互信息指导下的贝叶斯对抗训练

王一诺
2018011388
计85
wangyinu18@mails.tsinghua.edu.cn

郑凯文
2018011314
计82
zhengkw18@mails.tsinghua.edu.cn

张庶杰
2018011415
计85
sj-zhang18@mails.tsinghua.edu.cn

1 简介

在分类任务中，对抗训练[1]是一种提高模型鲁棒性和泛化性的方法，即通过计算特定方向的扰动生成对抗样本，再使用对抗样本进行训练。[2]提出了从无标签样本生成虚拟对抗样本的思想，并使用扰动前后预测分布的KL散度作为正则化项。

正则化项可以为模型引入额外的信息，提高模型泛化性能。由于虚拟对抗训练不依赖于标签，其可以作为一种半监督学习的方式，利用无标签数据提供正则化信息。

半监督学习大大降低了标注成本，且虚拟对抗训练的方式提供了一定的局部光滑性，使分类面变得更加鲁棒。但其存在一定问题，如[3]指出，扰动前后预测分布的KL散度并不能区分数据不确定性（如一个既像1又像7的图像）和模型不确定性。数据的不确定性是固有的，即使人也难以分辨，因此不应强行将其归于某一类。沿袭[3]的工作，我们将模型转化为贝叶斯网络后，针对互信息进行攻击，利用得到的对抗样本对模型进行与[2]类似的正则化。

我们的贡献包括：

1. 提出了一种贝叶斯对抗训练的方法，可以作为半监督学习方法使用。
2. 在小型数据集上定性探究了针对互信息攻击的效果。
3. 在MNIST上进行了充分的消融实验，证实了本方法可以达到与虚拟对抗训练相似的准确率，且可在已知的半监督学习方法基础上进一步提高性能。
4. 以误分类率为指标，在对抗攻击、虚拟对抗攻击与互信息攻击三种攻击手段下，对鲁棒性进行定量探究，证实了本方法可显著提高模型鲁棒性。

2 相关工作

2.1 虚拟对抗训练

虚拟对抗训练(VAT)为[2]中提出的一种半监督学习方式。通过寻找扰动使得扰动前后预测分布的KL散度最大，再最小化此扰动下扰动前后的KL散度，使得模型学习到一个鲁棒的分类面，在样本点的周围的概率分布保持一定的一致性。相对于添加随机扰动作为对抗样本的训练，VAT能够找到变化最大的方向进行优化，从而获得更高的准确率与收敛速度。具体算法如1所示。

Algorithm 1 虚拟对抗训练

输入: 有标签数据$(x_i^l, y_i) \in D^l$，无标签数据$x_i^{ul} \in D^{ul}$，极小正数ξ，超参数ϵ

Step1: 计算有标签损失

$$\mathcal{L}_l = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i|x_i^l, \theta) \tag{1}$$

Step2: 用标准高斯分布生成与x_i^{ul}维度相同的随机向量并进行单位化，得到$d^{(i)}$.

Step3: 计算攻击方向r_{adv}

$$g^{(i)} = \nabla_r D_{KL}[p(y|x_i^{ul}, \theta), p(y|x_i^{ul} + r, \theta)]|_{r=\xi d^{(i)}} \tag{2}$$

$$r_{adv}^{(i)} = g^{(i)}/||g^{(i)}||_2 \tag{3}$$

Step4: 计算无标签损失

$$\mathcal{L}_{ul} = \frac{1}{M}\sum_{i=1}^{M} D_{KL}[p(y|x_i^{ul}, \theta), p(y|x_i^{ul} + \epsilon r_{adv}^{(i)}, \theta)] \tag{4}$$

