解相关的批量标准化

## 抽象

批量标准化（BN）能够通过在小批量中进行居中和缩放激活来加速深度模型的训练。 在这项工作中，我们提出了Decorrelated Batch Normalization（DBN），它不仅可以集中和扩展激活，还可以对它们进行白化。 我们探索了多种美白技术，并发现PCA美白会导致我们称之为随机轴交换的问题，这对学习是不利的。 我们表明，ZCA美白不会受到这个问题的影响，允许成功学习。 DBN保留了BN的理想品质，进一步提高了BN的优化效率和泛化能力。 我们设计了全面的实验来证明DBN可以提高BN在多层感知器和卷积神经网络上的性能。 此外，我们不断提高CIFAR-10，CIFAR-100和ImageNet上残留网络的准确性。

## 简介

批量标准化[25]是一种加速深度网络训练的技术。由Ioffe和Szegedy介绍，它已被广泛用于各种最先进的系统[17,19,19,54,48,20]。批量标准化通过标准化小批量内的深层网络的激活来工作 - 转换层的输出，或者等效地输入到下一层，以具有零均值和单位方差。具体来说，让{xi∈R，i = 1,2，...，m}成为小批量m个例子中单个神经元的原始输出.Batch Normalization产生变换后的输出，其中是均值和方差。小批量，ǫ> 0是一个小数字，以防止数值不稳定，γ，β是额外的可学习参数。至关重要的是，在训练期间，批量标准化是推理计算（前向传递）以及梯度计算（后向传递）的一部分。批量标准化可以广泛地插入到网络中，通常在线性映射和非线性之间。

批量标准化的动机是众所周知的事实，即白化输入（即居中，去相关和缩放）加速了训练[32]。已经表明，更好地调节输入的协方差矩阵可以更好地调节Hessian更新权重，使梯度下降更新更接近牛顿更新[32,50]。批量标准化通过寻求不仅美白网络第一层的输入，而且还美白网络中每个内部层的输入，进一步利用了这一事实。但批量标准化仅执行标准化，而不是白化。也就是说，激活是集中和缩放的，但不是解相关的。这种选择在[25]中是合理的，引用了美白的成本和可区分性，但没有实际尝试推导或试验美白操作。

虽然标准化已被证明对批量标准化有效，但是对于批量标准化的完全增白 - 添加去相关是否可以进一步帮助仍然是一个有趣的问题。从概念上讲，有明显的情况，美白是有益的。例如，当激活接近完全相关时，标准化几乎不会改善协方差矩阵的条件，而白化仍然有效。此外，先前的工作表明，去相关的激活导致更好的特征[3,43,5]和更好的泛化[9,53]，这表明有进一步改进批量标准化的空间。

在本文中，我们提出了Decorrelated Batch Normalization，其中我们在小批量中对每层的激活进行白化。令xi∈Rd为小批量m的第i个示例的层的输入。白化输入由小批量平均值和Σ=给出

  是小批量协方差

矩阵。

实现Decorrelated Batch Normalization会产生几个问题。一个是如何进行反向传播，特别是如何反向传播通过矩阵的逆平方根（即），其关键步骤是特征分解。这种矩阵变换的可区分性是批量标准化论文[25]中没有追求白化的原因之一。 Desjardins等人。 [12]通过将平均μ和白化矩阵作为模型参数处理而不是作为输入示例的函数，使激活变白但避免通过它进行反向传播。但是，正如已经指出的那样[25,24]，这样做可能会导致培训不稳定。

在这项工作中，我们对激活进行去相关并在训练期间进行适当的反向传播。我们通过使用特征分解是可微分的这一事实来实现这一点，并且可以使用矩阵微积分来获得其导数，如先前的工作[15,26]所示。我们以这些现有结果为基础，推导出Decorrelated Batch Normalization的反向传播更新。

另一个问题是，或许令人惊讶的是，如何计算白化矩阵的选择。白化矩阵不是唯一的，因为白化输入在任意旋转后保持白化[27]。事实证明，PCA美白是一种标准选择[13]，根本不会加速训练，实际上会造成重大伤害。原因是PCA白化通过执行旋转然后进行缩放来工作，但是旋转可能导致我们称为随机轴交换的问题，如将在3.1节中讨论的那样，实际上随机地置换每个批次的层的神经元。这种排列可以将数据表示从一个批次大幅改变到另一个批次，达到训练永不收敛的程度。

为了解决这个随机轴交换问题，我们发现使用ZCA白化[4,27]是至关重要的，它会使PCA白化激活旋转回来，使原始激活的失真最小。我们通过实验证明，去相关的好处只能通过ZCA增白的额外旋转来观察。

第三个问题是要进行的美白量。给定特定的批量大小，DBN可能没有足够的样本来获得完全协方差矩阵的合适估计。因此，我们通过将较小的激活组去相关而不是一起激活来控制白化的程度。也就是说，对于维度d的输出，我们将其划分为大小为kG <d的组并且在每个组内应用白化。该策略具有额外的益处，即降低从O（d2 max（m，d））到O（mdkG）的白化的计算成本，其中m是小批量大小。

我们对多层感知器和卷积神经网络进行了广泛的实验，并表明解相关批量归一化（DBN）在训练速度和泛化性能方面改进了原始批量归一化（BN）。特别是，实验表明，使用DBN可以持续改善残留网络[17,19,54]对CIFAR-10，CIFAR-100 [29]和ILSVRC-2012 [11]的性能。

## 结论

在本文中，我们提出了解相关批量标准化（DBN），它扩展了批量标准化以包括对小批量数据的白化。 我们发现PCA美白有时可能对训练有害，因为它会引起随机轴交换，并证明使用ZCA美白至关重要，这避免了这个问题。 DBN保留了批量标准化的优势，同时使用去相关表示来进一步提高模型的优化效率和泛化能力。 这是因为DBN可以保持近似动态等距并改善Fisher信息矩阵的条件。 这些属性经过实验验证，表明DBN在设计DNN架构时具有很大的潜力。