互信息目标下的贝叶斯对抗训练

王一诺 郑凯文 张庶杰

引言

我们提出了一种全新的半监督学习方法：**互信息指导的半监督学习**。将模型转化为变分贝叶斯网络后，通过将互信息作为不确定性度量，产生攻击样本并进行训练，进而获得鲁棒的分类面。在我们的测试中，此方法可以搭配**大部分已知的半监督学习方法**,并且能够获得准确率的提高。

什么是互信息?

互信息描述了在获得一部分真实信息之后对随机变量信息的增益[1]:

$$\begin{aligned} MI = I(X,Y) &= H[P(X)] - E_{P(y)}H[P(X|Y)] \\ &= H[P(Y)] - E_{P(x)}H[P(Y|X)] \end{aligned} \tag{1}$$

其中H为信息熵。在一个机器学习分类任务的问题背景中，Y对应数据分类的真实信息，即为Ground Truth；而随机变量X对应模型参数ω，$p(x)$对应于模型参数的后验分布$p(\omega|\mathcal{D})$。因此，互信息描述了在得知数据真实分类之后对于模型参数信息的信息增益，也就是**模型不确定性**的度量，在本问题中互信息可以表达为：

$$I(\omega, y|\mathcal{D}, x) = H[p(y|x,\mathcal{D})] - E_{p(\omega|\mathcal{D})}H[p(y|x,\omega)] \tag{2}$$

变分贝叶斯网络

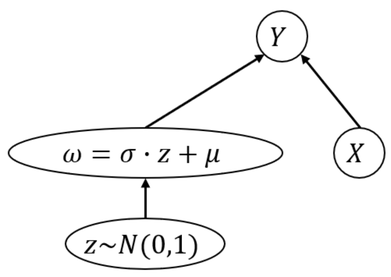


Figure 1: 贝叶斯网络

我们在此次任务中采用**变分贝叶斯网络**进行训练。变分贝叶斯网络的参数为结构参数随机变量ω对应的均值μ与方差σ^2，为了处理参数ω抽样无法追踪优化的问题，采用重参数化的手段使得参数服从$\mathcal{N}(\mu, \sigma^2)$的分布，如图1所示，此时即可对ELBO进行优化。变分贝叶斯网络学习了模型参数的后验分布$p(\omega|\mathcal{D})$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：

$$p(y|\mathcal{D}, x) \simeq \frac{1}{T}\sum_{i=1}^{T} p(y|\omega_i, x) := p_{MC}(y|x) \tag{3}$$

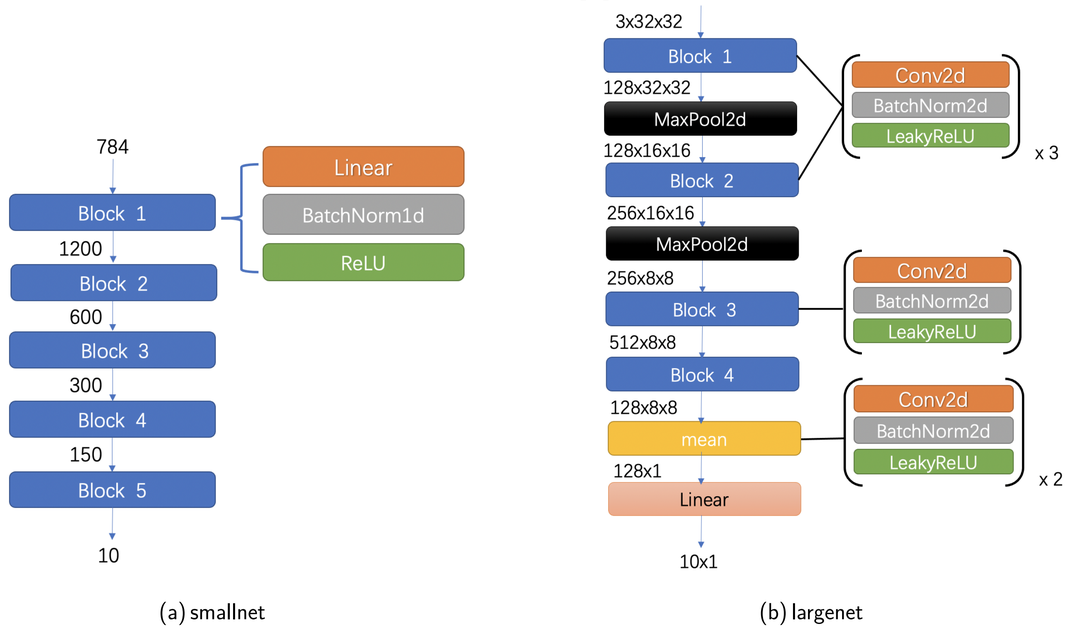
$$I(\omega, y|\mathcal{D}, x) = H[p_{MC}(y|\mathcal{D}, x)] - \frac{1}{T}\sum_{i=1}^{T} H[p(y|\omega_i, x)] \tag{4}$$

虚拟对抗攻击

我们学习了VAT[2]的核心思想，在样本点邻域内寻找使得互信息变化最大的扰动r_{adv}，并且降低在虚拟样本点上的互信息值，希望模型能够学习到一个足够鲁棒的参数后验分布$p(\omega|\mathcal{D})$，使得模型在样本点的邻域能够得到较低的不确定性。由于互信息的计算并不依赖于具体的样本标签值，因此可以利用大量的未标注样本点进行训练。

