Mixup 技术分析

11821043 陈晋泰

摘要

Mixup 技术是 2017 年提出了一种简单有效的新型模型正则化、数据增强方法，作用效果显著、执行简单，已经被广泛的应用在深度学习图像分类任务中。本文针对论文《Mixup: Beyond Empirical Risk Minimization》的思想，从数据增强、模型正则和流形学习三个角度展开分析：Mixup 对模型过拟合的缓解作用、Mixup 对模型线性正则的作用，和从流形空间中看到的 Mixup 的一些缺点。并通过简单的实验证实了 Mixup 的缺点。

一、简介与综述

深度学习在图像分类领域已经做出了显著从成绩。同其他模型相似，一般是从总体中抽样出训练样本，依照经验风险最小化原则（ERM）[1]进行模型训练。为了约束收敛空间，许多图像正则化技术被研究者提出。在图像分类领域，旋转、裁剪、放缩、翻转、随机擦除、梯度裁剪等技术被广泛使用。和 Mixup 相关的，Chawla 等[2]提出了采用插值的方法来增加不平衡数据集中的数据数量。这个方法通过在训练数据集中临近的数据点间采用插值方法，取得了更好的泛化性。然而这个方法没有扩展到不同类的数据之间的插值，也没有考虑非临近数据点之间的插值。目前，对输出项进行也有正则的方法，如标签平滑[3]等，也提倡信息混合。然而，这些方法与 Mixup 技术都有本质的不同，对输出项进行正则本质上不能帮助数据学习类别中和类别间的信息。因此，这些方法和 Mixup 所解决的问题从本质上而言是不同的。

实际上，深度学习网络有可能“学习”了数据，也有可能仅仅是“记住”了数据——即时在拥有很强的正则化的情况下。是记忆还是学习，就涉及到了泛化的问题。对抗性样本的出现[4]：将噪声加入图像中会使图像分类出现错误，这种现象超过了 ERM 能解释的范围。已提出的数据增强（data augmentation）技术[5]和邻域最小化风险理论在 VRM[6]中，需要专业知识描述训练数据中每个样本的邻域，从而可以从训练样本邻域中提取附加的虚拟样本以扩充对训练分布的支持。数据增强可以提高泛化能力，但这一过程依赖于数据集，而且需要专家知识。因此，这方法也不是具有适宜推广的特性。

Mixup[7]技术可以被视为一种图像数据扩增的手段，看似不合理，但操作简单，效果也很好。从图像的角度看来，这是一种数据增强技术；从对模型的影响来看，这是一种模型的正则化方案。它的操作方法很简单，即将一个数据对（(x_i, y_i)，图像和标签）进行混合，加入到模型训练中：

$$\begin{cases} y = \alpha y_i + (1-\alpha) y_j \\ x = \alpha x_i + (1-\alpha) x_j \end{cases}, \ \alpha \in [0,1]$$

和它一脉相承的有 Pair Sampling 技术[8]，它几乎是在和 Mixup[7]同期提出的，却更加反直觉。随机取一个样本x_j与样本x_i均匀混合$x = 0.5x_i + 0.5x_j$，让他们的标签仍是y_i。这个过程如图 1 展示：将一张猫的图片和一张狗的图片直接混合，要求标签仍是猫或狗。

图 1 将一张猫的图片和一张狗的图片按照各占 0.5 的比例混合

实际上，这是一种对称的做法，因为即使猫和狗的图片混合而标签不混合，每一个标签被抽到的概率依然是 0.5。在梯度下降法中，他们对总体模型的贡献依然是 0.5，相当于标签经过了混合。另外，与 Pair Sampling 不同的是，Mixup 技术不将权重限制在 0.5，而是通过在贝塔分布中采样。因此，与其认为 Mixup 和 Pair Sampling 是两项技术，不如认为这是一项技术的两种变体，而 Mixup 更加灵活、稳定，而 Pair Sampling 的做法加入一个抽样的过程，模型的方差会更大。

