《人工神经网络》大作业开题报告

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 任务定义

在虚拟对抗训练(VAT)[1]中，作者提出了对图像分类任务的虚拟对抗攻击。具体而言，将输入的数据x施加微小扰动δ前后，使用某种散度D（如KL散度）来计算网络预测概率的变化

$$\mathcal{R}_{adv} = D(p(y|x,\omega)|p(y|x+\delta,\omega)) \tag{1}$$

并寻找扰动δ使得KL散度最大

$$\delta = \arg\max_{|\delta|\leq\epsilon} D(p(y|x,\omega)|p(y|x+\delta,\omega)) \tag{2}$$

之后，针对这一方向的攻击进行对抗防御，由此来提升网络的鲁棒性。

虚拟对抗训练使得模型倾向于生成距离样本点较远的分类面，从而能够得到一个鲁棒的网络。在理论上，寻找使得KL散度最大的扰动等价于选择使得不确定性最大的扰动，但是在DNN中概率的熵并不完全与不确定性等价，而贝叶斯网络可以自然的输出预测的不确定性，使得在贝叶斯网络中应用虚拟对抗的思想，直接攻击不确定性成为一个自然的想法。

具体而言，我们使用变分贝叶斯网络，利用θ-parameterized 变分分布$q(\omega|\theta)$近似真实的后验分布$p(\omega|\mathcal{D})$，采用优化信心下界（ELBO）的方式获得最优的$q(\omega|\theta^*)$:

$$\max_{\theta} \mathcal{L}_{ELBO} = E_{q(\omega|\theta)}[\frac{1}{n}\sum_{i}\log(p(y_i|\mathbf{x}_i;\omega)) - \frac{1}{n}D_{KL}(q(\omega|\theta)||p(\omega)))] \tag{3}$$

而在半监督的对抗学习中，以互信息的方式引入对预测标签概率分布的不确定性度量:

$$\mathcal{I}(\omega, y|\mathbf{x},\mathcal{D}) = H(\frac{1}{S}\sum_{s=1}^{S} p(y|\mathbf{x};\omega^{(s)})) - \frac{1}{S}\sum_{s=1}^{S} H(p(y|\mathbf{x};\omega^{(s)})) \quad where\ \omega^{(s)} \sim q(\omega|\theta) \tag{4}$$

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:

$$\mathcal{L}_{unc} = \max_{|\delta|\leq\epsilon} \mathcal{I}(\omega, y|\mathbf{x}+\delta,\mathcal{D}) \tag{5}$$

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:

$$\max_{\theta} \mathcal{L}_{ELBO}(\mathcal{D}_l,\theta) - \alpha\mathcal{L}_{unc}(\mathcal{D}_l,\mathcal{D}_{ul},\theta) \tag{6}$$

我们将尝试上述的贝叶斯对抗训练，并在多个数据集上进行效果对比。

2 数据集

沿袭原论文的实验设置，共有3个候选数据集：

- MNIST[1]
- SVHN[2]
- CIFAR-10[3]

这些数据集均为流行的小型数据集，且有原始实验的基准供参考。

3 挑战和基线

3.1 问题和挑战

原始VAT模型是在Tensorflow平台上实现的，需要进行模型的迁移，复现原有结果具有一定难度。此外，BNN网络相对而言较为冷门，参考资料较少，训练成本也相对较高。

3.2 基线

虚拟对抗训练主要有三方面的应用：作为有监督学习的一种正则化、进行半监督学习、提升分类网络鲁棒性。

在有监督学习和半监督学习上，使用测试错误率作为衡量指标。表1列出了VAT在不同数据集下的基线。

Table 1: 基线

模型	策略	数据集	标注数	测试错误率(%)
VAT	有监督	MNIST	60000	0.64 (±0.05)
VAT	有监督	CIFAR-10	50000	5.81 (±0.02)
VAT	半监督	MNIST	100	1.36 (±0.03)
VAT	半监督	MNIST	1000	1.27 (±0.11)
VAT	半监督	SVHN	1000	6.83 (±0.24)
VAT	半监督	CIFAR-10	4000	14.87 (±0.13)

图1显示了使用/不使用虚拟对抗训练的模型在针对这两种模型的虚拟对抗样本下的误分类率。这张图展示的是VAT对模型鲁棒性的提升，ε较小时误分类率小的模型鲁棒性更强。

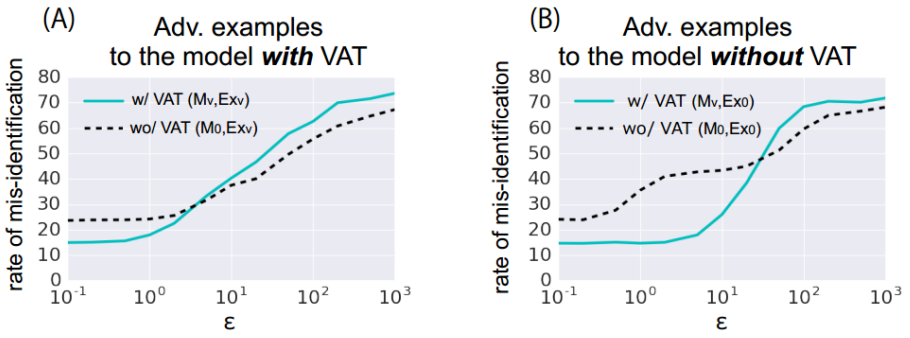

Figure 1: 在虚拟对抗攻击下的鲁棒性

模型的架构在原论文中给出。此外，原论文中还列出了一些其它有监督、半监督模型基线，此处不再赘述、

[1]http://yann.lecun.com/exdb/mnist/
[2]http://ufldl.stanford.edu/housenumbers/
[3]http://www.cs.toronto.edu/ kriz/cifar.html

4 研究计划

我们计划首先在Pytorch上复现出原始模型，并进行训练和测试。在确认能够成功复现出原论文的结论后，搭建结构一致的DNN网络并且将它转化为BNN网络，以式6作为优化目标进行训练，在同样的数据集上进行测试，与之前的模型指标进行对比。之后如果时间允许，我们计划继续进行额外测试，将原始的预训练模型通过BayesianAdapter[2]转化为BNN网络，进行对抗训练，测试这一种训练方式可能带来的性能提高。

5 可行性

首先，对于VAT背后的理论已经有了足够的研究，且BNN能够提供更加本质上的不确定性度量，加上BNN本身的特性能够在更小的标注样本集上进行训练，可以带来相对于DNN更加优越的性能。

此外，在训练BNN方面，已经有一些之前的工作，例如BayesianAdapter，能够以较低的代价将DNN网络转化为BNN网络，从而使我们的模型训练难度降低。同时，BayesianAdapter同样涉及了以互信息作为不确定性的度量，可以为我们的模型实现提供良好的参考。

在数据集和训练成本方面，我们使用的数据集规模较小，加上有专门的服务器GPU的算力，在计算资源和时间上是充足的。

6 申请MegStudio的额外算力

无需申请额外算力。

参考文献

References

[1] 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.

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