学界 | 虚拟对抗训练：一种监督和半监督学习的正则化方法

2017-11-14 机器海岸线

选自 arXiv

作者：Takeru Miyato , Shin-ichi Maeday, Masanori Koyamay and Shin Ishiiy 等

机器海岸线编译

参与：方建勇

Virtual Adversarial Training: a Regularization Method for Supervised and Semi-supervised Learning

Takeru Miyato*,†,‡, Shin-ichi Maeda†, Masanori Koyama†,§ and Shin Ishii†,‡

论文链接：https://arxiv.org/pdf/1704.03976

摘要：我们提出了一种新的基于虚拟对抗损失的正则化方法：产出分布局部平滑的一种新方法。虚拟对抗损失被定义为模型的后验分布与每个输入数据点周围局部扰动的鲁棒性。我们的方法类似于对抗训练，但与对抗训练不同，它只根据输出分布确定敌对方向，并且适用于半监督环境。因为我们平滑模型的方向实际上是对抗的，所以我们称之为虚拟对抗训练（VAT）。VAT 的计算成本相对较低。对于神经网络，虚拟对抗损失的近似梯度可以用不超过两对前向和后向传播来计算。在我们的实验中，我们将 VAT 应用于多个基准数

据集的监督和半监督学习。基于熵最小原则的附加改进，我们的 VAT 在 SVHN 和 CIFAR-10 上实现了半监督学习任务的最新性能。

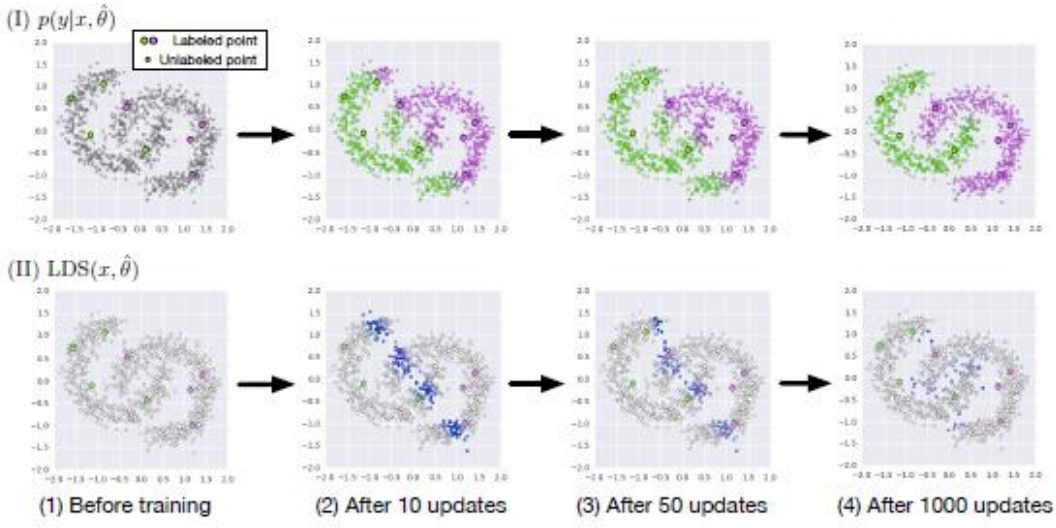


图 1：演示我们的 VAT 如何在半监督学习上工作。 我们在二维空间中生成了 8 个带标记的数据点（y = 1 和 y = 0 分别是绿色和紫色）和 1,000 个未标记的数据点。 第一行（I）中的面板显示算法的不同阶段（绿色，灰色和紫色，对应于 1.0,0.5 和 0.0 的值）的未标记输入点上的预测 $p(y=1|x,\theta)$ *。 第二行（II）中的面板是输入点上的正则化项* $\mathrm{LDS}(x,\hat{\theta})$ *的热图。 与灰色点相比，蓝色点上的 LDS 值相对较高。 我们使用 KL 散度作为方程（5）中的 D. 请注意，在训练开始时，所有数据点都以类似的方式对分类器作出了贡献。 经过 10 次更新，模型边界仍然出现在输入端。 随着训练的进展，VAT 推动边界远离标注的输入数据点。*

