Understanding deeply and improving VAT

Dongha Kim and Yongchan Choi
Speaker : Dongha Kim

Department of Statistics, Seoul National University, South Korea

July 5, 2018

1. Introduction
2. Related works
3. Explanations of VAT objective function
4. New methodologies improving VAT
5. Experiments

1. Introduction
2. Related works
3. Explanations of VAT objective function
4. New methodologies improving VAT
5. Experiments

Introduction

- Deep learning suffers from the **lack of labels**, since labeling is proceeded manually which results in a lot of expenditure in both money and time.
- Many researches have been proposed to deal with the lack of labels exploiting unlabeled data as well as labeled data to learn a optimal classifier (Weston et al., 2012; Rasmus et al., 2015; Kingma et al., 2014).
- Recently, two powerful methods have been proposed, one is called **VAT** method(Miyato et al., 2015, 2017) and the other is called **bad GAN** method(Dai et al., 2017).

Introduction

- VAT is efficient and powerful method, but its learning procedure is rather unstable and it is still not clear **why the VAT method also works well in semi-supervised case.**
- The method using bad GAN has clear principle and state-of-art prediction power, but **it needs additional architectures** which leads to heavy computational costs. So, it is infeasible to apply this to very large dataset.

Our contributions

- We give a **clear explanation why VAT works well in semi-supervised learning.**
- Based on our findings, we propose some simple and powerful **techniques to improve VAT.**
    - Especially we adopt the main idea of bad GAN which generates bad samples using bad generator, and apply this idea to VAT without any additional architectures.
- By using these methods, we can **achieve superior results** than other approaches, especially VAT, in both prediction power and efficiency aspects.

Adversarial training (AT, Goodfellow et al. (2014))

✓ **Smooth the model by using adversarial perturbations.**

- $p(\cdot|x;\theta)$: a conditional distribution of deep architecture parametrized by θ.
- Regularization term is the following:

$$L^{AT}(\theta;x,y,\epsilon) = KL\left[h(y), p(\cdot|x+r_{\text{advr}};\theta)\right]$$
$$\text{where } r_{\text{advr}} = \underset{r;||r||_2\leq\epsilon}{\operatorname{argmax}} KL\left[h(y), p(\cdot|x+r;\theta)\right]$$

where $h(y)$ is a one hot vector of y whose entries are all 0 except for the index corresponding to label y.

- The final objective function of AT is as follows:

$$\mathbb{E}_{(x,y)\sim\mathcal{L}^{tr}}\left[-\log p(y|x;\theta)\right] + \mathbb{E}_{(x,y)\sim\mathcal{L}^{tr}}\left[L^{AT}(\theta;x,y,\epsilon)\right]$$

where $\mathcal{L}^{tr}$ is labeled data and $\epsilon > 0$ is a hyperparameter.

Virtual adversarial training (VAT, Miyato et al. (2017))

✓ **VAT succeeds the key idea of AT.**

- VAT just **substitutes** $h(y)$ **by** $p(\cdot|x;\hat{\theta}^{cur})$ and this substitution allows VAT to be applicable to semi-supervised case.
- Regularization term of VAT is the following:

$$L^{VAT}(\theta;\hat{\theta}^{cur},x,\epsilon) = KL\left[p(\cdot|x;\hat{\theta}^{cur}), p(\cdot|x+r_{\text{advr}};\theta)\right]$$

$$\text{where } r_{\text{advr}} = \underset{r;||r||_2\leq\epsilon}{\text{argmax}}\, KL\left[p(\cdot|x;\hat{\theta}^{cur}), p(\cdot|x+r;\theta)\right]$$

where $\hat{\theta}^{cur}$ is current estimated parameters which is treated as constant and $p(\cdot|x;\hat{\theta}^{cur})$ is current conditional distribution.