二、从数据增强的角度分析 Mixup

以往的图像数据增强方式，是对图像数据进行翻转、裁剪等操作，使得模型对作者所提出的图像变换方法具有鲁棒性。正式地，数据增强是通过一个定义在图像上的变换$I(\cdot)$，使得模型能够对这个变换具有不变性：

$$f(x) = f(I(x))$$

其中$f(\cdot)$代表图像处理模型，这里我们集中讨论深度学习模型。显然，当加入数据增强项时，要学习的模型往往可以认为是从$f(\cdot)$变成了$f * I'(\cdot)$，这里，$I'(\cdot)$表示对图片不变性处理模块。这可能需要额外的参数。对于简单的数据增强方法，例如旋转、翻转等数据增强方式，由于原图像x和变换后的图像$I(x)$相似性较大，可以推测$I'(\cdot)$所需要的参数量较小，也就是模型通过改变一些参数，可以迅速的找到一个变动不大的模型来满足这项约束。因此，这些数据增强的方式给模型带来的促进并不充分。然而，对于 Mixup 和 Pair Sampling 的方法，$I(\cdot)$给图像带来的变化是巨大的。此时，一个简单的图像不变性处理模块$I'(\cdot)$是无法拟合的，此时，不能简单的认为模型从$f(\cdot)$变成了$f * I'(\cdot)$，这项正则化技术是不能被模型轻易的“消耗掉”的。模型需要做出更大的改变，才能很好的拟合这项数据增强技术提出的约束。从过拟合的角度讲，一个模型的参数越多，越容易出现过拟合。因此，要满足这项复杂的数据增强方式，需要付出更多的参数代价。对于同一个模型而言，用于满足数据增强的参数越多，用于拟合的参数则越少，其过拟合的风险就越低。

从另外一个角度讲，这项正则化方法，将数据从 N 扩充到了 N^2，当模型的参数量没有呈平方倍数增长时，记住所有训练数据的难度增加了，其过拟合的风险就随之降低了。

三、从模型正则的角度分析 Mixup

Mixup 技术不仅仅是一种数据增强技术，也是一种模型正则化技术。对于训练集$(x_i, y_i), i = 1,2,3,\dots,n.$ 要找到一个模型$f(\cdot)$，满足分类任务。由于深度学习模型有很强的非线性拟合能力，因此，就容易过拟合。而 Mixup 这项技术就在模型满足分类准确率的基础上，提出了线性约束。

对于不加正则项的模型要求：$y_i = f(x_i), i = 1,2,3,\dots,n$。而 Mixup 对模型提出了一个更高的要求，即满足$f\big(\alpha x_i + (1-\alpha)x_j\big) = \alpha y_i + (1-\alpha)y_j = \alpha f(x_i) + (1-\alpha)f(x_j)$。其中，$(x_i, y_i)$取遍训练集。由于训练集中的数据点是密集的，我们可以假设它是连续的。由于α是来自模型之外的随机变量，容易证得一个满足$f\big(\alpha x_i + (1-\alpha)x_j\big) = \alpha f(x_i) + (1-\alpha)f(x_j)$的$f(\cdot)$是一个线性函数。因此，我们可以认为 Mixup 对模型加入了一个强约束：在满足模型拟合的前提下，尽量保持线性。

而构造 Mixup 的思路又是十分精巧的。相对于使用拉格朗日乘子法需要考虑线性约束和模型拟合约束之间的比例从而引入超参数，Mixup 不需要显式地引入超参数。

四、从流形学习的角度分析 Mixup

从流形学习的角度讲，Mixup 引入了对流形空间的建模。该建模方式认为：流形空间之间两点连线上的点的标签，也应该在两者之间。从这点角度出发，我们同样可以认为，Mixup 为模型引入了线性范式。然而，这也暴露了模型的几个问题：

（1）当模型两点连线之中的某个点不属于流形空间，那么这项正则有可能会破坏了模型对流形空间的学习。

（2）当模型两点连线经过一些数据点，则可能给该数据点附上不一致的标签。

这两种情况如图 2 所示。

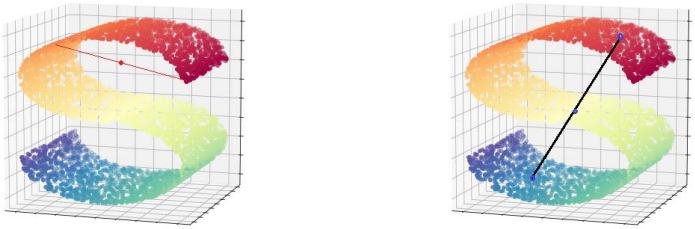

图 2　Mixup 造成的流形空间的混杂

而实际上，由 Mixup 造成的流形空间的变形可以看到。对于图 1 中猫狗图片的叠加得到的图片，显然是不存在于流形空间中的。这就造成了问题 1 所说的现象。在手写数字识别任务中 2 和 1 的叠加可能出现 4，6 和 9 的叠加可能出现 8。这些情况则属于问题 2 中体现的情况。这些情况表明了，Mixup 这项正则化手段所体现出来的直观上的“不合理性”，是确实存在不合理之处。