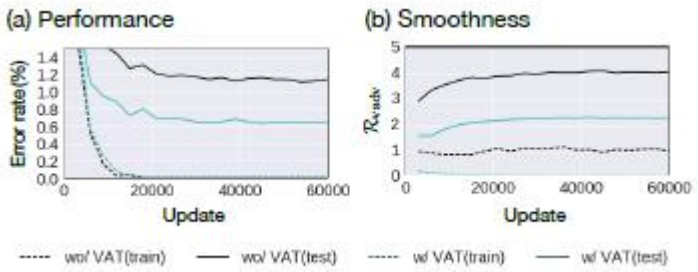


图2：（a）分类错误的转换和（b）MNIST 监督学习的 Rvadv。 我们设置 $\epsilon = 2.0$ 的基准和 VAT 评估 Rvadv。 这是 VAT 培训的模型在验证数据集上达到最佳性能的价值。

Method	Test error rate(%)
SVM (Gaussian kernel)	1.40
Dropout [35]	1.05
Adversarial, L_∞ norm constraint [11]	0.78
Ladder networks [30]	**0.57**
Baseline (MLE)	1.11
RPT	0.82
Adversarial, L_∞ norm constraint	0.79
Adversarial, L_2 norm constraint	0.71
VAT	0.64

表1：在置换不变设置下，用6万个标记的例子测试 MNIST 上的监督学习性能。 第一行引用了原文提供的结果。 第二行显示了作者实现的表现。

Method	Test error rate(%)
Network in Network [23]	8.81
All-CNN [34]	7.25
Deeply Supervised Net [22]	7.97
Highway Network [36]	7.72
Baseline (only with dropout)	6.76
RPT	6.25
VAT	**5.81**

表2：在有监督的 CIFAR-10 上使用50000 个标记的例子测试 CNN 的性能。

Method	Test error rate(%)	
	$N_l = 100$	$N_l = 1000$
TSVM [5]	16.81	5.38
Pseudo Ensembles Agreement (PEA) [3]	5.21	2.87
Deep Generative Model [19]	3.33	2.59
CatGAN [33]	1.91	1.73
Ladder Networks [30]	1.06	**0.84**
GAN with feature matching [32]	**0.93**	
RPT	6.81	1.58
VAT	1.36	1.27

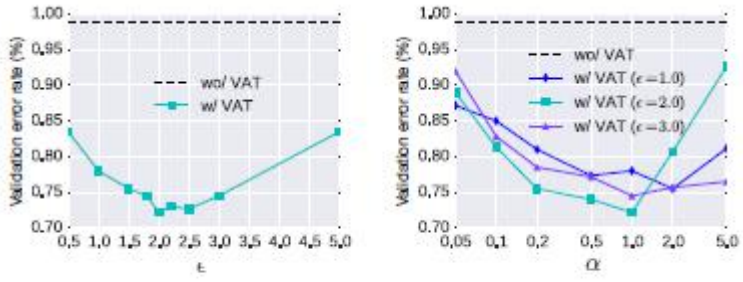
表3：在半监督 MNIST 上测试性能，在置换不变设置上使用100 个和1000 个标记示例。

Method	Test error rate(%) SVHN $N_l = 1000$	CIFAR-10 $N_l = 4000$
SWWAE [43]	23.56	
Skip Generative Model [24]	16.30	
Ladder networks, Γ model [30]		20.40
CatGAN [33]		19.58
GAN with feature matching [32]	8.11	18.63
∏ model [21]	5.43	16.55
(on Conv-Small used in [32])		
RPT	8.41	18.56
VAT	**6.83**	**14.87**
(on Conv-Large used in [21])		
VAT	5.77	14.82
VAT+EntMin	**4.28**	**13.15**