- The final objective function of VAT is as follows:

$$\mathbb{E}_{(x,y)\sim\mathcal{L}^{tr}}\left[-\log p(y|x;\theta)\right] + \mathbb{E}_{x\sim\mathcal{U}^{tr}}\left[L^{VAT}(\theta;\hat{\theta}^{cur},x,\epsilon)\right]$$

where $\mathcal{U}^{tr}$ is unlabeled data.

Virtual adversarial training (VAT, Miyato et al. (2017))

Remark

- Note that $p(\cdot|x;\hat{\theta}^{cur})$ is a constant vector, thus we can rewrite the regularization term as follows:

$$L^{VAT}(\theta;\hat{\theta}^{cur},x,\epsilon) = -\sum_{k=1}^{K}\left[p(k|x;\hat{\theta}^{cur})\log p(k|x+r_{\text{advr}};\theta)\right] + C,$$

which is equal to **cross-entropy** term between $p(\cdot|x;\hat{\theta}^{cur})$ and $p(\cdot|x;\theta)$.

bad GAN approach (Dai et al., 2017)

- Bad GAN approach is a method that **trains a good discriminator with a bad generator** which generates samples over the support with low density.
- This approach trains a generator $p(\cdot|x;\theta)$ and a bad generator $p_G(\cdot|\eta)$ simultaneously with their own objective functions.
- To train $p_G(\cdot|\eta)$, we need a pre-trained density estimation model, for instance PIXELCNN++ (Salimans et al., 2017).
- To train the discriminator, we consider K-class classification problem as $(K+1)$-class classification problem where $(K+1)$-th class is an artificial label of bad samples generated by bad generator.

bad GAN approach (Dai et al., 2017)

- The objective function of discriminator is as follows:

$$\mathbb{E}_{x,y\sim\mathcal{L}^{tr}}\left[-\log p(y|x;\theta,y\leq K)\right]+\mathbb{E}_{x\sim\mathcal{U}^{tr}}\left[-\log\left\{\sum_{k=1}^{K}p(k|x;\theta)\right\}\right]$$
$$+\mathbb{E}_{x\sim\mathcal{G}(\hat{\eta}^{cur})}\left[-\log p(K+1|x;\theta)\right]$$

where $\mathcal{G}(\hat{\eta}^{cur})$ is data generated by currently estimated generator $p_G(\cdot|\hat{\eta}^{cur})$.

1. Introduction

2. Related works

3. Explanations of VAT objective function

4. New methodologies improving VAT

5. Experiments

Notations

- $\mathcal{L}^{tr} = \{(x_i^l, y_i)\}_{i=1}^n$: labeled data ($x \in \mathbb{R}^p$ and $y \in \{1, ..., K\}$).
- $\mathcal{U}^{tr} = \{x_j^u\}_{j=1}^m$: unlabeled data.
- $y(x)$: ground-truth label of an input x. (of course, $y(x_i^l) = y_i$.)
- We can partition unlabeled data as following:

$$\mathcal{U}^{tr} = \cup_{k=1}^K \mathcal{U}_k^{tr}$$

where $\mathcal{U}_k^{tr} = \{x : x \in \mathcal{U}^{tr}, y(x) = k\}$.

Definition 1.

We define a tuple $(x, x^{'})$ is ϵ-*connected iff* $d(x, x^{'}) < \epsilon$, where $d(\cdot, \cdot)$ is Euclidean distance. And a set $\mathcal{X}$ is called ϵ-*connected iff* for all $x, x^{'} \in \mathcal{X}$, there exists a path $(x, x_1, ..., x_q, x^{'})$ such that $(x, x_1), (x_1, x_2), ..., (x_{q-1}, x_q), (x_q, x^{'})$ are all ϵ-*connected*.

Notations

- With definition 1, we can partition $\mathcal{U}_k^{tr}$ as disjoint union of clusters as following:

$$\mathcal{U}_k^{tr} = \cup_{l=1}^{n(\epsilon,k)} \mathcal{U}_{k,l}^{tr}(\epsilon)$$

where $\mathcal{U}_{k,l}^{tr}(\epsilon)$ is ϵ-*connected* for all l, $d(\mathcal{U}_{k,l}^{tr}(\epsilon), \mathcal{U}_{k,l'}^{tr}(\epsilon)) = \min_{x\in\mathcal{U}_{k,l}^{tr}, x'\in\mathcal{U}_{k,l'}^{tr}} d(x, x^{'}) \geq \epsilon$ for all $l \neq l^{'}$, and $n(\epsilon, k)$ is the number of clusters of $\mathcal{U}_k^{tr}$.

