Tangent-Normal Adversarial Regularization for Semi-Supervised Learning

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

Peking University Beijing Institute of Big Data Research

June, 2019

Semi-supervised learning (SSL)

- Suppose we have insufficient amount of labeled data (x_l, y_l) and large amount of unlabeled data x_{ul};
- How to learn a classifier fully utilizing the unlabeled data x_{ul}?

One important approach: Manifold Regularization!
The key motivation is that unlabeled data could help to identify a good data manifold.

Assumptions (informal)

The manifold assumption The observed data $x \in \mathbb{R}^D$ is almost concentrated on a low dimensional underlying manifold $\mathcal{M} \cong \mathbb{R}^d, d \ll D$.

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.

The semi-supervised learning assumption The true classifier, or the true condition distribution $p(y|X)$ varies smoothly along the underlying manifold $\mathcal{M}$.

Introduce TNAR: Tangent-Normal Adversarial Regularization

Based on the assumptions, a good classifier for semi-supervised learning should be:

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

To this end, we propose *tangent-normal adversarial regularization (TNAR)*.

TNAR: Tangent-Normal Adversarial Regularization

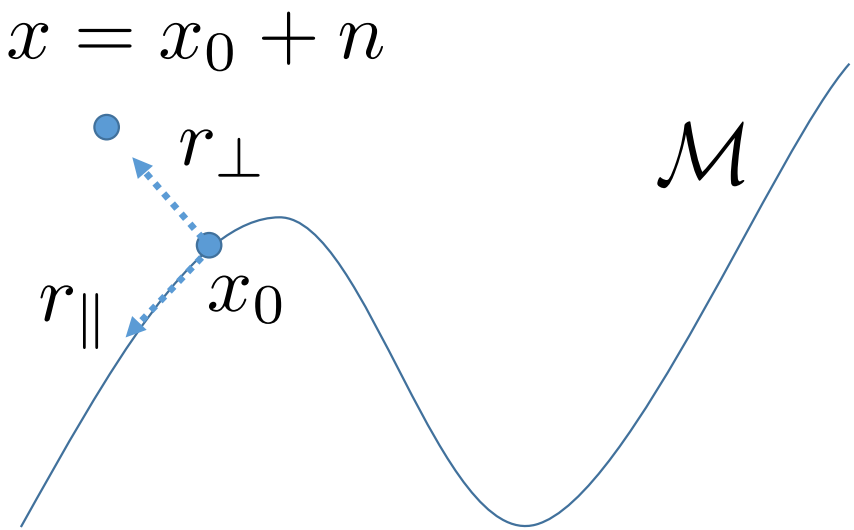

Figure: Illustration for the *tangent-normal adversarial regularization*. $r_\parallel$ is the adversarial perturbation along the tangent space to induce invariance of the classifier on manifold;
$r_\perp$ is the adversarial perturbation along the normal space to impose robustness on the classifier against 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.

Overview of the TNAR loss

The proposed loss for SSL is

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

ℓ is the supervised loss and TAR and NAR are:

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

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

Elaborate TNAR (= TAR + NAR)

Part 1: Manifold Identify the underlying data manifold $\mathcal{M}$ (or its tangent space $T_x\mathcal{M}$).

Part 2: Tangent Adversarial Regularization (TAR) Perform virtual adversarial training along $T_x\mathcal{M}$, to enforce the local smoothness of the classifier along the underlying manifold.

Part 3: Normal Adversarial Regularization (NAR) Perform virtual adversarial training along $(T_x\mathcal{M})^{\perp}$, to impose robustness on the classifier against the noise carried in the observed data.

Part 1: Identify the underlying manifold $\mathcal{M}$

Generative models with both encoder and decoder could be used to describe the data manifold

- VAE;
- Localized GAN;
- Other generative models like denoise AE, Flow, BiGAN, etc.

Key observation to Part 2 and 3

$$F(x, r, \theta) := dist(p(y|x;\theta), p(y|x+r;\theta)) \approx \frac{1}{2} r^T H r. \quad (4)$$

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

Thus

$$\mathcal{R}_{\text{tangent}}(x;\theta) = \max_{\substack{\|r\|_2 \le \epsilon, \\ r \in T_x\mathcal{M} = J_z g(\mathbb{R}^d)}} \frac{1}{2} r^T H r, \quad (5)$$

$$\mathcal{R}_{\text{normal}}(x;\theta) = \max_{\substack{\|r\|_2 \le \epsilon, \\ r \perp T_x\mathcal{M}}} \frac{1}{2} r^T H r. \quad (6)$$

Part 2: Tangent Adversarial Regularization

To optimize TAR

$$\mathcal{R}_{\text{tangent}}(x;\theta) = \max_{\substack{\|r\|_2 \le \epsilon, \\ r \in T_x\mathcal{M} = J_z g(\mathbb{R}^d)}} \frac{1}{2} r^T H r \tag{7}$$

is equivalent to solve:

$$\begin{aligned} \underset{r \in \mathbb{R}^D}{\text{maximize}} \quad & \frac{1}{2} r^T H r \\ \text{s.t.} \quad & \|r\|_2 \le \epsilon \\ & r = J \cdot \eta, \quad \eta \in \mathbb{R}^d. \quad (J := J_z g \in \mathbb{R}^{D \times d}) \end{aligned} \tag{8}$$

Part 2: Tangent Adversarial Regularization

Eliminate r, we have

$$\begin{aligned} \underset{\eta \in \mathbb{R}^d}{\text{maximize}} \quad & \frac{1}{2}\eta^T J^T H J \eta \\ \text{s.t.} \quad & \eta^T J^T J \eta \leq \epsilon^2. \end{aligned} \tag{9}$$

This is a *generalized eigenvalue problem* and could be solved by power iteration and conjugate gradient as

$$\begin{aligned} v &\leftarrow J^T H J \eta \\ \mu &\leftarrow (J^T J)^{-1} v \\ \eta &\leftarrow \frac{\mu}{\|\mu\|_2}. \end{aligned} \tag{10}$$

Fortunately, all the above update could be computed efficiently **in constant times of back-propagating**.