表 4：SVHN（1,000 标记）和 CIFAR-10（4000 标记）的测试性能没有图像数据增加。

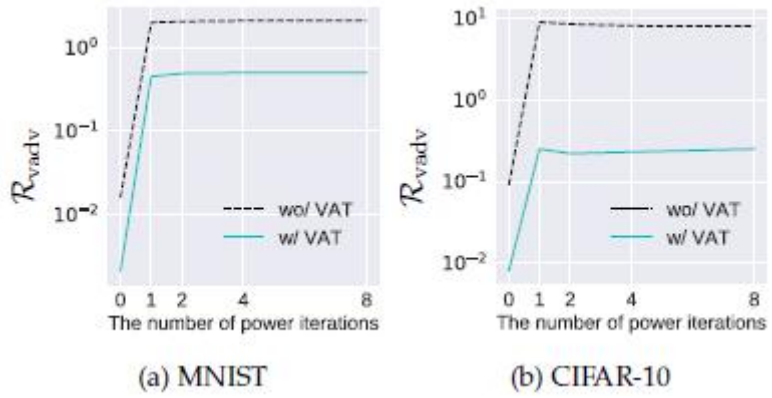
Method	Test error rate(%) SVHN $N_l = 1000$	CIFAR-10 $N_l = 4000$
Γ model [30] (experimented by [21])		≈ 16
∏ model [21].	4.84	12.36
Temporal ensembling [21]	4.43	12.16
Sajjadi et al. [31]		11.29
(On Conv-Large used in [21])		
VAT	5.42	11.36
VAT+EntMin	**3.86**	**10.55**

表 5：SVHN 和 CIFAR-10 的图像数据增强测试性能。 除 Sajjadi 等人以外的所有方法的表现。 [31]是基于中等数据增强翻译和翻转的实验（更多细节见附录 D）。 Sajjadi 等人 [31]使用广泛的图像增强，其中包括旋转，拉伸和剪切操作。



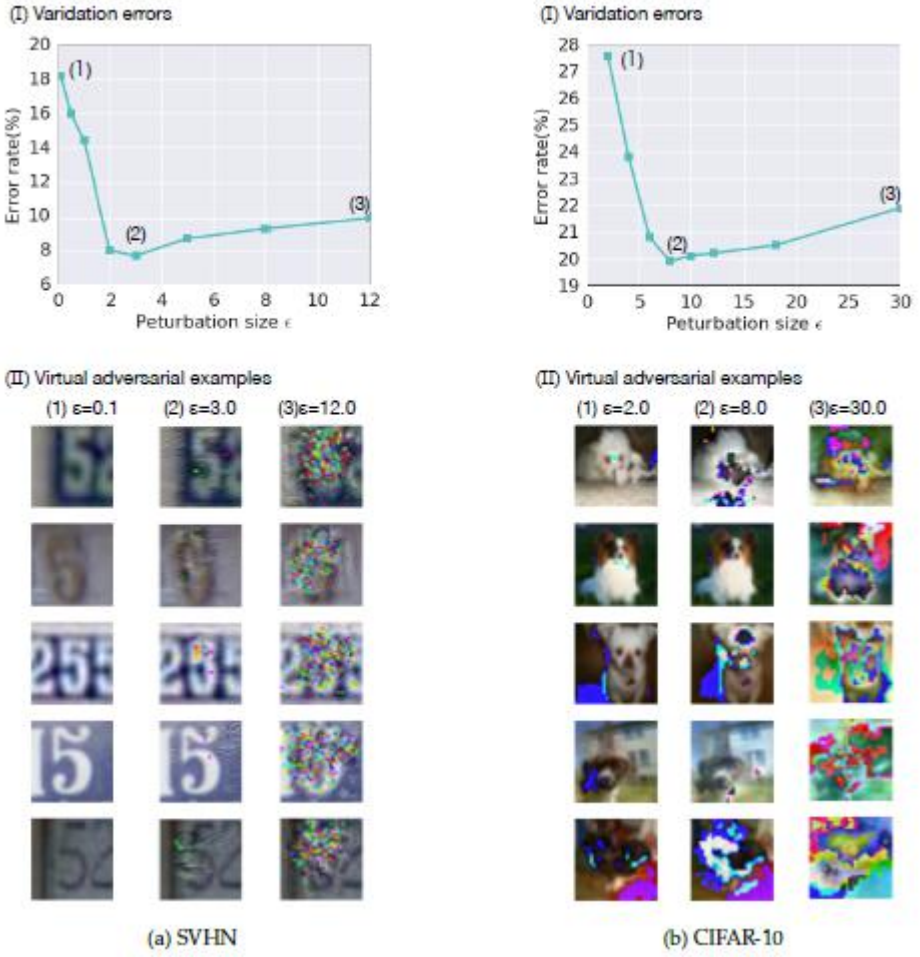
(a) Effect of ϵ ($\alpha = 1$). (b) Effect of α.

