Semi-Supervised Learning by Augmented Distribution Alignment

Abstract

- 本文揭示了一个存在于现在SSL方法里, 由于有限有标签数据集的原因而导致的sampling bias问题. 这个经常会导致在有标签和无标签数据上的empirical distribution mismatch问题. 为了解决这个问题, 本文提出了一种distribution alignment的方法
- 本文采用了adversarial training去最小化有标签和无标签数据的distribution distance, 同时为了解决只有很少有标签数据的问题, 作者还提出了一个简单的数据扩充方法去产生pseudo training samples

Introduction

- 半监督学习旨在通过使用有限的有标签数据和大量的无标签数据学习到一个robust的模型
- 近年来很多方法被提出用于解决SSL问题, 但essential sampling bias issue没有被多少研究. 这个问题指的是 : ***the empirical distribution of labeled data often deviates from the true samples distribution, due to the limited sampling size of labeled data*** 如下图所示

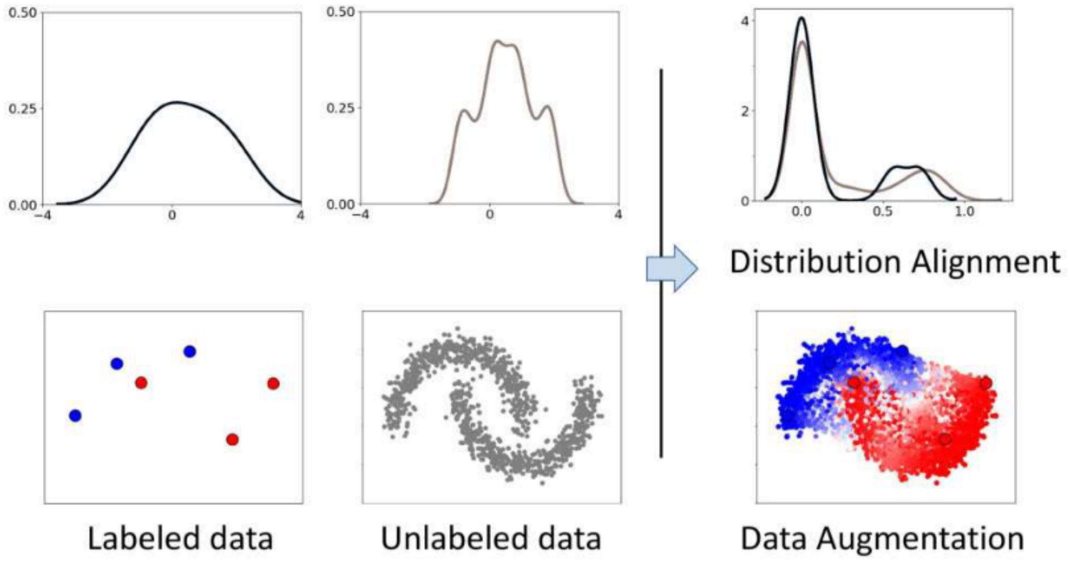


Figure 1. Illustration of the empirical distribution mismatch between labeled and unlabeled samples with the two-moon data. The labeled and unlabeled samples are shown in the **bottom left** and **bottom middle** figures, and the kernel density estimations of their x-axis projection are plotted in the **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**).

值得注意的是, 本文也是在feature space做distribution alignment

- 可以看到, 由于仅仅只有随机采样有标签数据的数量限制, 有标签数据的分布和无标签数据的分布是不一致的, 虽然它们也是从整体的分布中采样. 同样的empirical distribution mismatch在真实世界中也是很常见, 比如做domain adaption时, 可能会由于分布的差异而导致模型性能的下降. 因此SSL方法也会潜在的受到empirical distribution mismatch的影响
- 为了解决这个问题, 本文提出一种显式减少distribution mismatch的方法, Augmented Distribution Alignment. 一方面使用adversarial training strategy在feature space中去minimize有标签和无标签数据的distribution distance, 另一方面通过使用data augmentation产生pseudo samples增强有标签数据少的问题
- 本文方法可以方便的实现, 并集成到现有的SSL放上, 在SVHN和CIFAR10上进行实验, 达到了SOTA效果

