**周志华最新论文挑战深度学习 | 深度森林：探索深度神经网络以外的方法**

**新智元导读】**西瓜书《机器学习》作者、南京大学周志华日前发表论文，提出了一种基于树的方法，挑战深度学习。在设置可类比的情况下，新方法 gcForest 取得了和深度神经网络相当甚至更好的结果，而且更容易训练，小数据也能运行，更重要的是相比神经网络，基于树的方法不会存在那么困难的理论分析问题。周志华和冯霁在论文里写道，“我们认为，要解决复杂的问题，学习模型也需要往深了去。然而，当前的深度模型全部都是神经网络。这篇论文展示了如何构建深度树（deep forest），为在许多任务中使用深度神经网络之外的方法打开了一扇门”。**【进入新智元公众号，在对话框输入“170302”下载论文】**

深度神经网络的巨大成功掀起了一股深度学习热潮。

或许，这股热潮有些“太热”了。

就在这周，南京大学机器学习与数据挖掘研究所（LAMDA）负责人周志华教授将他和 LAMDA 博士生冯霁写的论文《深度森林：探索深度神经网络以外的方法》（Deep Forest: Towards An Alternative to Deep Neural Networks）上传到 arXiv。可以说，是用实际行动为这股热潮降了降温。

深度学习，或者说深度神经网络有一些明显的缺陷，除了我们都知道的必须要有大数据、大算力，他们在论文中写道，

“众所周知，表征学习（representation learning）能力对于深度神经网络至关重要。还应注意，为了利用大规模训练数据，学习模型的容量（capacity）应该很大；这一点非常好的说明了相比普通的学习模型，比如支持向量机，深度神经网络十分复杂的原因。

**“我们推测，若能将这些属性赋予其他一些形式合适的学习模型，我们可以在抛开上述缺陷的情况下，实现和深度神经网络相当的性能。”**

由此，周志华和冯霁提出了一种基于树的新方法——gcForest，用文中的术语说，就是“multi-Grained Cascade forest”，多粒度级联森林。此外，他们还提出了一种全新的决策树集成方法，使用级联结构让 gcForest 做表征学习。实验中，gcForest 使用相同的参数设置，在不同的域中都获得了优异的性能，并且无论是大规模还是小规模的数据，表现都很好。此外，由于是基于树的结构，gcForest 相比神经网络也更容易分析。

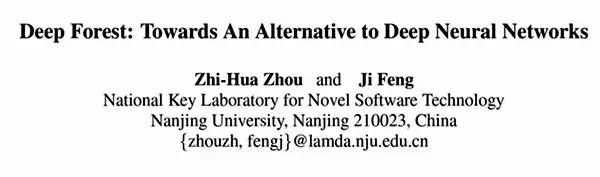
那么，作者是打算用 gcForest 取代深度学习？

答案是否定的。正如他们在论文中所说：

**“我们认为，要解决复杂的问题，学习模型也需要往深了走。然而，当前的深度模型全部都是神经网络。这篇论文展示了如何构建深度树（deep forest），为在许多任务中使用深度神经网络之外的方法打开了一扇门。”**

向着深度学习以外的方法进军

下面是新智元对论文《深度森林：探索深度神经网络以外的方法》的摘译。要查看完整的论文，参见文末的论文地址。



**摘要**

在这篇论文里，我们提出了 gcForest，这是一种决策树集成方法（decision tree ensemble approach），性能较之深度神经网络有很强的竞争力。深度神经网络需要花大力气调参，相比之下 gcForest 要容易训练得多。实际上，在几乎完全一样的超参数设置下，gcForest 在处理不同领域（domain）的不同数据时，也能达到极佳的性能。gcForest 的训练过程效率高且可扩展。在我们的实验中，它在一台 PC 上的训练时间和在 GPU 设施上跑的深度神经网络差不多，有鉴于 gcForest 天然适用于并行的部署，其效率高的优势就更为明显。此外，深度神经网络需要大规模的训练数据，而 gcForest 在仅有小规模训练数据的情况下也照常运转。不仅如此，作为一种基于树的方法，gcForest 在理论分析方面也应当比深度神经网络更加容易。