模型与方法

我们使用普通的全连接网络与卷积网络作为训练的backbone（图2）。左侧为5层MLP，用于数据集MNIST；右侧为使用了LeakyReLU作为激活函数的多层CNN，用于数据集CIFAR10。模型结构均与[2]保持一致。



(a) smallnet (b) largenet

Figure 2: 网络结构示意图

由于在邻域内最大化MI是一个复杂的优化问题，我们沿袭[2]的工作，使用近似的数值方法解决。 将梯度

$$\frac{\partial \mathcal{I}(\omega, y|x, D)}{\partial x} \tag{5}$$

的方向作为扰动的方向，将梯度向量单位化后乘以因子ϵ控制扰动的幅度。算法流程如下：

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

1.计算贝叶斯网络中有标签数据对数似然的ELBO

$$\mathcal{L}_{ELBO}(\mathcal{D}_l, \theta) = E_{q(\omega|\theta)}[\frac{1}{n}\sum_i \log(p(y_i^l|x_i^l;\omega)) - \frac{1}{n}D_{KL}(q(\omega|\theta)||p(\omega)))] \tag{6}$$

2.采样若干ω，使用有标签数据和无标签数据计算互信息

$$\mathcal{I}(\omega, y|x, \mathcal{D}) = H(\frac{1}{S}\sum_{s=1}^{S} p(y|x;\omega^{(s)})) - \frac{1}{S}\sum_{s=1}^{S} H(p(y|x;\omega^{(s)})) \quad \text{where } \omega^{(s)} \sim q(\omega|\theta) \tag{7}$$

3.寻找扰动以最大化互信息

$$r_{adv} = \epsilon \frac{\partial \mathcal{I}(\omega, y|x, D)}{\partial x} \Big/ \left\| \frac{\partial \mathcal{I}(\omega, y|x, D)}{\partial x} \right\| \tag{8}$$

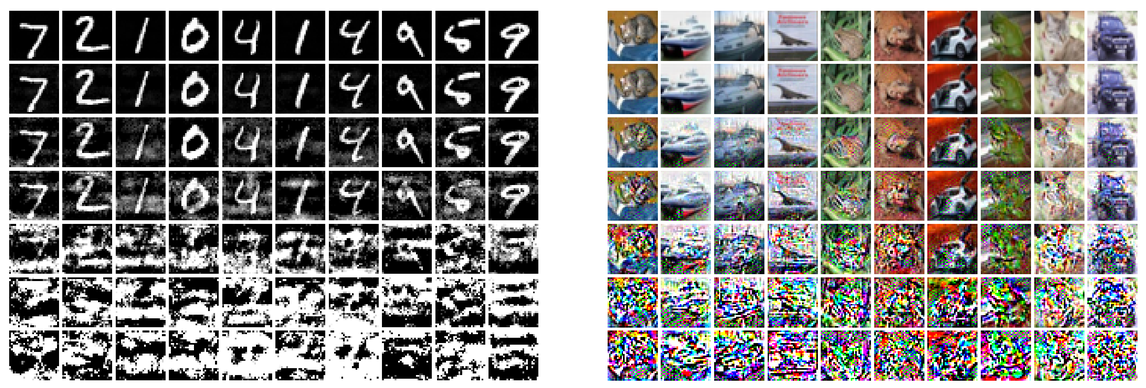
4.再次计算扰动后的互信息$\mathcal{I}(\omega, y|x + r_{adv}, \mathcal{D})$作为无标签数据的损失$\mathcal{L}_{unc}(\mathcal{D}_l, \mathcal{D}_{ul}, \theta)$

5.使用梯度下降法最小化损失

$$\mathcal{L} = -\mathcal{L}_{ELBO}(\mathcal{D}_l, \theta) + \alpha\mathcal{L}_{unc}(\mathcal{D}_l, \mathcal{D}_{ul}, \theta) \tag{9}$$

实验

实验采用数据集MNIST与CIFAR10，将使用有标签样本的交叉熵作为损失的裸模型作为Baseline；VAT在Baseline的基础上，添加了虚拟对抗损失。我们将确定性的预训练模型转化为贝叶斯网络，并按照上节流程进行对抗训练。图3展示了最大化MI攻击后的对抗样本，在$\epsilon < 3$时对抗样本的扰动不易被人眼察觉。



(a) MNIST (b) CIFAR10

Figure 3: 最大化MI攻击的对抗样本。原始图像为测试集中选取的10个样本，我们对同一模型进行ϵ值不同的攻击。从上到下依次为：$\epsilon = 0.1, 1, 5, 8, 20, 50, 100$。

表1给出了MNIST数据集上的初步实验结果。无论是有监督训练还是半监督训练，我们的方法均能在准确率上有所提升。

	标注数	Accuarcy	标注数	Accuarcy
Baseline	45000	98.4%	1000	91.6%
VAT	45000	98.7%	1000	97.4%
Baseline+MI	45000	98.7%	1000	95.0%
VAT+MI	45000	**98.9%**	1000	**97.8%**

Table 1: MNIST数据集上实验结果

Conclusion

我们提出了一种基于互信息的贝叶斯对抗训练方法，实验证明其可以作为一种有效的有监督正则化方法与半监督学习方法，对已有模型进行提升。进一步地，我们将采用对抗攻击、虚拟对抗攻击及MI最大化攻击，量化探究不同训练方式下模型的鲁棒性。

References

[1] Lewis Smith and Yarin Gal.
Understanding measures of uncertainty for adversarial example detection.
03 2018.

[2] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii.
Virtual adversarial training: A regularization method for supervised and semi-supervised learning.
IEEE Transactions on Pattern Analysis and Machine Intelligence, PP, 04 2017.