Semi-Supervised Learning by Augmented Distribution Alignment
(2019)

Qin Wang, Wen Li, Luc Van Gool

Summary

Contributions

Given the limited size of the labeled training examples, the empirical distribution of the labeled data often deviates from the true samples distribution. To correct for this distribution mismatch, at both the latent representations level and the small sampling size, the authors propose to add an adversarial loss to push the model to produce similar representations for both the labeled and unbalanced data, in addition to using Mixup to create new examples using the labeled and unlabeled examples to overcome the small sampling size.

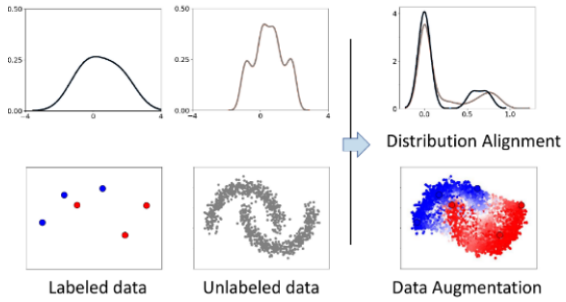

Figure 1. Illustration of the empirical distribution mismatch between labeled and unlabeled samples with the two-moon data. The labeled and unlabeled samples are shown in the **bottom left** and **bottom middle** figures, and the kernel density estimations of their x-axis projection are plotted in the **top left** and **top middle** figures, respectively. Our approach aims to address the empirical distribution mismatch by aligning sample distributions in the latent space (**top right**) and augmenting training samples with interpolation between labeled and unlabeled data (**bottom right**).

Method

To further show the distribution mismatch between the labeled and unlabeled examples, the authors compare the Maximum Mean Discrepancy (MMD), that was show to be heavily dependent on the sampling size, between the labeled and unlabeled sets, that are drawn from the same distribution. We see that the MMD vanishes only for a large number of labeled examples. The small sampling size thus causes a large MMD, for both its mean the its variance.

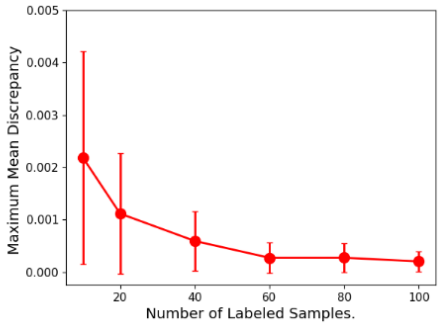

Figure 2. MMD between labeled and unlabeled samples in two-moon example with varying number of labeled samples. Number of unlabeled sample is fixed as $1,000$.

So a model trained on the labeled example is unlikely to generalize well to the test data. To overcome this, the authors propose to learn a model that minimizes both the classification loss and the distribution divergence between the unlabeled and labeled sets ($\mathcal{D}_u$ and $\mathcal{D}_l$).

$$\min_f \sum_{i=1}^{n} \ell\left(f\left(\mathbf{x}_i^l\right), y_i\right) + \gamma\Omega\left(\mathcal{D}_l, \mathcal{D}_u\right)$$

Distribution divergence To measure the distribution divergence, the authors use the discriminator that takes as input the feature extractor's g representation h of a given examples; and must be able to distinguish between the labeled and unlabeled examples, by predicting 1 for examples from $\mathcal{D}_l$ and 0 for examples from $\mathcal{D}_u$. The training objective is to minimize adversarial loss, which translates to minimizing the distribution divergence, and push the model to produce invariant representations across the sets.

$$\min_g d_{\mathcal{H}}\left(\mathcal{D}_l, \mathcal{D}_u\right) = \max_g \min_{h \in \mathcal{H}} \left[\text{err}\left(h, g, \mathcal{D}_l\right) + \text{err}\left(h, g, \mathcal{D}_u\right)\right]$$

Data augmentation The training objective above can be unstable from small number of training examples n. The authors propose a Mixup based data augmentation technique. New training samples can be produces by interpolating between labeled and unlabeled examples. For a labeled $\mathbf{x}_l$ and unlabeled examples $\mathbf{x}_u$. We can create a new input $\tilde{x}$, classification target $\tilde{y}$ and discriminator target $\tilde{z}$

$$\tilde{\mathbf{x}} = \lambda \mathbf{x}^l + (1-\lambda)\mathbf{x}^u$$
$$\tilde{y} = \lambda y^l + (1-\lambda)\hat{y}^u$$
$$\tilde{z} = \lambda \cdot 0 + (1-\lambda)\cdot 1$$

with $\lambda \sim \beta(\alpha, \alpha)$ and α as a hyper-parameter.