Main theorem

Main theorem

Let assume there exists $\epsilon > 0$ s.t.

1. $d(\mathcal{U}^{tr}_{k,l}(\epsilon), \mathcal{U}^{tr}_{k',l'}(\epsilon)) \geq 2\epsilon$ for all $k \neq k'$,
2. For all $\mathcal{U}^{tr}_{k,l}(\epsilon)$, there exist at least one $(x, y) \in \mathcal{L}^{tr}$ which have the same label s.t. $d(x, \mathcal{U}^{tr}_{k,l}) < \epsilon$.

And also let assume that there exists a classifier $f : \mathbb{R}^p \rightarrow \{1, ..., K\}$ s.t.

3. $f(x) = y$ for all $(x, y) \in \mathcal{L}^{tr}$ and $f(x) = f(x')$ for all $x' \in \mathcal{B}(x, \epsilon), x \in \mathcal{U}^{tr}$.

Then, the f classify the unlabeled set perfectly, that is:

$$f(x) = y(x) \text{ for all } x \in \mathcal{U}^{tr}.$$

Derivation of VAT loss function

- Let $f(x;\theta) = \underset{k=1,\ldots,K}{\operatorname{argmax}}\, p(k|x;\theta)$.
- We focus to find optimal θ satisfying the condition 3 in main theorem by using a suitable objective function.
- The most plausible candidate may be using indicator function:

$$\begin{aligned} &\mathbb{E}_{(x,y)\sim\mathcal{L}^{tr}}\left[I(f(x;\theta) \neq y)\right] \\ &+\mathbb{E}_{x\sim\mathcal{U}^{tr}}\left[I\left(f(x;\theta) \neq f(x^{'};\theta) \text{ for } \forall x^{'} \in \mathcal{B}(x,\epsilon)\right)\right] \quad (1) \end{aligned}$$

- $\hat{\theta}$ achieves 0 value $\iff$ $f(\cdot;\hat{\theta})$ satisfies the condition 3.
- Two problems to minimize the objective function (1):
  1. The indicator function is impossible to be optimized because of discontinuity.
  2. It is infeasible to search all $x^{'} \in \mathcal{B}(x,\epsilon)$ in order to calculate the second term of (1).

Derivation of VAT loss function

- To deal with these above problems,
  1. we exploit the **differentiable surrogate function**, which is called **cross-entropy**,
  2. and we **only search the most adversarial neighborhood**, which increases the cross entropy most rapidly, of each x in unlabeled set.

3. If we **replace** $p(\cdot|x;\theta)$ **to** $p(\cdot|x;\hat{\theta}^{cur})$, then modified version of (1) finally become the exactly same formula as that of VAT:

$$\mathbb{E}_{x,y\sim\mathcal{L}^{tr}}\left[-\log p(y|x;\theta)\right]+\mathbb{E}_{x\sim\mathcal{U}^{tr}}\left[-\sum_{k=1}^{K}p(k|x;\hat{\theta}^{cur})\log p(k|x+r_{\text{advr}};\theta)\right]$$

$$\text{where } r_{\text{advr}}=\underset{r;||r||_2\leq\epsilon}{\text{argmax}}\left[-\sum_{k=1}^{K}p(k|x;\hat{\theta}^{cur})\log p(k|x+r;\theta)\right]$$

Interpretation of VAT loss function

- Using the objective function (1), we can interpret and investigate the role of regularization term of VAT.
- Being positive value of second term of (1) means that there exists a cluster $\mathcal{U}_{k,l}^{tr}$ which is divided into at least two regions by current decision boundary.
- The only way to to minimize the above term is to prevent decision boundary to cut across inside of the cluster.
- Therefore, we may conclude that minimizing VAT regularization term is to **push decision boundary away from the inside of all clusters.**