针对第 2 项问题，Guo H[9]等人提出了一个双模型模型结构 AdaMixUp，一个用于生成混合参数α，一个用于 Mixup 分类，取得了比 Mixup 更加良好的效果。接着，Verma V[10]参考了 Mixup 技术，提出了 Manifold Mixup 技术。这项技术旨在使得类间界限（超平面）更加平滑。这项技术作用在不同类别之间，要求在空间中，不论类间界限（超平面）离哪个类别的点更近，都以 50%的概率认为该点共属于两个类。这项技术的逻辑在于：当加入这项正则之后，类别之间的界限会更加趋近于线性，于是模型表征就会更倾向于落入 c-1 维空间中（c 为类别数）。然而，这项技术同样面临着问题 2 类似的困境，例如：当空间中的一个点被属于四个类型的数据点包围。作者认为，这种不一致的情况可以被间接避免，如果每个类别的特征足够集中。为了达到这个目的，则要设置子空间的维度等于“隐藏层数-c+1”，使表征没有多余的自由度引入混杂。

五、实验说明

为了说明 Mixup 技术在流形空间中造成的混杂作用，我们从 MNIST 数据集中提取数据，构造了两个子数据集 $1\{c(8),c(6),c(9)\}$和子数据集 $2\{c(1),c(7),c(4)\}$，其中$\{c(\cdot)\}$表示 MNIST 数据集中某个类别所有的数据点。MNIST 数据集、子数据集 1 和子数据集 2 都按照官方划分为包含训练集和测试集。搭建一个简单的三层卷积神经网络，在层间使用 ReLU 激活函数，并在结尾用一层全连接网络作为分类器。对于数据，采用两组对比实验，一组是 MixUp 的数据、一组是无 MixUp 的数据。实验结果表 1 所示。

表 1 分类准确率记录表。其中，括号中的数据表示测试集分类准确率、括号外的数据表示训练集分类准确率。每个实验经过三次实验，取其平均值记录于表格中。

准确率	MNIST	子数据集 1	子数据集 2
Mixup	98.8%（97.9%）	95.6%（94.5%）	96.8%（96.1%）
Without Mixup	99.2%（98.1%）	97.7%（97.3%）	98.1%（97.2%）

实验可见，Mixup 技术总体上对 MNIST 数据集的分类准确率没有明显影响，但是对于特殊挑选的数据集却有较为明显的准确率下降的情况。究其原因在于我们所挑选的数据集容易出现问题 2 的混杂问题。因此该实验暴露了 Mixup 技术的问题。

参考文献：

[1] Vapnik V. Statistical Learning Theory[J]. Annals of the Institute of Statistical Mathematics, 2003,

55(2):371-389.
[2] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1):321-357.
[3] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[J]. CVPR, 2016:2818-2826.
[4] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. Computer Science, 2013.
[5] Simard P Y, Cun Y A L, Denker J S, et al. Transformation invariance in pattern recognition: Tangent distance and propagation[J]. Intl J of Imaging Systems & Technology, 2001, 11(3):181–197.
[6] Chapelle O, Weston J, Bottou L, et al. Vicinal risk minimization[C]//Advances in neural information processing systems. 2001: 416-422.
[7] Zhang H, Cisse M, Dauphin Y N, et al. Mixup: Beyond Empirical Risk Minimization[J]. ICLR. 2017.
[8] Inoue H. Data Augmentation by Pairing Samples for Images Classification[J]. 2018.
[9] Guo H, Mao Y, Zhang R. MixUp as Locally Linear Out-Of-Manifold Regularization[J]. 2018.
[10] Verma V, Lamb A, Beckham C, et al. Manifold Mixup: Better Representations by Interpolating Hidden States[J]. 2018.