. *IEEE TIP*, 2019.

[6] G. French, M. Mackiewicz, and M. Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018.

[7] Y. Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

[8] A. Graves and N. Jaitly. Towards End-To-End Speech Recognition with Recurrent Neural Networks. *JMLR Workshop Conference Proceedings*, 32(1):1764–1772, 2014.

[9] P. Haeusser, A. Mordvintsev, and D. Cremers. Learning by association-a versatile semi-supervised training method for neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] S. Hong, H. Noh, and B. Han. Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation. *Advances in Neural Information Processing Systems*, 28:1495—-1503, 2015.

[12] A. Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. PhD thesis, 2009.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances In Neural Information Processing Systems*, pages 1097–1105, 2012.

[14] S. Laine and T. Aila. Temporal Ensembling for Semi-Supervised Learning. In *International Conference on Learning Representation*, pages 1–13, 2017.

[15] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii. Virtual Adversarial Training: a Regularization Method for Supervised and Semi-supervised Learning. pages 1–14, 2017.

[16] Y. Netzer and T. Wang. Reading digits in natural images with unsupervised feature learning. In *Neural Information Processing Systems (NIPS)*, page 5, 2011.

[17] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko. Semi-Supervised Learning with Ladder Networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554, 2015.

[18] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. (Nips):1–9, 2016.

[19] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[20] A. Tarvainen and H. Valpola. Mean teachers are better role models : Weight-averaged consistency targets improve semi-supervised deep learning results. (Nips), 2017.