Step5: 使用梯度下降法最小化损失

$$\mathcal{L} = \mathcal{L}_l + \mathcal{L}_{ul} \tag{5}$$

2.2 变分贝叶斯网络

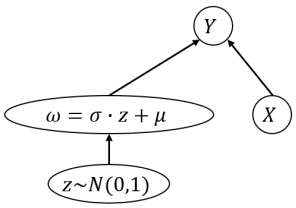


Figure 1: 贝叶斯网络

我们在此次任务中需要完成确定性模型和变分贝叶斯网络的转换。设模型参数为ω（贝叶斯学派假设其服从一定分布），变分贝叶斯网络采用变分分布$q(\omega|\theta) \sim \mathcal{N}(\mu, \sigma^2)$近似真实的后验分布$p(\omega|D)$，即$\theta = \{\mu, \sigma\}$。为了处理参数$\omega$抽样无法追踪优化的问题，采用重参数化的手段，如图1所示，此时即可对似然$\log p(y|x)$的ELBO进行优化:

$$\max_{\theta} \mathbb{E}_{q(\boldsymbol{w}|\theta)}[\frac{1}{n}\sum_{i} \log p(y_i|x_i; \boldsymbol{w})] - \frac{1}{n} D_{KL}(q(\boldsymbol{w}|\theta)||p(\boldsymbol{w})) \tag{6}$$

对于其中的KL散度项，在变分分布为对角正态分布时可以计算出它的表达式如下:

$$\mathcal{L}_c = -\frac{1}{n} D_{KL}(\mathcal{N}(\boldsymbol{w}; \mu, \Sigma)||\mathcal{N}(\boldsymbol{w}; 0, \sigma_0^2 \boldsymbol{I})) \tag{7}$$

$$= -\frac{\boldsymbol{\mu}^T \boldsymbol{\mu} + tr(\boldsymbol{\Sigma})}{2\sigma_0^2 n} + \frac{\log\det(\boldsymbol{\Sigma})}{2n} + c \tag{8}$$

进一步得到它的梯度为:

$$\nabla_{\boldsymbol{\mu}} \mathcal{L}_c = -\frac{\boldsymbol{\mu}}{\sigma_0^2 n} \tag{9}$$

$$\nabla_{\boldsymbol{\Sigma}} \mathcal{L}_c = \frac{\sigma_0^2 \boldsymbol{\Sigma}^{-1} - \boldsymbol{I}}{2\sigma_0^2 n} \tag{10}$$

设$\lambda = 1/(\sigma_0^2 n), \mathbf{\Sigma} = diag(\exp(2\boldsymbol{\psi}))$，则$\nabla_{\boldsymbol{\mu}} \mathcal{L}_c = -\lambda \boldsymbol{\mu}$可以通过权重衰减实现，$\nabla_{\boldsymbol{\psi}} \mathcal{L}_c = 1/n - \lambda \exp(2\boldsymbol{\psi})$可以通过修改优化器实现[4]。

对于确定性模型向贝叶斯网络的转换，将预训练参数作为均值$\boldsymbol{\mu}$后随机初始化$\boldsymbol{\psi}$；对于贝叶斯网络向确定性模型的转换，将均值$\boldsymbol{\mu}$作为确定性的网络参数。变分贝叶斯网络学习了模型参数后验分布$p(\omega|\mathcal{D})$的近似分布，可以自然的与互信息搭配。在实际中我们通过蒙特卡洛采样的形式近似计算模型的互信息：

$$p(y|\mathcal{D}, x) \simeq \frac{1}{T} \sum_{i=1}^{T} p(y|\omega_i, x) := p_{MC}(y|x) \tag{11}$$

$$I(\omega, y|\mathcal{D}, x) = H[p_{MC}(y|\mathcal{D}, x)] - \frac{1}{T} \sum_{i=1}^{T} H[p(y|\omega_i, x)] \tag{12}$$

2.3 Π模型

Π模型[5]是一种简单的半监督学习方法。

Algorithm 2 Π 模型

输入: 有标签数据$(x_i^l, y_i) \in D^l$，无标签数据$x_i^{ul} \in D^{ul}$，权重函数$w(t)$，随机的、有可训练参数θ的神经网络$f_\theta(x)$，类别个数C

Step1: 将无标签样本添加高斯噪声后先后两次通过网络，得到不同输出

$$z_i = f_\theta(x_i^{ul}), \quad \tilde{z}_i = f_\theta(x_i^{ul}) \tag{13}$$

Step2: 使用梯度下降法最小化损失

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i|x_i^l, \theta) + w(t) \frac{1}{CM} \sum_{i=1}^{M} ||z_i - \tilde{z}_i||^2 \tag{14}$$