Related Work

- **Semi-Supervised Learning** 由于在实际场景中, 有标签的数据获得需要有很高的代价, 因此SSL更受青睐. 不同于其他SSL的方法, 我们从empirical distribution mismatch的角度去解决SSL的问题. 并且由于我们从一个新的角度去解决这个问题, 我们的方法可能可以和其他SSL方法作为补充
- **Sampling bias problem** 通常在supervised learning和domain adapation场景中被考虑, 但在SSL中还较少的被考虑到. 有很多方法被提出在学习过程中解决sampling bias的问题. 近期根据GAN的想法, adversarial training strategy被广泛用于empirical distribution mismatch问题. 在domain adaption中, 考虑两个domain的数据是不同的分布, 在SSL中则是independent identical distribution, 但减少domain distribution mismatch的方法同样可以用于SSL
- **Other Related Works** 我们的方法也和最近提出的data augmentation的方法相关, 特别的, MixUp. 为了解决有标签数据样本少的问题, 我们使用了MixUp并且结合Pseudo-label在无标签数据中, 并且本文也展示了使用data augmentation的确可以让有标签和无标签的数据分布靠近

Problem Statement and Motivations

本文采用标准的SSL Setting. n个有标签数据集, m个无标签数据集, 其中$m > n$, 希望使用较少的有标签数据集去利用大量的无标签数据集去训练一个性能好的模型

- Empirical Distribution Mismatch in SSL

  在SSL的setting中, 有标签数据和无标签数据通常被假设从一个i. i. d的分布中被抽取, 然而由于有标签数据的数量较少, 无法反映出整个数据的真实分布, 这就造成了本文提出的empirical Distribution Mismatch问题.

  少的有标签的数据集不能represent真实的分布, 这点可以通过将分布投影到X轴上观察得出. 这个现象被称之为**sampling bias**, 这个可以被Maximum Mean Distance测量. 在SSL中, 有标签数据和无标签数据是被假设为I.I.D的, 虽然如此, 但由于有标签数据的数量很少, 并不能反映出和unlabeled data一样的分布模式, **因此在实际中其实并不是I.I.D** (和我做的挺相似哈哈)