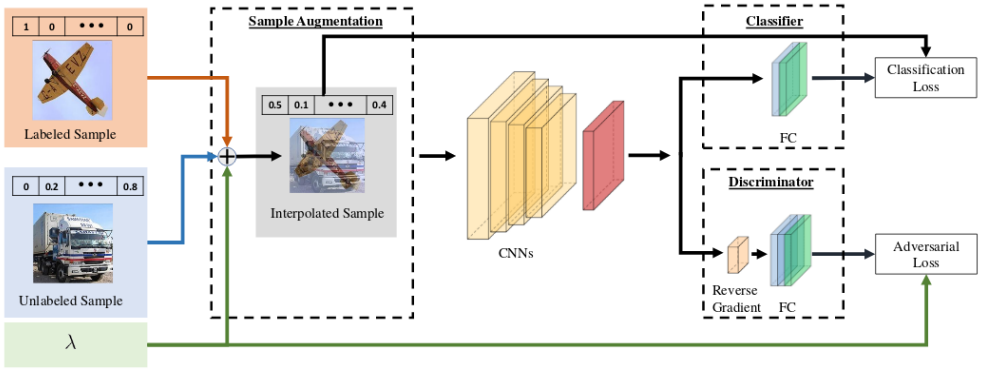

Figure 3. The network architecture of our proposed ADA-Net, in which we append an additional discriminator classifier branch with a gadient reverse layer to the vanilla CNN (shown in the bottom right part). In training time, the cross-set sample interpolation is performed between labeled and unlabeled samples, and we feed the interpolated samples into the network. Pesudo-labels of unlabeled samples are obtained using the classifier trained in last iteration (see explanation in Section 4.3) for details.

Results

Table 1. Classfication error rates of our proposed ADA-Net and its variants on the CIFAR10 and SVHN datasets. "dist" denotes the distribution alignment module, and "aug" denotes the cross-set sample augmentation module. PreAct-ResNet-18 [26] is used as the backbone network.

	dist	aug	CIFAR10	SVHN
Baseline			19.97%	13.80%
Ours	✓		18.67%	10.76%
		✓	13.79%	10.74%
	✓	✓	**8.87%**	**5.90%**

Table 1. Classfication error rates of our proposed ADA-Net and its variants on the CIFAR10 and SVHN datasets. "dist" denotes the distribution alignment module, and "aug" denotes the cross-set sample augmentation module. PreAct-ResNet-18 [26] is used as the backbone network.

	dist	aug	CIFAR10	SVHN
Baseline			19.97%	13.80%
Ours	✓		18.67%	10.76%
		✓	13.79%	10.74%
	✓	✓	**8.87%**	**5.90%**

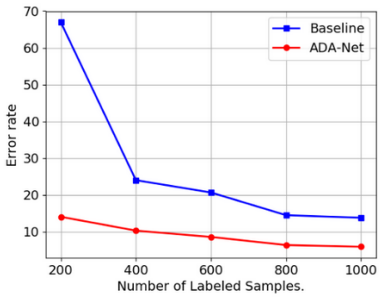

Figure 5. Classification Error rates on SVHN of our ADA-Net and baseline CNN when varying the number of labeled samples.

Table 2. Classification error rates of different methods on CIFAR10 and SVNH. Conv-Large [47] is used as the backbone network. Results of baseline methods are taken from the papers.

Method	**CIFAR10**	**SVHN**
Π Model [32]	12.36%	4.82%
Temporal ensembling [32]	12.16%	4.42%
Mean Teacher[47]	12.31%	3.95%
VAT [37]	11.36%	5.42%
VAT+Ent [37]	10.55%	3.86%
SaaS [11]	13.22%	4.77%
MA-DNN [10]	11.91%	4.21%
VAT+Ent+SNTG [35]	9.89%	3.83%
Mean Teacher+fastSWA* [2]	9.05%	-
ADA-Net (Ours)	10.30%	4.62%
ADA-Net+ (Ours)	10.09%	**3.54%**
ADA-Net*(Ours)	**8.72%**	-

* Larger translation range (4 instead of 2), and without ZCA whitening.

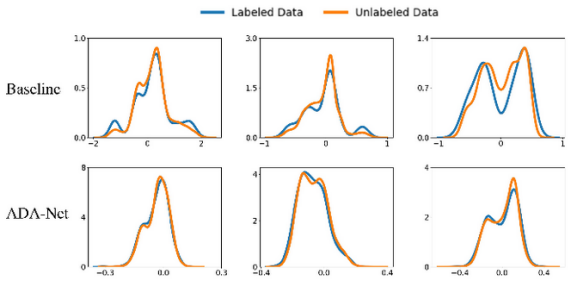

Figure 6. Kernel density estimation of labeled and unlabeled samples of the SVHN Dataset based on the first three feature activations of the baseline CNN and our ADA-Net. Considerable distribution mismatch between labeled and unlabeled data can be observed for the baseline CNN model (top row), while two distributions are generally aligned well with our ADA-Net (bottom row).