常用测试集带来过拟合？你真的能控制自己不根据测试集调参吗

机器之心　今天

选自arXiv

机器之心编译

在验证集上调优模型已经是机器学习社区通用的做法，虽然理论上验证集调优后不论测试集有什么样的效果都不能再调整模型，但实际上模型的超参配置或多或少都会受到测试集性能的影响。因此研究社区可能设计出只在特定测试集上性能良好，但无法泛化至新数据的模型。本论文通过创建一组真正「未见过」的同类图像来测量 CIFAR-10 分类器的准确率，因而充分了解当前的测试集是否会带来过拟合风险。

1 引言

过去五年中，机器学习成为一块实验田。受深度学习研究热潮的驱动，大量论文围绕这样一种范式——新型学习技术出现的主要依据是它在多项关键基准上的性能提升。同时，很少有人解释为什么这项技术是对先前技术的可靠改进。研究者对研究进展的认知主要依赖于少量标准基准，如 CIFAR-10、ImageNet 或 MuJoCo。这就引出了一个关键问题：

目前机器学习领域衡量研究进展的标准有多可靠？

对机器学习领域的进展进行恰当评估是一件非常精细的事情。毕竟，学习算法的目标是生成一个可有效泛化至未见数据的模型。由于通常无法获取真实数据的分布，因此研究人员转而在测试集上评估模型性能。只要不利用测试集来选择模型，这就是一种原则性强的评估方案。

不幸的是，我们通常只能获取具备同样分布的有限新数据。现在大家普遍接

受在算法和模型设计过程中多次重用同样的测试集。该实践有很多例子，包括一篇论文中的调整超参数（层数等），以及基于其他研究者的研究构建模型。尽管对比新模型与之前模型的结果是非常自然的想法，但很明显当前的研究方法论削弱了一个关键假设：分类器与测试集是独立的。这种不匹配带来了一种显而易见的危险，研究社区可能会轻易设计出只在特定测试集上性能良好，但无法泛化至新数据的模型 [1]。

1.1 在 CIFAR-10 上的复现性研究

为了了解机器学习当前进展的可靠性，本文作者设计并实施了一种新型复现性研究。主要目标是衡量现在的分类器泛化至来自同一分布的未见数据的性能。研究者主要使用标准 CIFAR-10 数据集，因为它的创建过程是透明的，尤其适合这项任务。此外，近十年的大量研究使用 CIFAR-10。由于该过程的竞争性本质，这是一项调查适应性（adaptivity）是否导致过拟合的优秀测试用例。

该研究分为三步：

1. 首先，研究者创建一个新的测试集，将新测试集的子类别分布与原始 CIFAR-10 数据集进行仔细匹配。

2. 在收集了大约 2000 张新图像之后，研究者在新测试集上评估 30 个图像分类模型的性能。结果显示出两个重要现象。一方面，从原始测试集到新测试集的模型准确率显著下降。例如，VGG 和 ResNet 架构 [7, 18] 的准确率从 93% 下降至新测试集上的 85%。另一方面，研究者发现在已有测试集上的性能可以高度预测新测试集上的性能。即使在 CIFAR-10 上的微小改进通常也能迁移至留出数据。

3. 受原始准确率和新准确率之间差异的影响，第三步研究了多个解释这一差距的假设。一种自然的猜想是重新调整标准超参数能够弥补部分差距，但是研究者发现该举措的影响不大，仅能带来大约 0.6% 的改进。尽管该实验和未来实验可以解释准确率损失，但差距依然存在。

总之，研究者的结果使得当前机器学习领域的进展意味不明。适应 CIFAR-10 测试集的努力已经持续多年，模型表现的测试集适应性并没有太大提升。顶级模型仍然是近期出现的使用 Cutout 正则化的 Shake-Shake 网络 [3, 4]。此外，该模型比标准 ResNet 的优势从 4% 上升至新测试集上的 8%。这说明当前对测试集进行长时间「攻击」的研究方法具有惊人的抗过拟合能力。

但是该研究结果令人对当前分类器的鲁棒性产生质疑。尽管新数据集仅有微小的分布变化，但广泛使用的模型的分类准确率却显著下降。例如，前面提到的 VGG 和 ResNet 架构，其准确率损失相当于模型在 CIFAR-10 上的多年进展 [9]。注意该实验中引入的分布变化不是对抗性的，也不是不同数据源的结果。因此即使在良性设置中，分布变化也对当前模型的真正泛化能力带来了严峻挑战。