  通过下图可以发现, MMD随着labeled data的增加, labeled data和unlabeled data的分布在减小

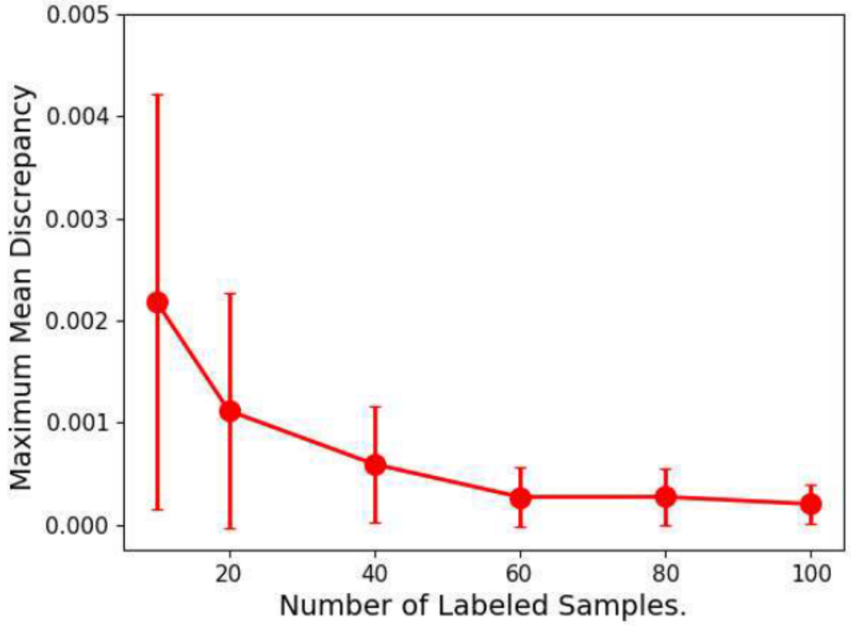
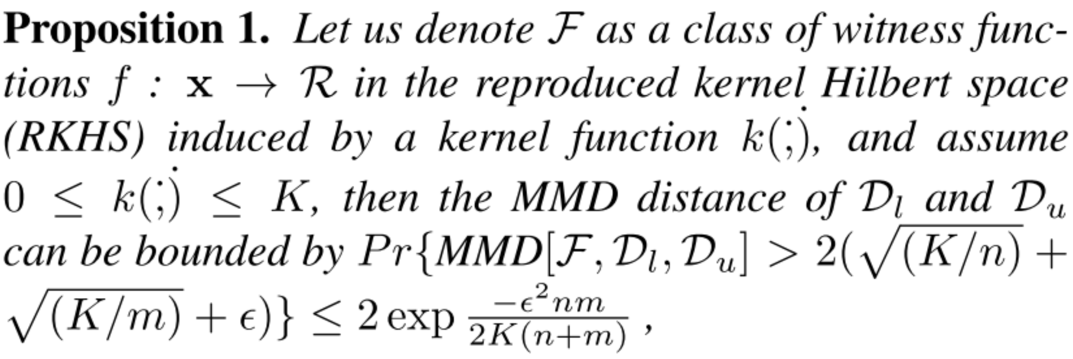
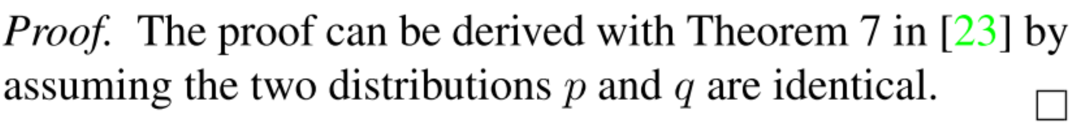


Figure 2. MMD between labeled and unlabeled samples in two-moon example with varying number of labeled samples. Number of unlabeled sample is fixed as $1,000$.

两个数据集如果I.I.D, 则它们的Maximum Mean Distance距离的界可以被计算为



Proposition 1. *Let us denote $\mathcal{F}$ as a class of witness functions $f : \mathbf{x} \to \mathcal{R}$ in the reproduced kernel Hilbert space (RKHS) induced by a kernel function $k(\dot{;})$, and assume $0 \le k(\dot{;}) \le K$, then the MMD distance of $\mathcal{D}_l$ and $\mathcal{D}_u$ can be bounded by* $Pr\{MMD[\mathcal{F}, \mathcal{D}_l, \mathcal{D}_u] > 2(\sqrt{(K/n)} + \sqrt{(K/m)} + \epsilon)\} \le 2\exp\frac{-\epsilon^2 nm}{2K(n+m)}$,

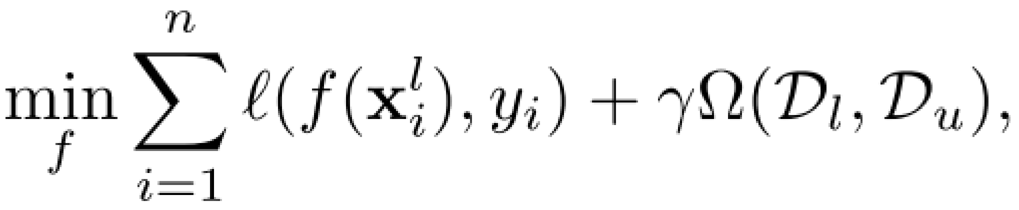


Proof. The proof can be derived with Theorem 7 in [23] by assuming the two distributions p and q are identical. □

这一点表明了在SSL中, labeled data由于采样数量的限制, 会导致empirical approximatetion of labeled data偏离真实的数据分布, 会导致在empirical distribution上训练的模型不能有很好的generalize 在test data上. 并且有很多方法都是基于pseudo labeling的, 如果原本模型无法学习到一个近似分布, 则pseudo-label的准确性会大幅降低. 这可能是在常规SSL方法中最重要的几个导致潜在不稳定的原因之一. [39]也证实了SSL方法的性能可能会降低, 如果labeled dataset数量下降