**引言**

深度神经网络尽管性能强大，但却拥有明显的不足。首先，众所周知深度神经网络在训练时需要大量的训练数据，也因此让深度神经网络无法被用于小规模数据任务。需要指出，就算在如今这个大数据时代，由于高昂的标记成本，有很多真实世界任务仍然没有足够多的带标签的数据，因此也使深度神经网络在这些任务上表现欠佳。

其次，深度神经网络是非常复杂的模型，因此训练过程通常需要强大的计算设施，这致使身处大公司外的个人无法充分发挥其学习的潜力。更重要的是，深度神经网络超参数太多，其学习的性能严重依赖于调参的过程。举例说，当几位作者都使用神经网络[LeCun et al., 1998; Krizhenvsky et al., 2012; Simonyan and Zisserman, 2014]，他们实际使用的也是不同的模型，这是由于卷积层结构等很多不同的选择造成的。这不仅使深度神经网络的训练十分麻烦——与其说是科学或工程，更像是艺术，而且导致针对深度神经网络的力量分析也极为困难，因为有太多的干扰因素，几乎无限的配置组合。

众所周知，表征学习（representation learning）能力对于深度神经网络至关重要。还应注意，为了利用大规模训练数据，学习模型的容量（capacity）应该很大；这一点非常好的说明了相比普通的学习模型，比如支持向量机，深度神经网络十分复杂的原因。我们推测，若能将这些属性赋予其他一些形式合适的学习模型，我们可以在抛开上述缺陷的情况下，实现和深度神经网络相当的性能。

在这篇论文里，我们提出了 gcForest（multi-Grained Cascade forest，多粒度级联森林），以及一种全新的决策树集成方法。这种方法生成一个深度树集成方法（deep forest ensemble method），使用级联结构让 gcForest 做表征学习。当输入带有高维度时，通过多粒度扫描，其表征学习能力还能得到进一步的提升，而这有望使 gcForest 能注意到上下文或结构（contextual or structural aware）。级联的数量能够根据情况进行调节，从而使 gcForest 在只有小数据的情况下也表现出优异的性能。需要指出，gcForest 的超参数比深度神经网络少得多；更好的是 gcForest 对于超参数设定性能鲁棒性相当高，因此在大多数情况下，即使遇到不同领域的不同数据，也能使用默认设定取得很好的结果。这不仅使 gcForest 训练起来很容易，也使其理论分析更为简单，虽然本文并没有涉及这方面的讨论（不消说，树学习模型［tree learner］通常比神经网络更容易分析）。在我们的实验中，gcForest 取得了比深度神经网络相当甚至更好的成绩，而训练时间方面，gcForest 在 PC 上的训练时间与使用 GPU 设置训练的深度神经网络相当。需要指出，gcForest 天然就更适用于并行部署，因此这种效率优势就更为明显。

我们认为，要解决复杂的问题，学习模型也需要往深了走。然而，当前的深度模型全部都是神经网络。这篇论文展示了如何构建深度树（deep forest），为在许多任务中使用深度神经网络之外的方法打开了一扇门。

接下来，我们会介绍 gcForest 及其实验结果，在分析相关工作后得出结论。

**方法介绍**

这里，我们选取了 gcForest，也即“多粒度级联森林”关系最密切的内容，包括级联森林、多粒度扫描，以及实验结果和相关的示意图作介绍。

**2.1 级联森林（Cascade Forest）**

深度神经网络中的表征学习（representation learning）主要依赖于对原始特征进行逐层处理。受此启发，gcForest 采用级联结构（cascade structure），如图1所示，其中级联中的每一级接收到由前一级处理的特征信息，并将该级的处理结果输出给下一级。