1. Introduction
2. Related works
3. Explanations of VAT objective function
4. **New methodologies improving VAT**
5. Experiments

Usage of virtual labels

- Note that the primary purpose of the second term of (1) is to get x and $x + r$ to have **equal predicted label, not conditional distribution.**
- So, the regularization term of VAT which leads to having almost identical conditional distribution between x and x + r seems to include some superfluous calculations.
- We modify this regularization term by using virtual label, not conditional distribution.

Proposed regularization term

- Then our modified version is as follows:

$$L^{\text{mod}}(\theta; \hat{\theta}, x, \epsilon) = -\sum_{y=1}^{K} I\left(y(x;\hat{\theta}) = k\right) \log p(k|x + r_{\text{advr}}; \theta), \qquad (2)$$

where $y(x;\hat{\theta}) = \text{argmax}_k p(k|x;\hat{\theta})$.

Generation of adversarial data without any generator

- We adopt **the role of adversarial data** in bad GAN and generate these directly **using only discriminator.**
- The crucial property is that **the optimal** r which maximize the KL divergence in VAT **is towards decision boundary**.
- By the above property, we can choose a suitable ϵ such that the perturbed input $x + r$ exists in the support with low density.

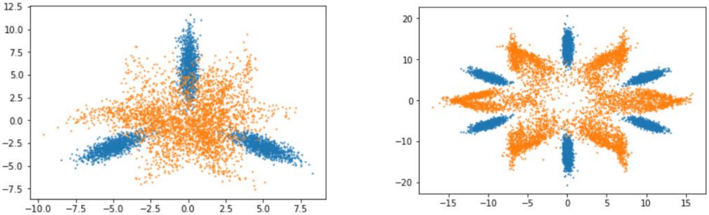


Figure : Generated data using KL divergence with suitable ϵ in 3-class and 6-class classification problems. True data are colored to blue and fake data are colored to orange.

Generation of adversarial data without any generator

Proposed regularization term

- The additional regularization term we propose newly is as follows:

$$Adv(\theta;\hat{\theta}^{cur},x,\epsilon) := -\log\frac{1}{1+\sum_{k=1}^{K}\exp\left\{f_k\left(x+r_{\text{advr}};\theta\right)\right\}} \tag{3}$$

$$\text{where } r_{\text{advr}} = \operatorname*{argmax}_{r;||r||_2\leq\epsilon}\sum_{k=1}^{K}p(k|x;\hat{\theta}^{cur})\log p(k|x+r;\theta)$$

  where $f_k(\cdot;\theta)$ is k-th output before softmax.

- Minimizing the above term enforce **decision boundary to be pulled from the support with high density to low density.**

Final objective function

- Combining two newly proposed methods (2) and (3), we achieve the final objective function as follows:

$$\begin{aligned}&\mathbb{E}_{x,y\sim\mathcal{L}^{tr}}\left[-\log p(y|x;\theta)\right]\\&+\mathbb{E}_{x\sim\mathcal{U}^{tr}}\left[L^{\text{mod}}(\theta;\hat{\theta},x,\epsilon_1)\right]+\mathbb{E}_{x\sim\mathcal{U}^{tr}}\left[Adv(\theta;\hat{\theta},x,\epsilon_2)\right]\quad(4)\end{aligned}$$

  where $\epsilon_2 > \epsilon_1 > 0$ are hyperparameters.
- We expect that our proposed objective function can obtain a dicriminator superior to that learned by VAT, and further, accelerate training step.

1 Introduction

2 Related works

3 Explanations of VAT objective function

4 New methodologies improving VAT

5 Experiments

Prediction accuracy

Synthetic data case

- We generate 1000 unlabeled data (gray) and 4 labeled data for each class (red and blu with black edge).
- Two discriminators which are 2-layered NN with 100 hidden units each are trained by our method and VAT respectively.
- Our best model achieves **99.9%** accuracy while the best VAT achieves **96.1%**.