- Healing the Empirical Distribution MisMatch

为了解决empirical distribution mismatch问题, 本文提出了Augmented Distribution Alignment方法. 形式化的, 整体objective function如下



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

其中第二项是分布之间的差异, γ是超参数. 同时数量少的labeled data可能潜在的导致训练过程unstable, 因此本文中还提出了数据增强的方法, 使用MixUp同时对labeled data和unlabeled data进行数据增强

Augmented Distribution Alignment

整个流程由两个方法组成 adversarial distirbution alignment和cross-set sample augmentation

- Adversarial Distribution Alignment

  使用[4, 12]的*H*-Divergence去测量两个数据集的分布 (MMD?), 使用adversarial training, 在labeled data中一个binary discriminator预测为0 , 对 unlabeled data预测为1, 整体objective function为

  $$d_{\mathcal{H}}(\mathcal{D}_l, \mathcal{D}_u)=2\left\{1-\min_{h\in\mathcal{H}}\left[err(h,g,\mathcal{D}_l)+err(h,g,\mathcal{D}_u)\right]\right\}$$

  只觉来说, 当empirical distribution 的差异很大的时候, discriminator能够容易的区分来自两个数据集的数据. 因此*err*的值会小, *H*-divergence就会变大. 因此我们训练去最小化*H*, 这样就会让feature extractor将两个数据集的特征很好的对齐, 如下是adversarial training的mix-max objective function. [17]表明这个可以使用一个简单的gradient reverse layer去实现

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

- Cross-set Sample Augmentation

  在SSL中, 有限采样的labeled data会导致unstable in optimization使得性能下降. 因此本文设计了对labeled/unlabeled数据进增强的方法. 特别地, 对于每个unlabeled数据, 使用模型预测一个pseudo-label. 设x^l为labeled数据, x^u为unlabeled数据, cross-set augment可以被表示为如下, $\tilde{z}$为distribution discriminator的label

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

  这样做的好处有两点 1) 增加了数据集 2) 由于是从两个数据集生成的pseudo sample, 因此它们的分布比原来的label training sample更加接近