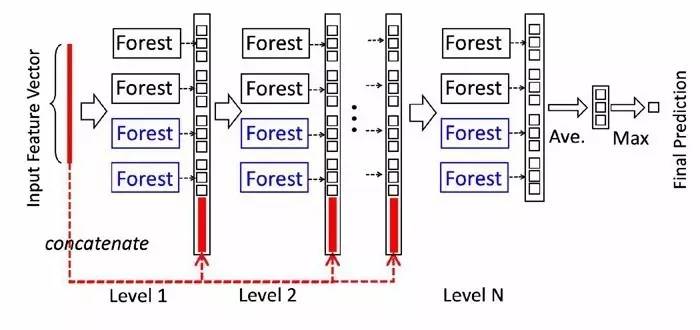


图1：级联森林结构的图示。级联的每个级别包括两个随机森林（蓝色字体标出）和两个完全随机树木森林（黑色）。假设有三个类要预测; 因此，每个森林将输出三维类向量，然后将其连接以重新表示原始输入。

每个级是决策树森林的一个集合，即集成的集成（ensemble of ensembles）。我们包含了不同类型的森林来鼓励多样性，因为众所周知，多样性是集合结构的关键[Zhou，2012]。简单起见，在我们的实现中，我们使用了两个完全随机的树森林（complete-random tree forests）和两个随机森林[Breiman，2001]。每个完全随机的树森林包含1000个完全随机树[Liu et al。，2008]，通过随机选择一个特征在树的每个节点进行分割实现生成，树一直生长，直到每个叶节点只包含相同类的实例或不超过10个实例。类似地，每个随机森林也包含1000棵树，通过随机选择√ d 数量的特征作为候选（d是输入特征的数量），然后选择具有最佳 gini 值的特征作为分割。每个森林中的树的数值是一个超参数，这将在2.3节中讨论。

给定一个实例，每个森林会通过计算在相关实例落入的叶节点处的不同类的训练样本的百分比，然后对森林中的所有树计平均值，以生成对类的分布的估计。如图2所示，其中红色部分突出了每个实例遍历到叶节点的路径。

