Tangent-Normal Adversarial Regularization for Semi-supervised Learning

Bing Yu*, **Jingfeng Wu***, Jinwen Ma, **Zhanxing Zhu**

{byu, pkuwjf, zhanxing.zhu}@pku.edu.cn

Peking University & Beijing Institute of Big Data Research

Abstract

We propose a **tangent-normal adversarial regularization** for semi-supervised learning (SSL). It is composed by

1. tangent adversarial regularization, which enforces the local smoothness of the classifier along the underlying manifold;
2. normal adversarial regularization, which imposes robustness on the classifier against the noise carried in the observed data.

Empirically, TNAR achieves state-of-the-art performance for semi-supervised learning.

Motivation

Semi-supervised Learning

Input

- Insufficient amount of labeled data (x_l, y_l);
- Sufficient amount of unlabeled data x_{ul}.

Output A learned classifier fully utilizing both labeled and unlabeled data.

Assumptions

1. **The manifold assumption** The observed data $x \in \mathbb{R}^D$ is almost concentrated on a low dimensional underlying manifold $\mathcal{M} \cong \mathbb{R}^d, d \ll D$.
2. **The noisy observation assumption** The observed data can be decomposed as $x = x_0 + n$, where x_0 is exactly supported on the manifold $\mathcal{M}$ and n is some noise independent of x_0.
3. **The semi-supervised learning assumption** The true classifier, or the true condition distribution $p(y|X)$ varies smoothly along the underlying manifold $\mathcal{M}$.

Thus a good classifier for SSL should be

- Smooth along the underlying manifold $\mathcal{M}$; $\leftarrow$ TAR
- Robust to the off manifold noise n. $\leftarrow$ NAR

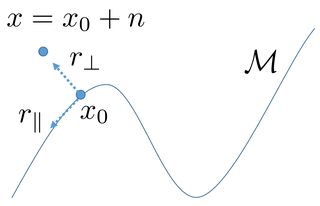

Figure 1: Illustration for tangent-normal adversarial regularization. $x = x_0 + n$ is the observed data, where x_0 is exactly supported on the underlying manifold $\mathcal{M}$ and n is the noise independent of x_0. $r_\parallel$ is the adversarial perturbation along the tangent space to induce invariance of the classifier on manifold; $r_\perp$ is the adversarial perturbation along the normal space to impose robustness on the classifier against noise n.

Notations

$(x_l, y_l), x_{ul}$ labeled example, unlabeled example.

$\mathcal{D}, \mathcal{D}_l, \mathcal{D}_{ul}$ full dataset, labeled dataset, unlabeled dataset.

$p(y|x;\theta)$ **or** $f(x;\theta)$ the classifier to be optimized.

$\mathbb{R}^D, \mathcal{M}$ the observed space and the data manifold.

x, z the coordinates of an example in the observed space $\mathbb{R}^D$ and on the manifold $\mathcal{M}$ respectively.

g, h the generator (decoder) and the encoder.

$T_x\mathcal{M} = J_z g(z) \cong \mathbb{R}^d, z = h(x)$, the tangent space, or the span of the columns of the Jacobian of g.

Tangent-Normal Adversarial Regularization (TNAR)

The proposed loss

$$L(D_l, D_{ul}; \theta) := \mathbb{E}_{(x_l,y_l)\in\mathcal{D}_l} \ell\left(y_l, p(y|x_l;\theta)\right) + \alpha_1 \mathbb{E}_{x\in\mathcal{D}} \mathcal{R}_{\text{tangent}}(x;\theta) + \alpha_2 \mathbb{E}_{x\in\mathcal{D}} \mathcal{R}_{\text{normal}}(x;\theta) \left(+\alpha_3 \mathbb{E}_{x\in\mathcal{D}} \mathcal{R}_{\text{entropy}}(x;\theta)\right) \tag{1}$$

- **Supervised loss**: ℓ

- **Tagent adversarial regularization (TAR)**:

$$\mathcal{R}_{\text{tangent}}(x;\theta) = \max_{\substack{\|r\|_2\le\epsilon, \\ r\in T_x\mathcal{M}=J_z g(\mathbb{R}^d)}} dist(p(y|x;\theta), p(y|x+r;\theta)), \tag{2}$$

- **Normal adversarial regularization (NAR)**:

$$\mathcal{R}_{\text{normal}}(x;\theta) = \max_{\substack{\|r\|_2\le\epsilon, \\ r\perp T_x\mathcal{M}}} dist(p(y|x;\theta), p(y|x+r;\theta)). \tag{3}$$

Key components for TNAR

1. Identify manifold;
2. Perform *virtual adversarial regularization* along tangent space;
3. Perform *virtual adversarial regularization* along normal space.

Manifold

Generative models with both encoder and decoder could be used to describe the data manifold, e.g.,

- Variational Autoencoder;
- Localized GAN;
- Other generative models like Denoise Autoencoder, Flow, BiGAN, etc.

Tangent adversarial regularization

Using the trick introduced by Miyato et.al and taking Taylor's expansion we have

$$\mathcal{R}_{\text{tangent}}(x;\theta) = \max_{\substack{\|r\|_2\le\epsilon, \\ r\in T_x\mathcal{M}=J_z g(\mathbb{R}^d)}} \frac{1}{2} r^T H r, \tag{4}$$

The vanishing of the first two terms in Taylors expansion occurs because that $dist(\cdot,\cdot)$ is some distance measure with 1) minimum zero and 2) $r = 0$ is the optimal value.