- Summary

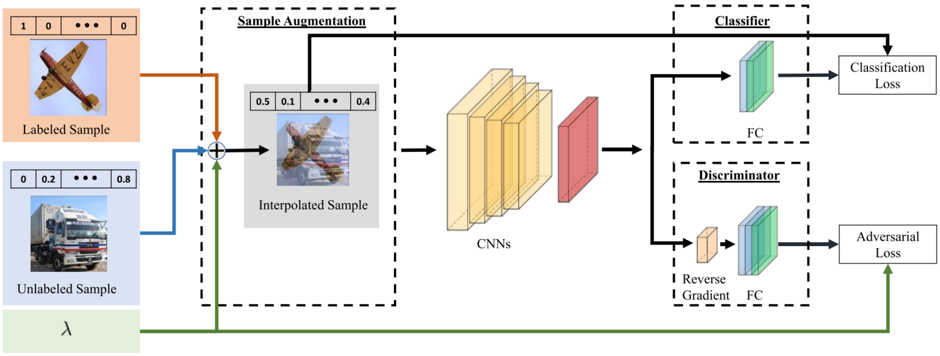
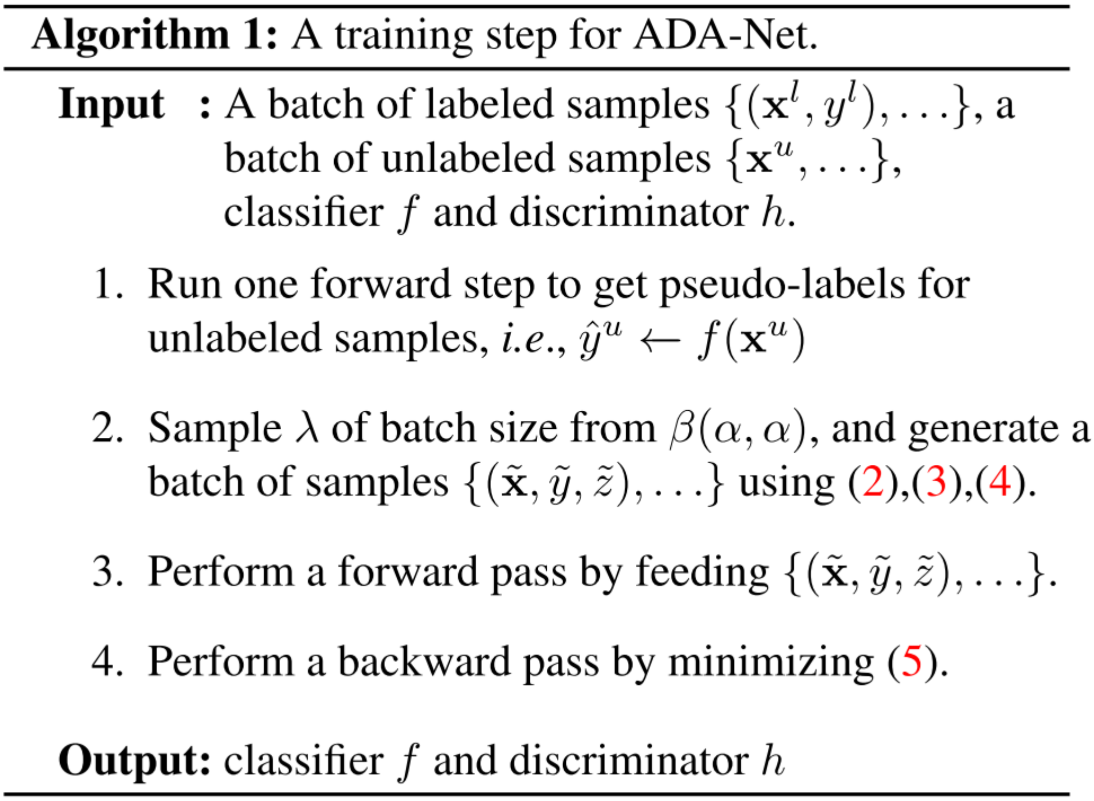


Figure 3. The network architecture of our proposed ADA-Net, in which we append an additional discriminator classifier branch with a gadient reverse layer to the vanilla CNN (shown in the bottom right part). In training time, the cross-set sample interpolation is performed between labeled and unlabeled 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.



```
Algorithm 1: A training step for ADA-Net.
Input  : A batch of labeled samples $\{(\mathbf{x}^l, y^l), \ldots\}$, a
         batch of unlabeled samples $\{\mathbf{x}^u, \ldots\}$,
         classifier $f$ and discriminator $h$.
    1. Run one forward step to get pseudo-labels for
       unlabeled samples, i.e., $\hat{y}^u \leftarrow f(\mathbf{x}^u)$
    2. Sample $\lambda$ of batch size from $\beta(\alpha, \alpha)$, and generate a
       batch of samples $\{(\tilde{\mathbf{x}}, \tilde{y}, \tilde{z}), \ldots\}$ using (2),(3),(4).
    3. Perform a forward pass by feeding $\{(\tilde{\mathbf{x}}, \tilde{y}, \tilde{z}), \ldots\}$.
    4. Perform a backward pass by minimizing (5).
Output: classifier $f$ and discriminator $h$
```

Experiment

数据集使用SVHN和CIFAR10. 对于SVHN, 使用1000张图片作为supervised learning, 仅使用Random translation作为增强方法. 对于CIFAR10, 使用4000张图片作为labeled data