4 模型性能结果

完成新测试集构建之后，研究者评估了多种不同的图像分类模型。主要问题在于如何对原始 CIFAR-10 测试集上的准确率和新测试集上的准确率进行比较。为此，研究者对机器学习研究领域中出现多年的多种分类器进行了实验，这些模型包括广泛使用的卷积网络（VGG 和 ResNet [7,18]）、近期出现的架构（ResneXt、PyramidNet、DenseNet [6,10,20]）、已发布的当前最优模型 Shake-Drop[21]，以及从基于强化学习的超参数搜索而得到的模型 NASNet [23]。此外，他们还评估了基于随机特征的「浅层」方法 [2,16]。总体来说，原始 CIFAR-10 测试集上的准确率的范围是 80% 到 97%。

对于所有深层架构，研究者都使用了之前在线发布的代码来实现（参见附录 A 的列表）。为了避免特定模型 repo 或框架带来的偏差，研究者还评估了两个广泛使用的架构 VGG 和 ResNet（来自于在不同深度学习库中实现的两个不同来源）。研究者基于随机特征为模型编写实现。

主要的实验结果见表 1 和图 2 上，接下来将介绍结果中的两个重要趋势，然后在第 6 部分中讨论结果。

	Original Accuracy	New Accuracy	Gap	Δ Rank
shake_shake_64d_cutout [3, 4]	97.1 [96.8, 97.4]	93.0 [91.8, 94.0]	4.1	0
shake_shake_96d [4]	97.1 [96.7, 97.4]	91.9 [90.7, 93.1]	5.1	-2
shake_shake_64d [4]	97.0 [96.6, 97.3]	91.4 [90.1, 92.6]	5.6	-2
wide_resnet_28_10_cutout [3, 22]	97.0 [96.6, 97.3]	92.0 [90.7, 93.1]	5	+1
shake_drop [21]	96.9 [96.5, 97.2]	92.3 [91.0, 93.4]	4.6	+3
shake_shake_32d [4]	96.6 [96.2, 96.9]	89.8 [88.4, 91.1]	6.8	-2
darc [11]	96.6 [96.2, 96.9]	89.5 [88.1, 90.8]	7.1	-4
resnext_29_4x64d [20]	96.4 [96.0, 96.7]	89.6 [88.2, 90.9]	6.8	-2
pyramidnet_basic_110_270 [6]	96.3 [96.0, 96.7]	90.5 [89.1, 91.7]	5.9	+3
resnext_29_8x64d [20]	96.2 [95.8, 96.6]	90.0 [88.6, 91.2]	6.3	+3
wide_resnet_28_10 [22]	95.9 [95.5, 96.3]	89.7 [88.3, 91.0]	6.2	+2
pyramidnet_basic_110_84 [6]	95.7 [95.3, 96.1]	89.3 [87.8, 90.6]	6.5	0
densenet_BC_100_12 [10]	95.5 [95.1, 95.9]	87.6 [86.1, 89.0]	8	-2
neural_architecture_search [23]	95.4 [95.0, 95.8]	88.8 [87.4, 90.2]	6.6	+1
wide_resnet_tf [22]	95.0 [94.6, 95.4]	88.5 [87.0, 89.9]	6.5	+1
resnet_v2_bottleneck_164 [8]	94.2 [93.7, 94.6]	85.9 [84.3, 87.4]	8.3	-1
vgg16_keras [14, 18]	93.6 [93.1, 94.1]	85.3 [83.6, 86.8]	8.3	-1
resnet_basic_110 [7]	93.5 [93.0, 93.9]	85.2 [83.5, 86.7]	8.3	-1
resnet_v2_basic_110 [8]	93.4 [92.9, 93.9]	86.5 [84.9, 88.0]	6.9	+3
resnet_basic_56 [7]	93.3 [92.8, 93.8]	85.0 [83.3, 86.5]	8.3	0
resnet_basic_44 [7]	93.0 [92.5, 93.5]	84.2 [82.6, 85.8]	8.8	-3
vgg_15_BN_64 [14, 18]	93.0 [92.5, 93.5]	84.9 [83.2, 86.4]	8.1	+1
resnet_preact_tf [7]	92.7 [92.2, 93.2]	84.4 [82.7, 85.9]	8.3	0
resnet_basic_32 [7]	92.5 [92.0, 93.0]	84.9 [83.2, 86.4]	7.7	+3
cudaconvnet [13]	88.5 [87.9, 89.2]	77.5 [75.7, 79.3]	11	0
random_features_256k_aug [2]	85.6 [84.9, 86.3]	73.1 [71.1, 75.1]	12	0
random_features_32k_aug [2]	85.0 [84.3, 85.7]	71.9 [69.9, 73.9]	13	0
random_features_256k [2]	84.2 [83.5, 84.9]	69.9 [67.8, 71.9]	14	0
random_features_32k [2]	83.3 [82.6, 84.0]	67.9 [65.9, 70.0]	15	-1
alexnet_tf	82.0 [81.2, 82.7]	68.9 [66.8, 70.9]	13	+1