图 3：监督 MNIST 的效果和验证性能。



(a) MNIST (b) CIFAR-10

图 4：（a）有监督的 MNIST 和（b）半监督的 CIFAR-10 的功率迭代次数对 Rvadv 的影响。



(a) SVHN (b) CIFAR-10

图 5：不同价值的 VAT 表现。 半监督学习的性能的影响，以及一个 VAT 训练模型产生的具有相应值的典型虚拟对抗实例。

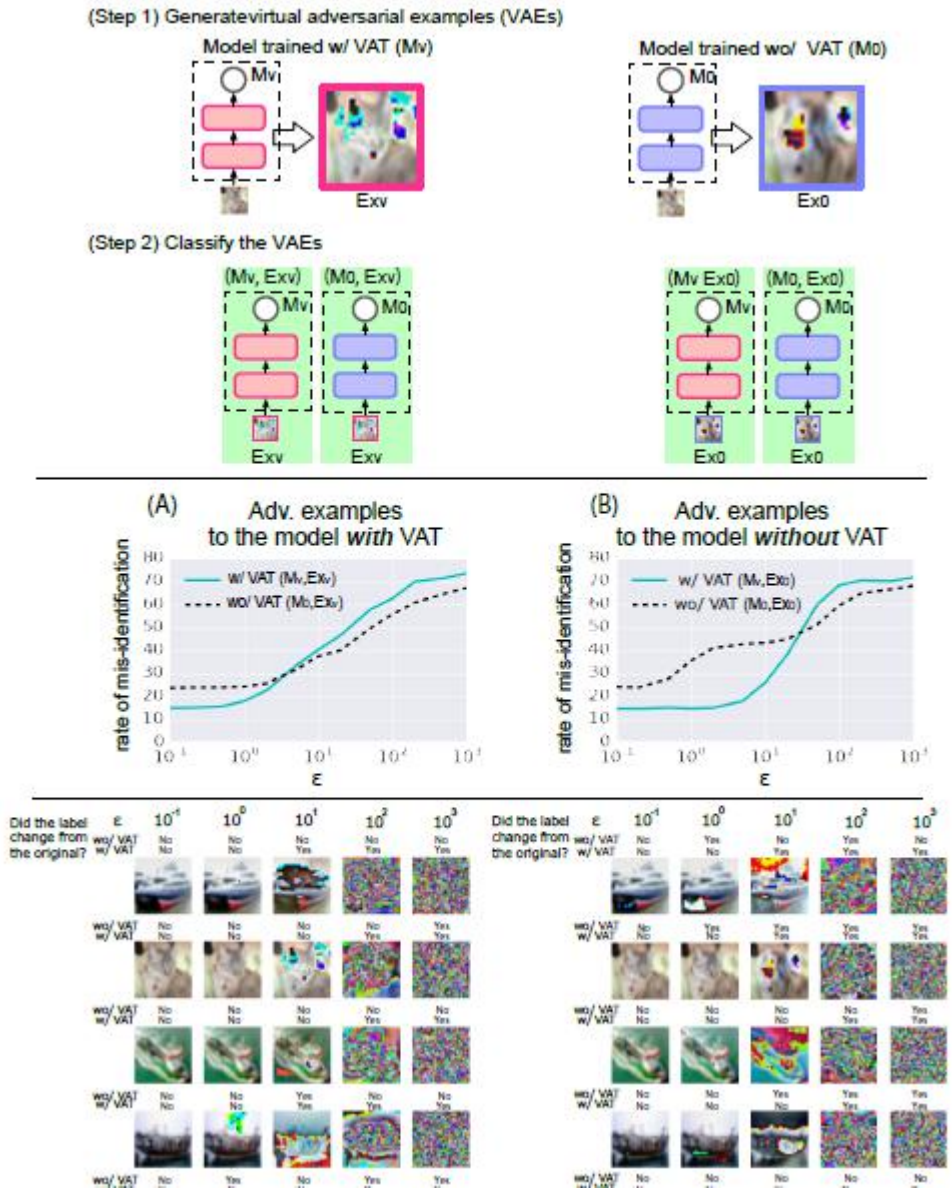


图 6：VAT-training 模型对干扰图像的鲁棒性。上图显示了评估鲁棒性的程序。在第一步中，我们准备了两个分类器 - 一个用 VAT（Mv）和另一个用 VAT（M0）进行训练，并从每个分类器（Exv 和 Ex0）中生成一个虚拟的敌对示例。在步骤 2 中，我们将 Exv 和 Ex0 分类到这两个模型，从而总共产生四个分类结果。中间面板（图 A 和 B）绘制了在这四个分类任务中做出的误识别率与在步骤 1 中用于产生虚拟对抗示例（VAEs）的摄动（）的大小的关系。底部面板的左半部分对齐（A）显示了一系列不同的 Exv 值，以及图像上 Mv 和 M0 的结果。