- result

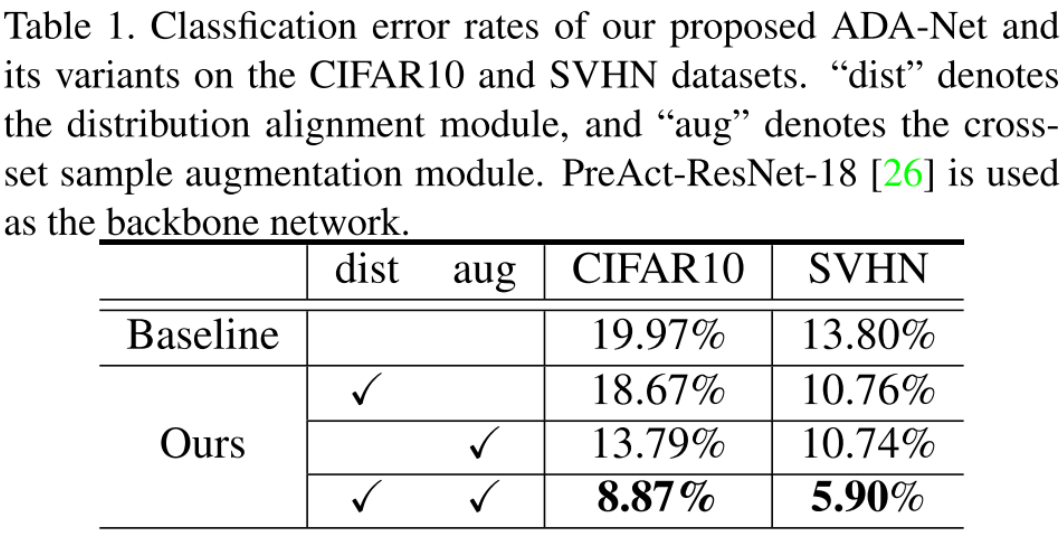


Table 1. Classfication error rates of our proposed ADA-Net and its variants on the CIFAR10 and SVHN datasets. “dist” denotes the distribution alignment module, and “aug” denotes the cross-set sample augmentation module. PreAct-ResNet-18 [26] is used as the backbone network.

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

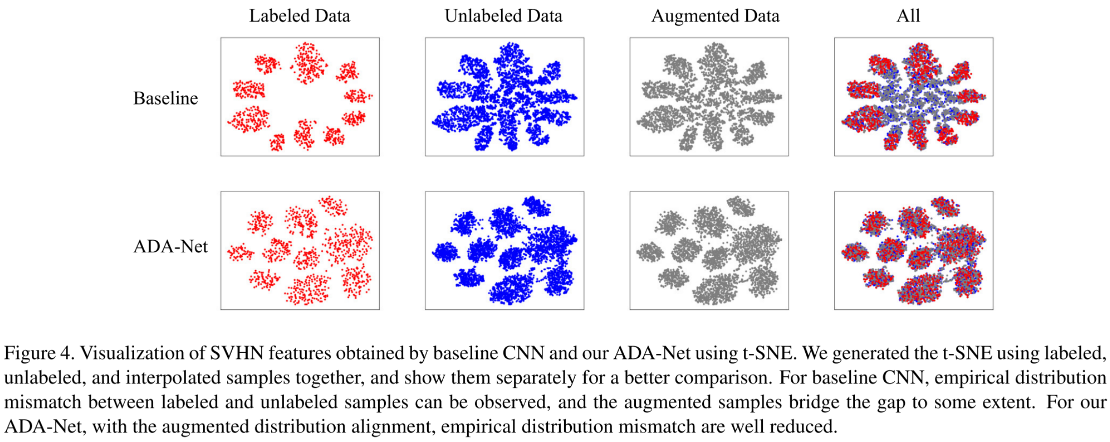


Figure 4. Visualization of SVHN features obtained by baseline CNN and our ADA-Net using t-SNE. We generated the t-SNE using labeled, unlabeled, and interpolated samples together, and show them separately for a better comparison. For baseline CNN, empirical distribution mismatch between labeled and unlabeled samples can be observed, and the augmented samples bridge the gap to some extent. For our ADA-Net, with the augmented distribution alignment, empirical distribution mismatch are well reduced.