Problem (4) is equivalent to

$$\begin{aligned} \underset{r\in\mathbb{R}^D}{\text{maximize}} \quad & \frac{1}{2} r^T H r \\ \text{s.t.} \quad & \|r\|_2 \le \epsilon, \quad r = J\cdot\eta, \quad \eta \in \mathbb{R}^d. \quad (J := J_z g \in \mathbb{R}^{D\times d}) \end{aligned} \tag{5}$$

This is a *generalized eigenvalue problem* and could be solved by *power iteration* and *conjugate gradient* **in constant times of back-propagating**.

Normal adversarial regularization

Similarly, we approximately reformate NAR (3) as

$$\begin{aligned} \underset{r\in\mathbb{R}^D}{\text{maximize}} \quad & \frac{1}{2} r^T H r - \lambda r^T (r_\parallel r_\parallel^T) r \\ \text{s.t.} \quad & \|r\|_2 \le \epsilon, \end{aligned} \tag{6}$$

where $r_\parallel$ is the perturbation obtained in TAR. It is again an *eigenvalue problem* and could be solved by *power iteration* **in constant times of back-propagating**.

Experiments

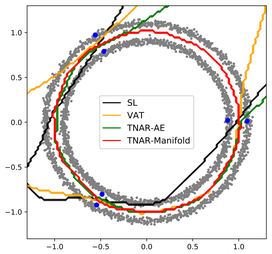

Figure 2: The decision boundaries of compared methods on two-rings artificial dataset. Gray dots distributed on two rings: the unlabeled data. Blue dots (3 in each ring): the labeled data. Colored curves: the decision boundaries found by compared methods.

Table 1: Classification errors (%) of compared methods on SVHN and CIFAR-10 with data augmentation.

Method	SVHN 1,000 labels	CIFAR-10 4,000 labels
VAT (large)	3.86 ± 0.11	10.55 ± 0.05
VAT + SNTG	3.83 ± 0.22	9.89 ± 0.34
Π model	4.82 ± 0.17	12.36 ± 0.31
Temporal ensembling	4.42 ± 0.16	12.16 ± 0.24
Mean Teacher	3.95 ± 0.19	12.31 ± 0.28
LGAN	-	9.77 ± 0.13
TNAR-VAE (large)	$\mathbf{3.74 \pm 0.04}$	$\mathbf{8.85 \pm 0.03}$

Table 2: Classification errors (%) of compared methods on SVHN and CIFAR-10 without data augmentation.

Method	SVHN 1,000 labels	CIFAR-10 4,000 labels
VAT (small)	6.83 ± 0.24	14.87 ± 0.13
VAT (large)	4.28 ± 0.10	13.15 ± 0.21
VAT + SNTG	4.02 ± 0.20	12.49 ± 0.36
Π model	5.43 ± 0.25	16.55 ± 0.29
Mean Teacher	5.21 ± 0.21	17.74 ± 0.30
CCLP	5.69 ± 0.28	18.57 ± 0.41
ALI	7.41 ± 0.65	17.99 ± 1.62
Improved GAN	8.11 ± 1.3	18.63 ± 2.32
Tripple GAN	5.77 ± 0.17	16.99 ± 0.36
Bad GAN	4.25 ± 0.03	14.41 ± 0.30
LGAN	4.73 ± 0.16	14.23 ± 0.27
Improved GAN + JacobRegu + tangent	4.39 ± 1.20	16.20 ± 1.60
Improved GAN + ManiRegu	4.51 ± 0.22	14.45 ± 0.21
TNAR-LGAN (small)	4.25 ± 0.09	12.97 ± 0.31
TNAR-LGAN (large)	4.03 ± 0.13	12.76 ± 0.04
TNAR-VAE (small)	3.99 ± 0.08	12.39 ± 0.11
TNAR-VAE (large)	$\mathbf{3.80 \pm 0.12}$	$\mathbf{12.06 \pm 0.35}$
TAR-VAE (large)	5.62 ± 0.19	13.87 ± 0.32
NAR-VAE (large)	4.05 ± 0.04	15.91 ± 0.09

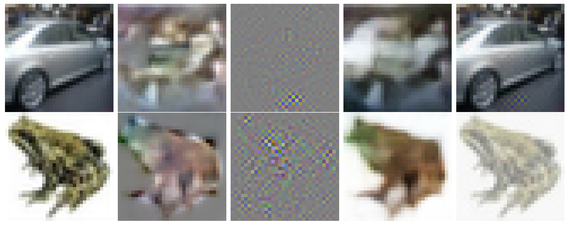

Figure 3: The perturbations and adversarial examples in tangent space and normal space for CIFAR-10 dataset. Note that the perturbations is actually too small to distinguish easily, thus we show the scaled perturbations. From left to right: original example, tangent adversarial perturbation, normal adversarial perturbation, tangent adversarial example, normal adversarial example.

Conclusion

We present the tangent-normal adversarial regularization for semi-supervised learning, a novel regularization strategy based on virtual adversarial training and manifold regularization. TNAR is composed of regularization on the tangent and normal space separately. The tangent adversarial regularization enforces manifold invariance of the classifier, while the normal adversarial regularization imposes robustness of the classifier against the noise contained in the observed data. Experiments on synthetic and real datasets demonstrate the effectiveness of our method.

References

[1] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

[2] Abhishek Kumar, Prasanna Sattigeri, and Tom Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. In *Advances in Neural Information Processing Systems*, pages 5540–5550, 2017.

[3] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *arXiv preprint arXiv:1704.03976*, 2017.