Π模型需要输出具有一定随机性的网络（如具有Dropout层）。在训练中，对输入的数据添加一个微小的高斯噪声之后通过Π模型得到预测结果，对这个过程重复多次，通过优化分类损失与多次预测之间的方差进行半监督学习。从本质上看，Π模型通过内在的随机性对输入和模型的计算过程施加扰动，在微小扰动下的预测一致性描述了模型关于参数空间和样本空间的光滑性，将这种光滑性作为正则项能够为在有限样本的情况下提供更多的信息，从而达到半监督学习的目的。

2.4 互信息与对抗样本检测

互信息(MI)是描述随机变量X与随机变量Y之间的关系的数学工具，它的表达式为：

$$\begin{aligned} I(X, Y) &= H[P(X)] - \mathbb{E}_{P(y)} H[P(X|Y)] \\ &= H[P(Y)] - \mathbb{E}_{P(x)} H[P(Y|X)] \end{aligned} \tag{15}$$

互信息描述了得知随机变量Y的真实分布后对于随机变量X的信息的增益，可以用以描述随机变量X的不确定性。在本次实验中，Y对应数据的真实分布$p(y|x)$，X对应模型参数的参数ω，因此在本问题中模型的互信息表达式为：

$$I(\omega, y|\mathcal{D}|x) = H[p(y|x, \mathcal{D})] - \mathbb{E}_{p(\omega|\mathcal{D})} H[p(y|x, \omega)] \tag{16}$$

它描述了真实分布$p(y|x)$对于参数分布$p(\omega|\mathcal{D})$的信息增益。在实践中可以利用变分贝叶斯网络在参数空间中进行抽样，将$\sum_{i=0}^{n} p(y|\omega_i, x)$作为真实分布的近似，如式12所示

[3]中指出，不确定性分为两类，第一类不确定性为模型不确定性，第二类不确定性为数据不确定性。若数据本身在真实分类面周围，那么每一次参数空间抽样的ω都可以使得预测

熵$H(y|\omega,x)$较大，因而使得互信息变小；而对于某一个数据点由于模型本身缺乏足够的信息而无法作出判断，那么将会得到较高的互信息值。因此，互信息可以良好的描述模型不确定性。互信息在$p=1$处的泰勒展开式可以看作Softmax Variance的一种近似，推导如下：

$$\begin{aligned}\hat{I} &= H(p) - \frac{1}{T}\hat{\sum_i} H(p_i) \\ &= \sum_j (\frac{1}{T}\sum_i p_{ij} log p_{ij}) - \hat{p}_j log \hat{p}_j \\ &= \sum_j (\frac{1}{T}\sum_i p_{ij}(p_{ij}-1)) - \hat{p}_j(\hat{p}_j - 1) + ... \qquad (Taylor Expansion) \\ &= \sum_{j=1}^{C} (\frac{1}{T}\sum_{i=1}^{T} p_{ij}^2) - \hat{p}_j^2 + ... \\ &\simeq \frac{1}{C}(\sum_{j=1}^{C}(\frac{1}{T}\sum_{i=1}^{T} p_{ij}^2) - \hat{p}_j^2) \\ &= C\hat{\sigma}^2\end{aligned} \tag{17}$$

因此互信息同样可以看作对于参数抽样的预测一致性的正则项。

3 可行性与合理性的进一步分析

3.1 与Π模型的联系

在正常情况下，机器学习的目标为获得在数据集上损失函数最小化的一组参数ω^*：

$$\omega^* = \arg\min_{\omega \in p(\omega|\mathcal{D})} \mathcal{L}(Y|\omega, X) \tag{18}$$

而在Π模型中，加入了对预测一致性的正则项，若我们将神经网络在数据点处进行泰勒展开，我们可以将Π模型的输出近似表达为：

$$f(y|\omega, x+\delta_x) = f(y|\omega,x) + \frac{\partial f}{\partial x}\delta_x \tag{19}$$

正则项的优化目标为：

$$\min_{\omega} \text{norm}(\frac{\partial f}{\partial x})||\delta_x|| \tag{20}$$