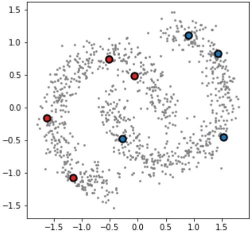


Figure : Scatter plot of synthetic data

Prediction accuracy

Benchmark data case

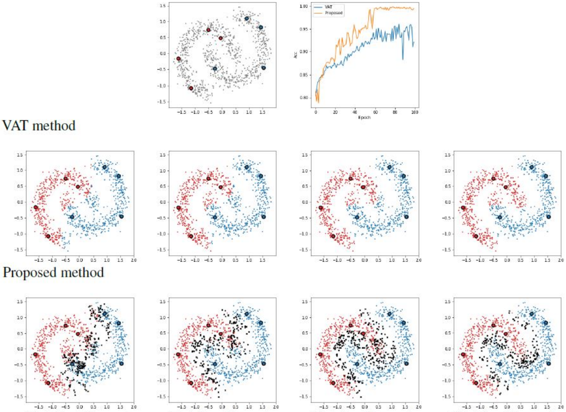
- We randomly sample 100, 1000, 4000 labeled data for MNIST, SVHN, CIFAR10 respectively and use them as labeled data, while the rest are used as unlabeled data.

Method	Test acc.(%)		
	MNIST	SVHN	CIFAR10
DGN (Kingma et al., 2014)	96.67	63.98	-
Ladder (Rasmus et al., 2015)	98.94	-	79.6
GAN with FM (Salimans et al., 2016)	99.07	91.89	81.37
Bad GAN (Dai et al., 2017)	99.20	95.75	85.59
VAT(paper) (Miyato et al., 2017)	98.64	93.17	85.13
VAT(our code)	98.55	93.6	84.19
Proposed	98.74	94.03	84.69

Table : Test performances on three benchmark datasets.

Effects of generated adversarial data

Synthetic data case



VAT method

Proposed method

(1) After 20 epochs (2) After 40 epochs (3) After 60 epochs (4) After 80 epochs

Figure : Learning procedure at each steps. Our method substantially improves the convergence speed as well as prediction power. Besides, prediction accuracies of ours are stable while those of VAT are tend to oscillate.

Effects of generated adversarial data

Benchmark data case (MNIST)

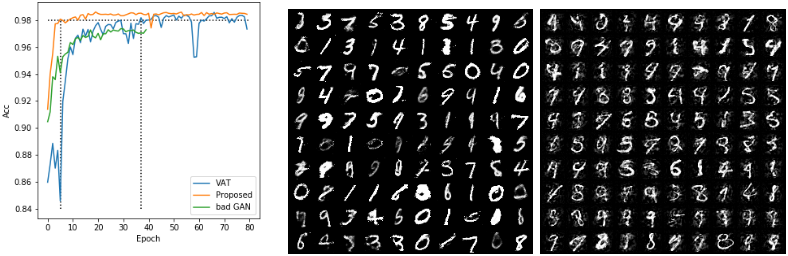


Figure : **(Left)** Test accuracy of each epoch for tree methods(ours, VAT and bad GAN). Our method achieves the same accuracy with 6 times fewer training steps and beat the best performance of VAT. **(Middle)** Adversarial images using bad GAN. Bad generator still generates realistic images. **(Right)** Adversarial images using our method. As can be seen, our method consistently generates diverse 'bad' samples.

References

Dai, Z., Yang, Z., Yang, F., Cohen, W. W., and Salakhutdinov, R. R. (2017). Good semi-supervised learning that requires a bad gan. In *Advances in Neural Information Processing Systems*, pages 6513–6523.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2017). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *arXiv preprint arXiv:1704.03976*.

Miyato, T., Maeda, S.-i., Koyama, M., Nakae, K., and Ishii, S. (2015). Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. (2015). Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554.