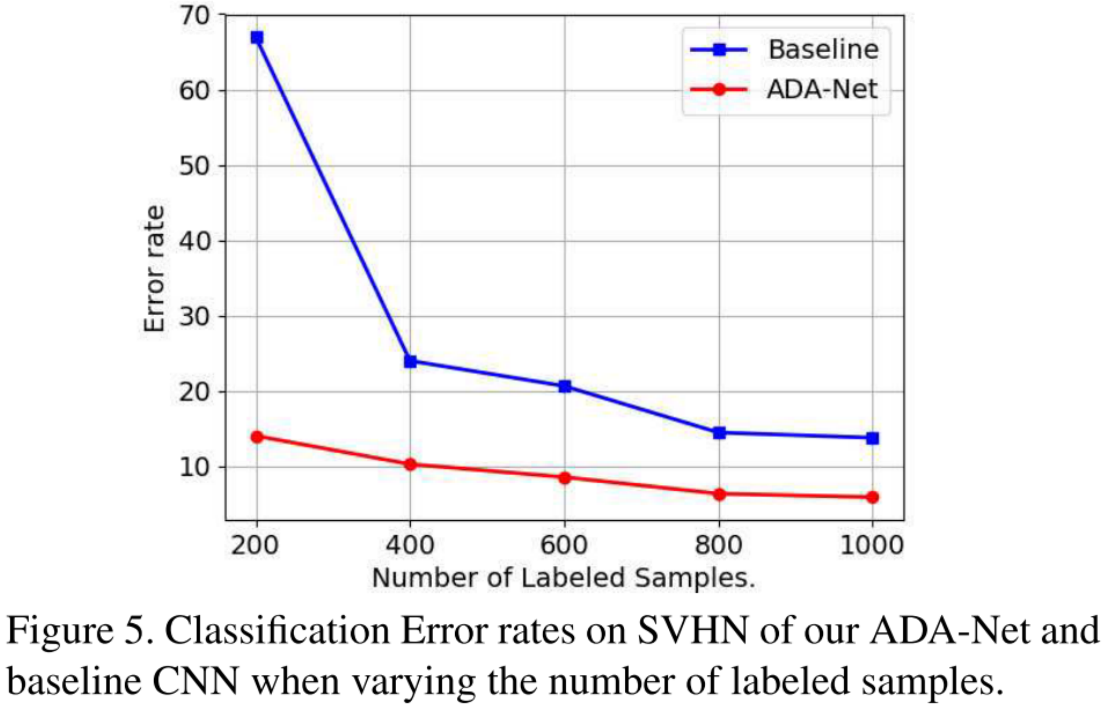


Figure 5. Classification Error rates on SVHN of our ADA-Net and baseline CNN when varying the number of labeled samples.