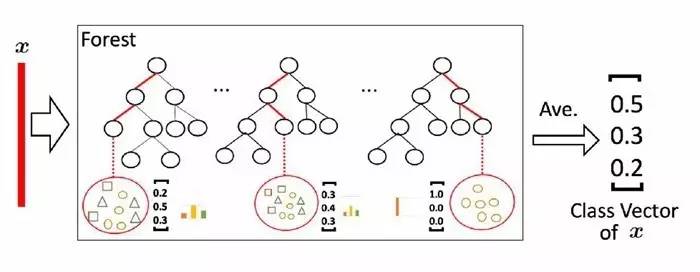


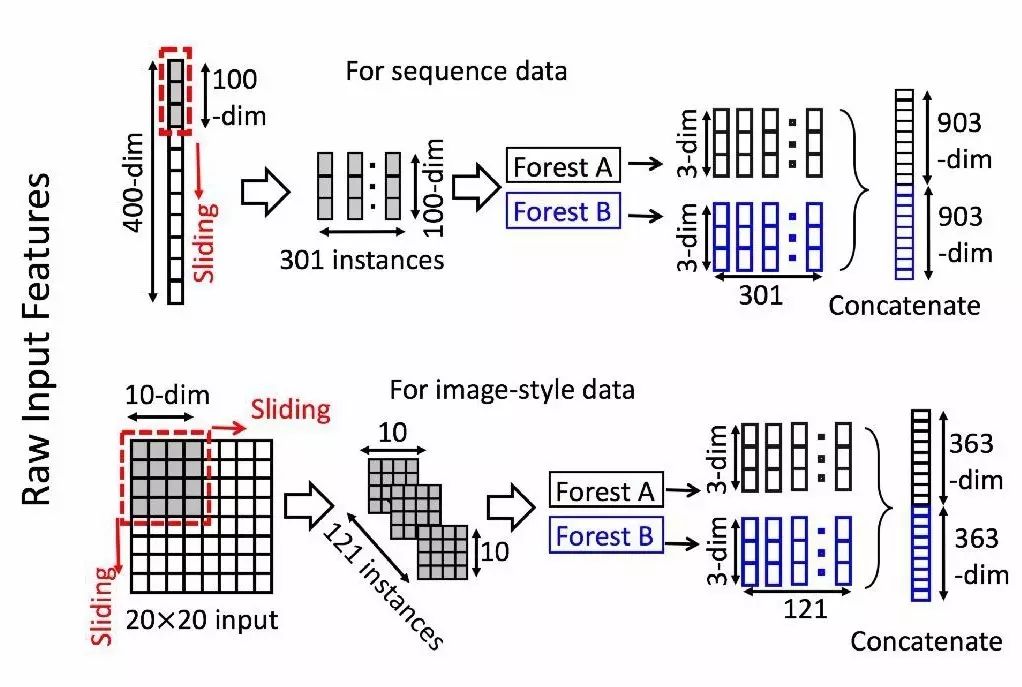
图2：类向量生成图示。叶节点中的不同标记表示了不同的类。

被估计的类分布形成类向量（class vector），该类向量接着与输入到级联的下一级的原始特征向量相连接。例如，假设有三个类，则四个森林每一个都将产生一个三维的类向量，因此，级联的下一级将接收12 = 3×4个增强特征（augmented feature）。

为了降低过拟合风险，每个森林产生的类向量由k折交叉验证（k-fold cross validation）产生。具体来说，每个实例都将被用作 k − 1 次训练数据，产生 k − 1 个类向量，然后对其取平均值以产生作为级联中下一级的增强特征的最终类向量。需要注意的是，在扩展一个新的级后，整个级联的性能将在验证集上进行估计，如果没有显着的性能增益，训练过程将终止；因此，级联中级的数量是自动确定的。与模型的复杂性固定的大多数深度神经网络相反，gcForest 能够适当地通过终止训练来决定其模型的复杂度。这使得 gcForest 能够适用于不同规模的训练数据，而不局限于大规模训练数据。

**2.2 多粒度扫描（Multi-Grained Scanning）**

深度神经网络在处理特征关系方面是强大的，例如，卷积神经网络对图像数据有效，其中原始像素之间的空间关系是关键的。（LeCun et al., 1998; Krizhenvsky et al., 2012），递归神经网络对序列数据有效，其中顺序关系是关键的（Graves et al., 2013; Cho et al.，2014）。受这种认识的启发，我们用多粒度扫描流程来增强级联森林。



**图3：使用滑动窗口扫描的特征重新表示的图示。 假设有三个类，原始特征是400-dim，滑动窗口是100-dim。**

如图3所示，滑动窗口用于扫描原始特征。假设有400个原始特征，并且使用100个特征的窗口大小。对于序列数据，将通过滑动一个特征的窗口来生成100维的特征向量；总共产生301个特征向量。如果原始特征具有空间关系，比如图像像素为400的20×20的面板，则10×10窗口将产生121个特征向量（即121个10×10的面板）。从正/负训练样例中提取的所有特征向量被视为正/负实例；它们将被用于生成如 2.1节中所说的类向量：从相同大小的窗口提取的实例将用于训练完全随机树森林和随机森林，然后生成类向量并连接为转换后的要素。如图3所示，假设有3个类，并且使用100维的窗口；然后，每个森林产生301个三维类向量，导致对应于原始400维原始特征向量的1,806维变换特征向量。