列出的所有 Exv 都是从干净的例子中产生的图像，Mv 和 M0 都是正确的标识。底部面板的右半部分与图（B）对齐，显示了使用不同值生成的 Ex0 集合。标签"是"表示当对图像施加扰动时，模型改变了标签分配。标号"否"表示该模型在扰动的图像上保持标签分配。请注意，Mv 模型在图像上占主导地位，与人眼清晰的图像几乎没有区别。

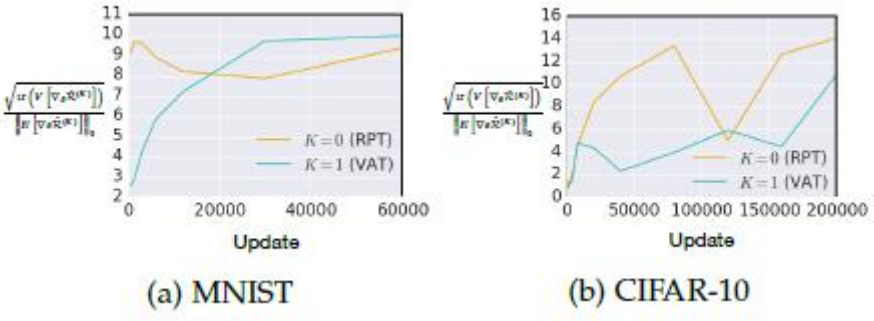


图 7：在 MNIST 监督学习的 NNS VAT 培训和 CIFAR-10 半监督学习期间 R（0）和 R（1）的归一化 SD 范数的转换。

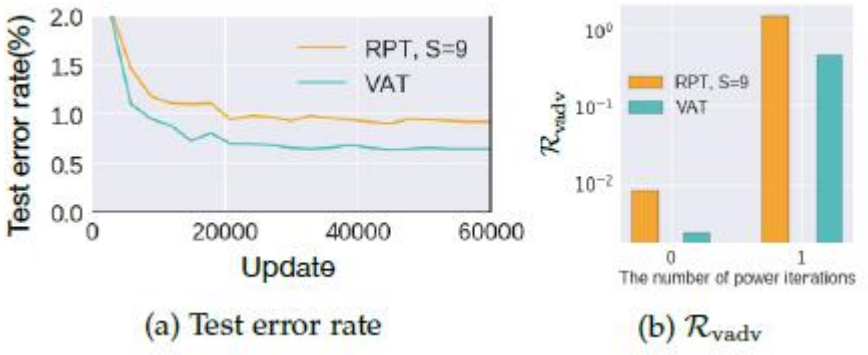


图 8：a）学习曲线和（b）VAT 的 Rvadv 实施了 $\alpha = 1$ and $S = 1$，*RPT 实施最佳* $\alpha(= 7)$ and $S = 9$.。*其余超参数* ϵ *设置为 2.0 为 VAT 和 RPT。*

本文为机器海岸线编译，转载请联系 fangjianyong@zuaa.zju.edu.cn 获得授权。

✂----------