Negative sampling in semi-supervised learning (2020)

John Chen, Vatsal Shah, Anastasios Kyrillidis

Summary

Contributions

The authors propose to integrate negative sampling/contrastive estimation into existing semi-supervised learning (SSL) frameworks. Adding Negative Sampling for SSL (NS3L) loss term does not introduce any significant computational overhead, and is easy and fast to tune to existing SSL algorithms. With NS3L, we favor good models that discriminate a supervised example from its neighborhood, comprised of unsupervised examples, where the model must only implicitly explain why the observed, supervised example is better than its unsupervised neighbors.

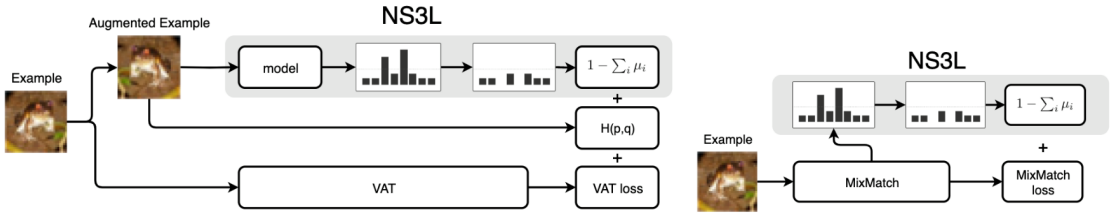

Figure 1: Left: Diagram of NS^{3}L with VAT. For NS^{3}L, an augmented example is fed into the model, which outputs a probability for each class. A threshold T is used to determine classes with sufficiently low probability, and these classes are fed into the NS^{3}L loss. The NS^{3}L loss is combined with the existing VAT loss and Cross Entropy loss. Right: Similar diagram of NS^{3}L with MixMatch; the NS^{3}L loss is combined with the existing MixMatch loss.

Method

To derive the NS3L loss, we start from the standard cross entropy loss:

$$\mathcal{L}\left(\{x_i, y_i\}_{i=1}^n\right) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\log\mu_{ik}$$

where there are n labeled samples, K classes, y_{ik} equals 1 when k is the ground-truth class, and μ_{ik} is the output of the classifier for sample i for class k. By considering two sets, a labeled and unlabeled sets, with $n+n_u$ examples, the CE loss can be written as the sum of the above supervised term, and the unsupervised term where we use the output probabilities of the classes other than the ground-truth (note the two are equivalent since $\mu_{\text{True label}} = 1 - \sum_{j\neq\text{True label}}\mu_j$):

$$\mathcal{L}\left(\{x_i, y_i\}_{i=1}^{n+n_u}\right) = -\frac{1}{n}\sum_{i_1=1}^{n}\sum_{k=1}^{K} y_{ik}\log\mu_{ik} + \underbrace{\frac{1}{n_u}\sum_{i_2=1}^{n_u}\log\left(1-\sum_{j\neq\text{ True label}}\mu_{i_2 j}\right)}_{:=\text{unsupervised part}}$$

In the loss above, we still need to know the ground-truth label to avoid summing over it. However, the authors suggest that using some simple heuristics, like randomly choosing a given class to be the

true class, or using a threshold works well. The table bellow shows that the results obtained with such heuristics is close the optimal performance:

F-MNIST	2,000
Supervised	17.25 ± .22
Uniform - 1	18.64 ± .38
Uniform - 3	19.35 ± .33
Exclude class of NN - 1	17.12 ± .15
Exclude 4 nearest classes with NN - 1	17.13 ± .21
Furthest class with NN - 1	16.76 ± .15
Threshold $T = 0.03$	16.47 ± .18
Threshold $T = 0.05$	16.59 ± .19
Oracle - 1	16.37 ± .12
Oracle - 3	15.20 ± .66

Let P be the classes that we choose to consider as negatives, i.e., the sample does not belong to, and by optimizing over batches rather than the full dataset, we can write the N3SL loss as follows:

$$\hat{\mathcal{L}}_{B_1,B_2}\left(\{x_i,y_i\}_{i=1}^{n+n_u}\right) = -\frac{1}{|B_1|}\sum_{i_1\in B_1}\sum_{k=1}^{K} y_{ik}\log\mu_{ik} - \underbrace{\frac{1}{|B_2|}\sum_{i_2\in B_2}\log\left(1-\sum_{j=1}^{P_{i_2}}\mu_{i_2 j}\right)}_{:=\text{NS}^3\text{L loss}}$$

The loss can then be added into any existing SSL loss: for example, if we take VAT, the total loss will be:

$$\mathcal{L} = \mathcal{L}_{\text{supervised}} + \lambda_2\mathcal{L}_{\text{VAT}} + \lambda_1\mathcal{L}_{\text{NS3L}}$$

for some weighting $\lambda_i, i \in \{1,2\}$.