- comparison with SOTA

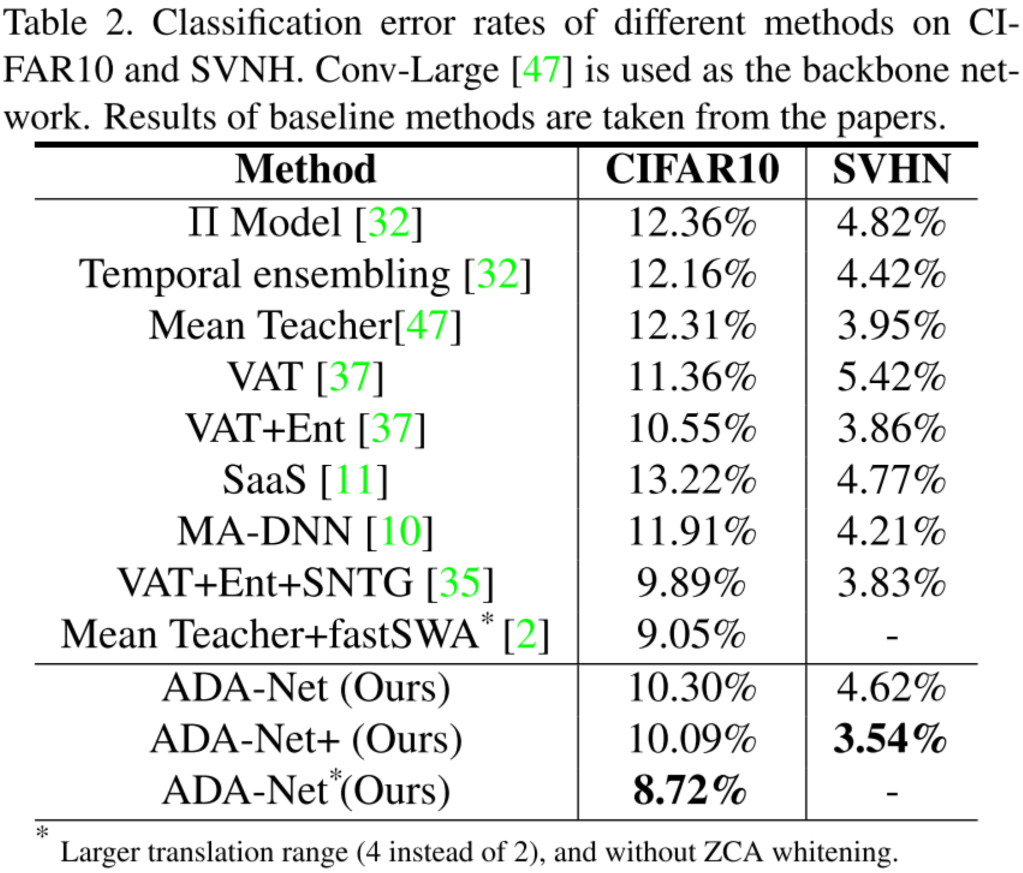


Table 2. Classification error rates of different methods on CIFAR10 and SVHN. Conv-Large [47] is used as the backbone network. Results of baseline methods are taken from the papers.

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.

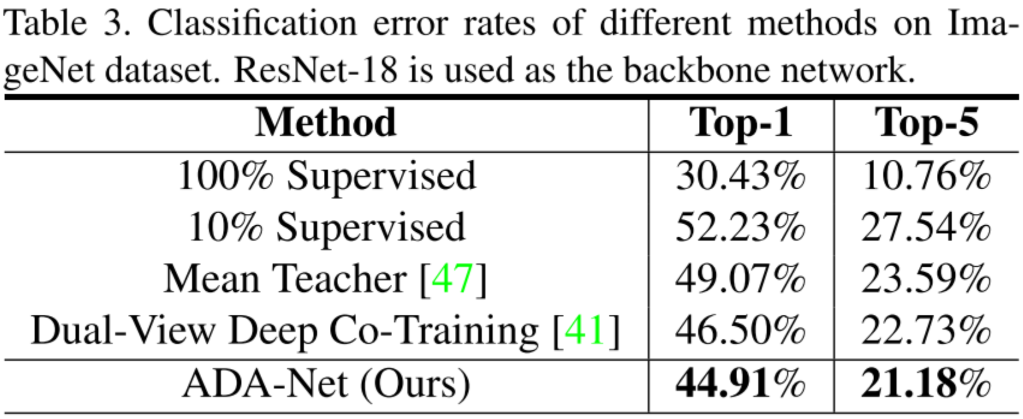


Table 3. Classification error rates of different methods on ImageNet dataset. ResNet-18 is used as the backbone network.

Method	Top-1	Top-5
100% Supervised	30.43%	10.76%
10% Supervised	52.23%	27.54%
Mean Teacher [47]	49.07%	23.59%
Dual-View Deep Co-Training [41]	46.50%	22.73%
ADA-Net (Ours)	**44.91%**	**21.18%**

个人小结

文章大概意思和之前那一篇free lunch distribution calibration 想法相似, 提出了两个数据集的数据虽然假设是I.I.D, 但实际中并不是. 因此提出了distribution alignment, calibration和alignment都可以吧. 本文展示了两个数据集分布之间的差异, 通过理论证明了在I.I.D的情况下, 两个分布的maximum mean discrepancy的bound. 并且提出使用adversarial training的方法来使得extractor能够学习到相同的feature, 也就是distribution alignment. 然后针对labeled数据较少问题, 提出了一种数据增强的机制, 其中有部分是来自labeled data一部分来自unlabeled data, 作者认为这样产生的数据在分布上更加靠近

这两天看的文章都有一个核心, 直指distribution alignment, 值得继续跟进