它优化了数据点附近的关于输入变量x的光滑性。而正如前文指出的那样，互信息也可以看作预测一致性的度量，若参数ω的正态分布$\mathcal{N}(\mu,\sigma)$的方差较小，可以将每一次抽样的参数写为$\mu+\delta$，其中δ为服从正态分布$\mathcal{N}(0,\sigma)$的随机变量，因此它的正规项可以写为：

$$\min_{\mu} \text{norm}(\frac{\partial f}{\partial \omega})||\delta|| \tag{21}$$

它优化了位于数据点附近的关于参数ω的光滑性，而由于ELBO中KL散度的存在，它将会使得模型学习到一个较为简单，且对不同参数抽样满足预测一致性的分布。

3.2 互信息作为不确定性度量的进一步探究

我们在一个toy dataset上探究了互信息与预测熵分布的区别。如下为在二维平面上构造的4个从正态分布中抽样的簇，如图2所示。 我们分别用一个简单的确定型模型与一个变分贝叶斯模型在这个toy dataset上进行了训练，得到预测熵和互信息的分布如图3所示。 由图3b可见，预测熵在分类面上取得了较大的值，而图3a表明互信息在远离样本的方向取得了较大的值，反而在分类面内互信息较小。因此，如VAT等将KL散度作为不确定性的度量，虚拟样本的方向将指向决策面；而以互信息作为不确定性的度量，虚拟样本的方向将指向out of dataset的方向，它能够补全模型的信息，使得模型在各种输入下都能获得较为一致的结果。相对于VAT而言，它能够良好的区分模型不确定性与数据不确定性。若一个数据点本身就比较模糊，如在真实分类面上的数据，以互信息作为指导能够保持一定的不确定性。

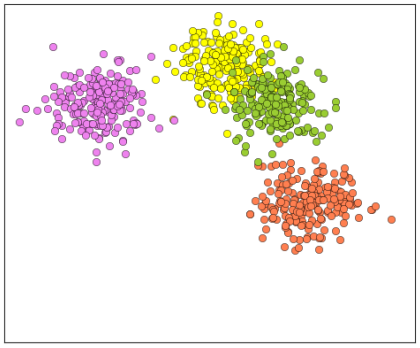
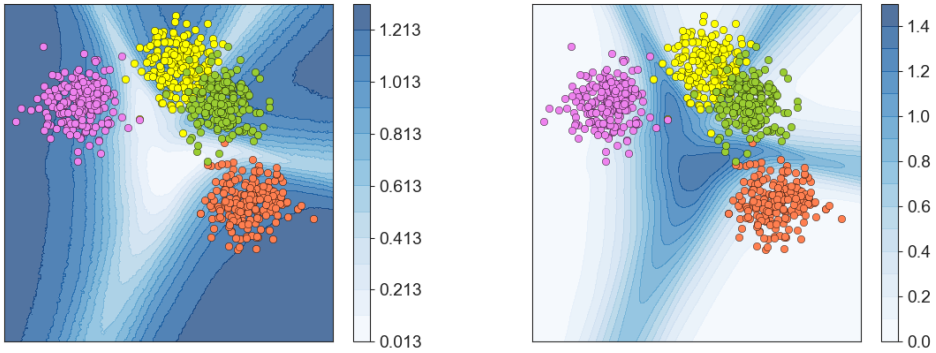


Figure 2: toy dataset



(a) 互信息分布　　(b) 预测熵分布

Figure 3: 两种不确定性度量的分布

4 方法

4.1 攻击方法

4.1.1 对抗攻击

对抗攻击的方法为在有标签数据(x_i, y_i)上直接寻找扰动使得分类概率$p(y_i|x_i, \omega)$最低，从而制造对抗攻击样本，具体过程如算法3所示。

4.1.2 虚拟对抗攻击

虚拟对抗攻击为在无标签数据x_i上直接直接寻找扰动使得扰动前后的分类分布的KL散度最大。寻找虚拟对抗样本的过程与算法1中完全一致。

4.1.3 互信息攻击

互信息攻击样本的寻找过程受到VAT样本的寻找过程的启发，在变分贝叶斯网络中对参数采样，多次计算最后的分类分布并以此计算互信息。在得到在样本点的互信息之后寻找扰动使得互信息最大，由此得到互信息攻击样本，具体算法与算法4中完全一致