图3仅仅表示了一个滑动窗口的大小。通过使用多个尺寸的滑动窗口，最终的变换特征矢量将包括更多的特征，如图4所示。

**实验及结果**

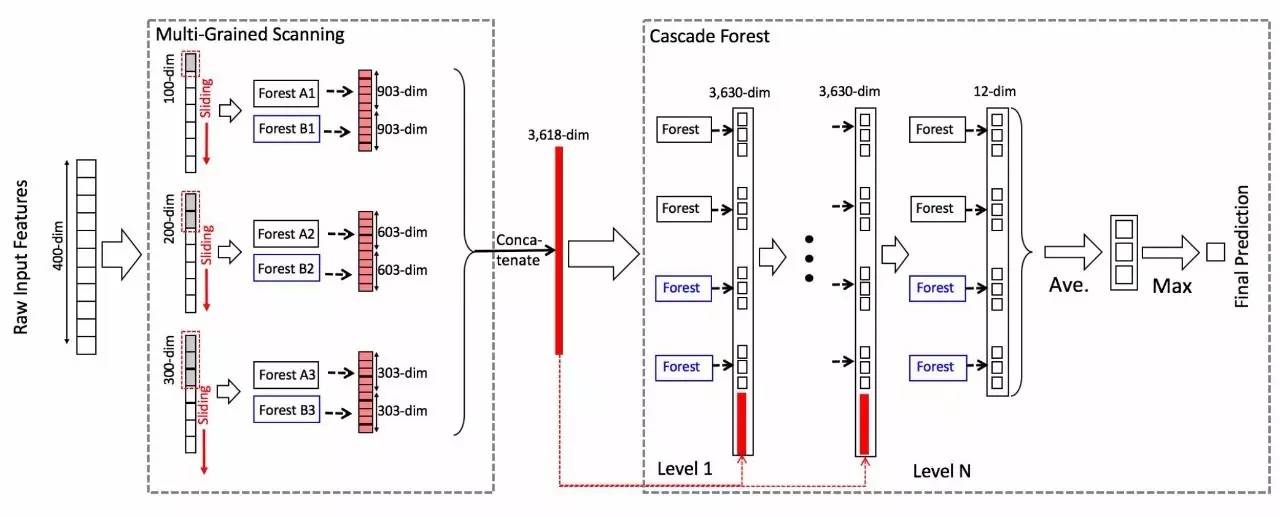
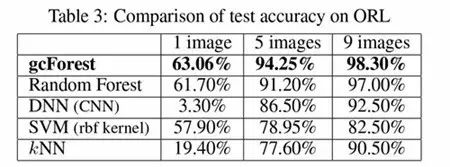
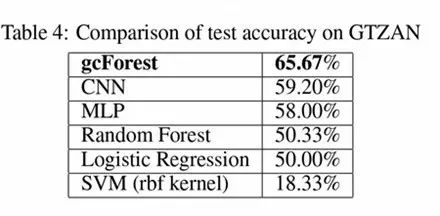


图4：gcForest的整体流程。假设有三个类要预测，原始特征是400-dim，使用了三个尺寸的滑动窗口。

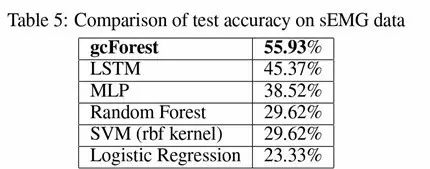
**人脸识别结果对比**



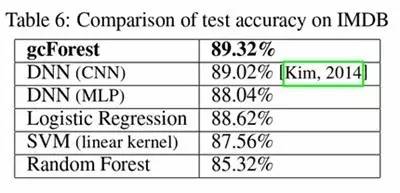
**音乐分类结果对比**



**手部运动识别结果对比**



**情感分类结果对比**



**结论**

带着深度学习的关键在于特征学习和巨大模型的能力这一认识，我们在本文中试图赋予树集成这些属性，并提出了 gcForest 方法。与深度神经网络相比，gcForest在我们的实验中表现了极高的竞争力或更好的性能。更重要的是，gcForest 具有少得多的超参数，并且对参数设置不太敏感；实际上在我们的实验中，通过使用相同的参数设置在不同的域中都获得了优异的性能，并且无论是大规模还是小规模的数据，它的工作都很好。此外，作为一种基于树的方法，gcForest 应该比深度神经网络更容易进行理论分析，不过这超出了本文的讨论范围。我们很快会提供 gcForest 的代码。

结论

作为结尾，请允许我们直接使用周志华和冯霁在论文中的最后一段：

“构建深度森林还存在其他可能性。作为一个会议论文，我们只朝这个方向进行了一点点探索。如果我们有更强大的计算设施，我们想尝试大数据和深度森林，这将留待以后讨论。原则上，深度森林应该能够展示出深度神经网络的其他能力，如充当特征提取器或预训练模型。 值得一提的是，为了解决复杂的任务，学习模型可能需要更深入。然而，当前的深度模型总是神经网络。本文说明了如何构建深度森林，我们相信它是一扇门，可能替代深度神经网络的许多任务。”

**论文地址：https://arxiv.org/abs/1702.08835**