Part 3: Normal Adversarial Regularization

In a same spirit with TAR and some relaxation, we could solve NAR

$$\mathcal{R}_{\text{normal}}(x;\theta) = \max_{\substack{\|r\|_2 \le \epsilon, \\ r \perp T_x\mathcal{M}}} \frac{1}{2} r^T H r \tag{11}$$

by

$$\begin{aligned} \underset{r \in \mathbb{R}^D}{\text{maximize}} \quad & \frac{1}{2} r^T H r - \lambda r^T (r_{\|} r_{\|}^T) r \\ \text{s.t.} \quad & \|r\|_2 \le \epsilon, \end{aligned} \tag{12}$$

where $r_{\|}$ is the perturbation obtained in TAR.
It is again an *eigenvalue problem* and could be solved **in constant times of back-propagating**.

The final loss

As suggested by Miyato et al., *entropy regularization* benefits VAT hence TNAR since it ensures the model to predict more determinately,

$$\mathcal{R}_{\text{entropy}}(x;\theta) := -\sum_{y} p(y|x;\theta)\log p(y|x;\theta). \tag{13}$$

The final proposed loss for SSL is

$$\begin{aligned} L(D_l, D_{ul}, \theta) :=& \mathbb{E}_{(x_l,y_l)\in\mathcal{D}_l}\ell\left(y_l, p(y|x_l;\theta)\right) \\ &+ \alpha_1\mathbb{E}_{x\in\mathcal{D}}\mathcal{R}_{\text{tangent}}(x;\theta) \\ &+ \alpha_2\mathbb{E}_{x\in\mathcal{D}}\mathcal{R}_{\text{normal}}(x;\theta) \\ &+ \alpha_3\mathbb{E}_{x\in\mathcal{D}}\mathcal{R}_{\text{entropy}}(x;\theta). \end{aligned} \tag{14}$$

Two-rings artificial dataset

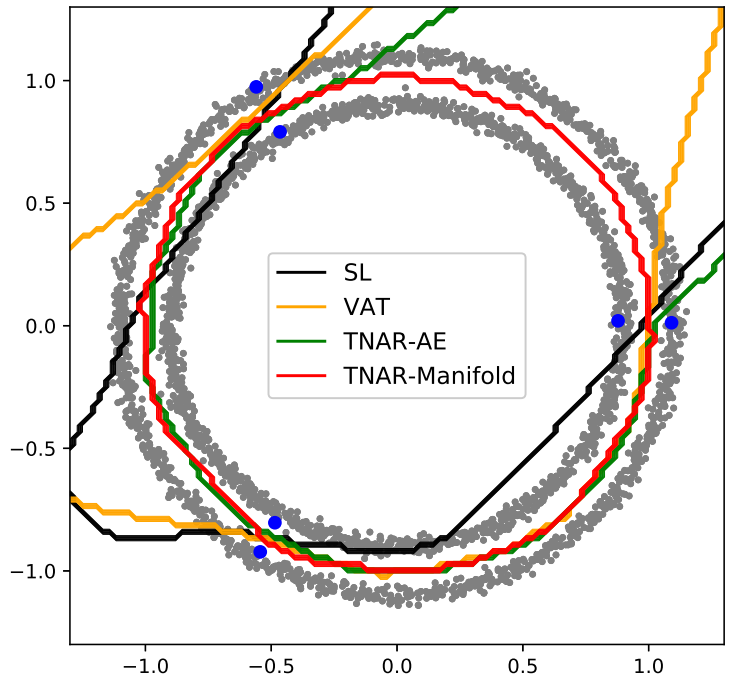

Figure: The decision boundaries of compared methods on two-rings artificial dataset. Gray dots distributed on two rings: the unlabeled data. Blue dots (3 in each ring): the labeled data. Colored curves: the decision boundaries found by compared methods.

SVHN and CIFAR-10 (without data augmentation)

Table: 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	14.87
VAT (large)	4.28	13.15
VAT + SNTG	4.02	12.49
Π model	5.43	16.55
Mean Teacher	5.21	17.74
CCLP	5.69	18.57
ALI	7.41	17.99
Improved GAN	8.11	18.63
Tripple GAN	5.77	16.99
Bad GAN	4.25	14.41
LGAN	4.73	14.23
Improved GAN + JacobRegu + tangent	4.39	16.20
Improved GAN + ManiReg	4.51	14.45
TNAR-LGAN (small)	4.25	12.97
TNAR-LGAN (large)	4.03	12.76
TNAR-VAE (small)	3.99	12.39
TNAR-VAE (large)	**3.80**	**12.06**
TAR-VAE (large)	5.62	13.87
NAR-VAE (large)	4.05	15.91

SVHN and CIFAR-10 (with data augmentation)

Table: 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	10.55
VAT + SNTG	3.83	9.89
Π model	4.82	12.36
Temporal ensembling	4.42	12.16
Mean Teacher	3.95	12.31
LGAN	-	9.77
TNAR-VAE (large)	**3.74**	**8.85**

Thanks!