Algorithm 3 对抗攻击

Step1: 从数据集$\mathcal{D}$中随机抽取M组样本$x^{(i)}, y^{(i)}(i = 1, 2, ...M)$.
Step2: 利用以下公式计算攻击方向$g^{(i)}$

$$g^{(i)} \leftarrow \nabla_r p(y^{(i)}|x^{(i)}, \omega) \tag{22}$$

Step3: 将$g^{(i)}$正规化并乘以系数ϵ得到最后的攻击样本扰动r_{adv}

$$r_{adv}^{(i)} \leftarrow -\epsilon g^{(i)}/||g^{(i)}||_2 \tag{23}$$

Step4: 返回最后的对抗样本$x^{(i)} + r_{adv}^{(i)}, y^{(i)}(i = 1, 2, ...M)$

4.2 训练方法

我们的基于互信息的半监督学习算法如下。首先在参数空间中对模型参数进行MC=S蒙特卡洛采样，利用取得的结果进行互信息的计算：

$$\mathcal{I}(\omega, y|\mathbf{x}, \mathcal{D}) = H(\frac{1}{S}\sum_{s=1}^{S} p(y|\mathbf{x}; \omega^{(s)})) - \frac{1}{S}\sum_{s=1}^{S} H(p(y|\mathbf{x}; \omega^{(s)})) \quad \text{where } \omega^{(s)} \sim q(\omega|\theta) \tag{24}$$

作为虚拟对抗样本攻击的对象，每一次攻击时，在有标注或无标注样本上寻找$x + \delta$ 使得$\mathcal{I}(\omega, y|\mathbf{x} + \delta, \mathcal{D})$能够最大：

$$\mathcal{L}_{unc} = \max_{|\delta| \leq \epsilon} \mathcal{I}(\omega, y|\mathbf{x} + \delta, \mathcal{D}) \tag{25}$$

由上，将有标签数据集记为$\mathcal{D}_l$，无标签数据集记为$\mathcal{D}_{ul}$，可将目标优化表达式写为：

$$\max_{\theta} \mathcal{L}_{ELBO}(\mathcal{D}_l, \theta) - \alpha \mathcal{L}_{unc}(\mathcal{D}_l, \mathcal{D}_{ul}, \theta) \tag{26}$$

通过优化此表达式得到最后的结果。但在实验中我们发现这样的效果并不理想，它在MNIST上的1000个有标注样本的训练中相对于baseline提升并不明显。分析问题的原因，我们发现在扰动之后的分类的概率分布相对于扰动之前已经发生了巨大的变化，此时仅仅降低扰动之后的互信息值并不能保证扰动前后分布的一致性，因此我们学习VAT的思想加入KL散度作为扰动前后距离的度量，因此最终无监督的损失函数$\mathcal{L}_{unc}$改写为表达式为：

$$\mathcal{L}_{unc} = \max_{|\delta| \leq \epsilon} \mathcal{I}(\omega, y|\mathbf{x} + \delta, \mathcal{D}) - KL(p(y|\mathbf{x} + \delta)||p(y|\mathbf{x})) \tag{27}$$

由此得到了最后的算法设计如下：

Algorithm 4 互信息指导的贝叶斯对抗训练

Step1: 从数据集$\mathcal{D}$中随机抽取M组样本$x^{(i)}(i = 1, 2, ...M)$.
Step2: 在样本上通过对参数空间蒙特卡洛采样计算互信息：

$$\mathcal{I}(\omega, y|\mathbf{x}) = H(\frac{1}{S}\sum_{s=1}^{S} p(y|\mathbf{x}; \omega^{(s)})) - \frac{1}{S}\sum_{s=1}^{S} H(p(y|\mathbf{x}; \omega^{(s)})) \quad \text{where } \omega^{(s)} \sim q(\omega|\theta)$$

Step3: 利用以下公式计算攻击方向r_{adv}

$$g \leftarrow \nabla_x \mathcal{I}(\omega, y|\mathbf{x}) \tag{28}$$

$$r_{adv} \leftarrow \epsilon g/||g||_2 \tag{29}$$

Step4: 计算扰动点的互信息与扰动前后的KL散度，返回梯度并利用梯度下降进行训练

$$\nabla_\theta(\mathcal{I}(\omega, y|\mathbf{x} + r_{adv}) - KL(p(y|\mathbf{x} + r_{adv})||p(y|\mathbf{x}))) \tag{30}$$