Results

Table 1: Test errors achieved by various SSL approaches on the standard benchmarks of CIFAR10, with all but 4,000 labels removed, SVHN, with all but 1,000 labels removed, and CIFAR100, with all but 10,000 labels removed. "Supervised" refers to using only 4,000, 1,000, and 10,000 labeled samples from CIFAR10, SVHN, and CIFAR100 respectively without any unlabeled data. VAT refers to Virtual Adversarial Training.

Dataset	Supervised	PL	NS^{3}L	VAT	VAT + EntMin	Π model	Π + NS^{3}L	VAT + NS^{3}L
CIFAR10	20.76 ± .28	17.56 ± .29	16.03 ± .05	14.72 ± .23	14.34 ± .18	17.12 ± .19	16.06 ± .21	**13.94** ± .10
SVHN	12.39 ± .53	7.70 ± .22	6.52 ± .22	6.20 ± .11	6.10 ± .02	8.48 ± .15	7.98 ± .18	**5.51** ± .14
CIFAR100	48.26 ± .25	46.91 ± .31	46.34 ± .37	44.38 ± .56	43.92 ± .44	47.87 ± .34	46.98 ± .41	**43.70** ± .19

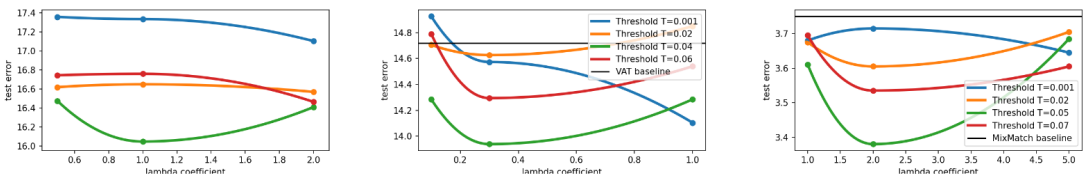

Figure 3: Parameter sensitivity study. Left: Test errors achieved by NS^{3}L on the standard benchmark of CIFAR10, with all but 4,000 labels removed. Middle: Test errors achieved by VAT + NS^{3}L on the standard benchmark of CIFAR10, with all but 4,000 labels removed. Right row: Test errors achieved by Mixmatch + NS^{3}L on the standard benchmark of SVHN, with all but 250 labels removed.

Table 2: Test errors achieved by various SSL approaches on top of VAT on the standard benchmarks of CIFAR10, with all but 4,000 labels removed, and CIFAR100, with all but 10,000 labels removed. VAT, EntMin and PL refer to Virtual Adversarial Training, Entropy Minimization, and Pseudo-Labeling respectively.

Dataset	VAT	VAT + EntMin	VAT + PL	VAT + NS^{3}L
CIFAR10	14.72 $\pm$ .23	14.34 $\pm$ .18	14.15 $\pm$ .14	**13.94** $\pm$ .10
CIFAR100	44.38 $\pm$ .56	43.92 $\pm$ .44	43.93 $\pm$ .33	**43.70** $\pm$ .19

Table 3: Test errors achieved by MixMatch and MixMatch + NS^{3}L on the standard benchmark of CIFAR10, with all but 250 labels removed and all but 4,000 labels removed.

CIFAR10	250	4,000
MixMatch	14.49 $\pm$ 1.60	7.05 $\pm$ 0.10
Mixmatch + NS^{3}L	**12.48** $\pm$ 1.21	**6.92** $\pm$ 0.12

Table 5: Test errors achieved by MixMatch and MixMatch + NS^{3}L on the standard benchmark of STL10, with all but 1,000 labels removed.

STL10	1,000
MixMatch	22.20 $\pm$ 0.89
Mixmatch + NS^{3}L	**21.74** $\pm$ 0.33