5 实验

5.1 实验设置

实验采用数据集MNIST，包含有70000张28*28的手写数字图像，其中训练集大小为60000，测试集大小为10000。我们抽取训练集中1000个样本作为验证集，剩余共59000个样本作为有标注和无标注训练集。模型为(Linear-BatchNorm-ReLU)*n-Linear-BatchNorm形式的全连接网络，各层大小为(784, 1200, 600, 300, 150, 10)。

我们选取了全标注、标注数为1000和100作为有监督和半监督学习的参数，其中100标注可以作为少次学习（Few-Shot Learning）的典型。有标注数据和无标注数据采用独立的batch，无标注数据为所有的训练样本，有标注数据从10个类别中均匀选择。在所有实验中，有标注的batch大小为100，无标注的batch大小为256。优化器选用Adam，学习率为0.001，均不采用诸如学习率衰减、warm up等技巧。

我们共选取5种方法进行实验：Baseline为以预测值和标签的交叉熵作为损失的有监督模型；Π模型和VAT如前述算法所示；Baseline+MI和VAT+MI分别将训练好的Baseline和VAT作为预训练模型，转化为贝叶斯网络后按照前述算法训练。

VAT和MI对抗训练中的超参数ϵ在区间[0.1,20]内进行粗粒度的网格搜索（见附录：原始数据）。对于Baseline、Π模型和VAT，训练持续100000步，每1000步在验证集上进行一次评估。对于MI对抗训练，共进行10000步，每100步进行一次评估。验证集准确率最高的模型将被保存并用于测试。

5.2 实验结果

方法	标注数	准确率	标注数	准确率	标注数	准确率
Baseline	59000	98.66%	1000	92.56%	100	75.03%
Π Model	59000	98.47%	1000	94.48%	100	84.67%
VAT	59000	99.19%	1000	98.71%	100	98.60%
Baseline+MI	59000	99.12%	1000	98.70%	100	98.22%
VAT+MI	59000	**99.26**%	1000	**98.72**%	100	**98.78**%

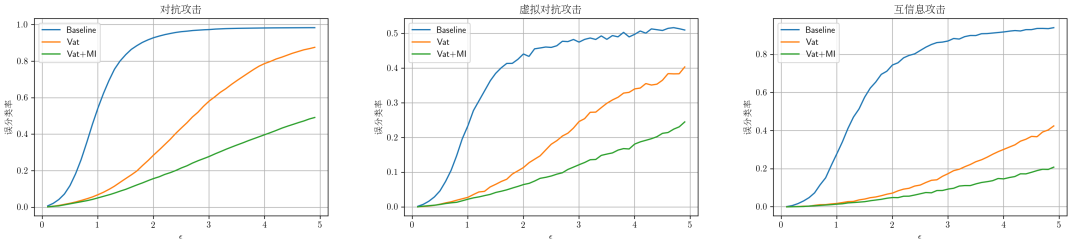
Table 1: MNIST数据集上的实验结果

表4展示了各种方法的最优准确率。随着标注数的下降，有监督模型的性能急剧下降，而VAT或Baseline+MI均能维持98%以上的准确率，这证明了MI对抗训练是一种行之有效的半监督学习方法。在VAT预训练模型的基础上，辅以MI对抗训练，可进一步提高模型的泛化性能。

为展示MI对抗训练对模型鲁棒性的影响，我们在全标注训练的最优Baseline、VAT和VAT+MI上测试三种攻击下误分类率随攻击半径ϵ的变化曲线。对抗攻击和虚拟对抗攻击针对确定性的模型，而MI对抗攻击针对贝叶斯网络，因此前者需要将贝叶斯网络VAT+MI转化为确定性模型，后者需要将确定性模型Baseline、VAT转化为贝叶斯网络。

图4展示了$\epsilon \in [0.1, 50)$时的误分类率曲线。从中可以得出以下结论：

- Baseline由于未利用对抗样本进行正则化，在三种攻击下均不堪一击，在ϵ很小时误分类率便急剧上升。
- VAT虽然只利用了虚拟对抗样本，但其保证了一部分局部光滑性，即对于微小扰动的稳定性，因此即使没有进行对抗训练和MI对抗训练，对小ϵ下的三种攻击方式均表现出抗性，远强于Baseline。
- MI对抗训练进一步提升了VAT的鲁棒性。在$\epsilon < 1$时，VAT误分类率仅微小高于VAT+MI；而随着ϵ的增大，VAT+MI显著地维持了远低于VAT的误分类率。这表明VAT+MI甚至可以对一定程度的噪声鲁棒。



(a) 对抗攻击　　(b) 虚拟对抗攻击　　(c) 互信息攻击

Figure 4: 误分类率随攻击半径ϵ的变化

6 总结

结合学长的工作[4]，我们沿袭[3]对互信息在对抗样本上应用的讨论，在MNIST进行了充分的实验，并验证了互信息指导下贝叶斯对抗训练在提高泛化性能和鲁棒性上的有效性。在设计完算法后我们进行了额外的探究，在toy dataset上进一步探究了互信息的特性，并对它的表现作出一些可能的解释。由于训练成本较大，我们没有多余时间在较大数据集上进行尝试，因此CIFAR10或更大数据集上的效果还需进一步验证。

附录：原始数据

ϵ	Baseline	VAT	Baseline+MI	VAT+MI
0	98.66%	-	-	-
0.1	-	98.87%	98.74%	99.14%
0.3	-	98.90%	98.87%	99.09%
0.5	-	**99.19**%	98.93%	99.08%
1.0	-	99.04%	98.98%	99.20%
2.0	-	99.16%	98.99%	**99.26**%
3.0	-	99.10%	**99.12**%	99.21%
5.0	-	99.14%	99.11%	99.01%
8.0	-	99.10%	98.82%	98.85%
10.0	-	99.09%	-	-
12.0	-	99.07%	-	-
15.0	-	98.95%	-	-
20.0	-	98.77%	-	-

Table 2: 准确率-全标注

ϵ	**Baseline**	**VAT**	**Baseline+MI**	**VAT+MI**
0	92.56%	-	-	-
0.1	-	95.84%	96.59%	98.66%
0.3	-	98.29%	97.66%	98.69%
0.5	-	98.48%	98.09%	98.65%
1.0	-	98.34%	98.05%	98.56%
2.0	-	98.37%	**98.70**%	**98.72**%
3.0	-	98.46%	98.37%	**98.72**%
5.0	-	**98.71**%	98.39%	98.71%
8.0	-	98.46%	98.14%	98.64%
10.0	-	98.20%	-	-
12.0	-	98.01%	-	-
15.0	-	97.58%	-	-

Table 3: 准确率-1000标注

ϵ	**Baseline**	**VAT**	**Baseline+MI**	**VAT+MI**
0	75.03%	-	-	-
0.1	-	-	75.11%	97.99%
0.3	-	-	84.34%	98.15%
0.5	-	-	96.07%	98.52%
1.0	-	72.90%	97.63%	98.72%
1.5	-	94.82%	-	-
2.0	-	98.30%	98.10%	98.73%
3.0	-	98.51%	**98.22**%	**98.78**%
4.0	-	**98.60**%	-	-
5.0	-	97.04%	97.91%	98.65%
6.0	-	97.60%	-	-
8.0	-	96.21%	97.65%	98.63%
10.0	-	95.90%	-	-
12.0	-	95.40%	-	-
15.0	-	91.13%	-	-

Table 4: 准确率-100标注

参考文献

References

[1] Goodfellow, I. J., J. Shlens, C. Szegedy. Explaining and harnessing adversarial examples, 2015.

[2] Miyato, T., S. Maeda, M. Koyama, et al. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019.

[3] Smith, L., Y. Gal. Understanding measures of uncertainty for adversarial example detection. 2018.

[4] Deng, Z., X. Yang, H. Zhang, et al. Bayesadapter: Being bayesian, inexpensively and robustly, via bayeisan fine-tuning. *arXiv preprint arXiv:2010.01979*, 2020.

[5] Laine, S., T. Aila. Temporal ensembling for semi-supervised learning. 2016.