表 1：在原始 CIFAR-10 测试集和新测试集上的模型准确率，其中 Gap 表示两个准确率之间的差距。Δ Rank 是从原始测试集到新测试集的排名的相对变化。例如，Δ Rank = −2 表示模型在新测试集中的准确率排名下降了两位。

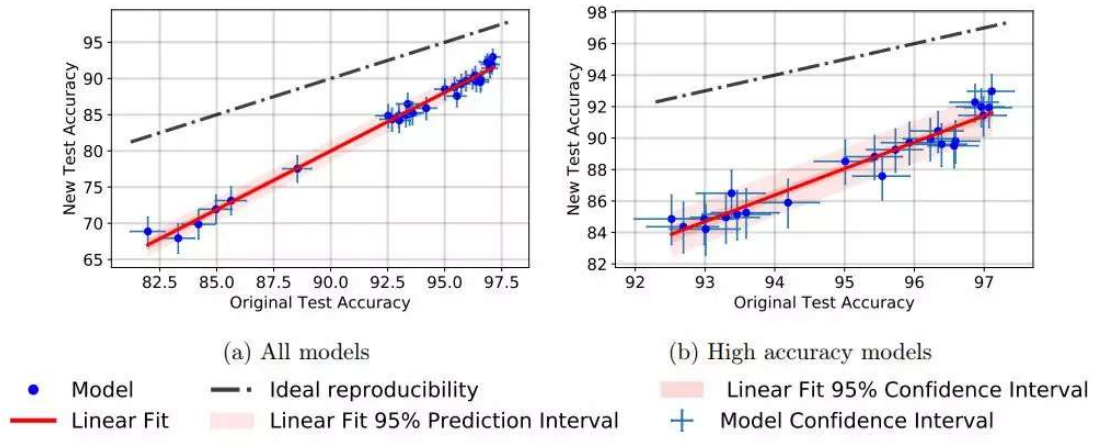

图 2：新测试集上的模型准确率 vs 原始数据集上的模型准确率。

4.1 准确率显著下降

所有模型在新测试集上的准确率都有显著的下降。对于在原始测试集上表现较差的模型，这个差距更大；对于在原始测试集上表现较好的模型，这个差距较小。例如，VGG 和 ResNet 架构的原始准确率（约 93%）和新准确率（约 85%）的差距大约为 8%。最佳准确率由 shake_shake_64d_cutout 得到，其准确率大致下降了 4%（从 97% 到 93%）。虽然准确率下降幅度存在变化，但没有一个模型是例外。

关于相对误差，拥有更高原始准确率的模型的误差可能有更大的增长。某些模型例如 DARC、shake_shake_32d 和 resnext_29_4x64d 在误差率上有 3 倍的增长。对于较简单的模型例如 VGG、AlexNet 或 ResNet，相对误差增长在 1.7 倍到 2.3 倍之间。参见附录 C 中的全部相对误差的表格。

4.2 相对顺序变化不大

按照模型的新旧准确率顺序对其进行分类时，总体排序结果差别不大。具有类似原始准确率的模型往往出现相似的性能下降。实际上，如图 2 所示，从最小二乘法拟合中派生出的线性函数可以对新旧准确率之间的关系做出很好的解释。模型的新准确率大致由以下公式得出：

$$\mathrm{acc}_{\mathrm{new}} = (1.62 \pm 0.04) \cdot \mathrm{acc}_{\mathrm{orig}} - 65.51\% \pm 3.16\%$$

另一方面，值得注意的是一些技术在新测试集上有了持续的大幅提升。例如，将 Cutout 数据增强 [3] 添加到 shake_shake_64d 网络，在原始测试集上准确率只增加了 0.12%，而在新测试集上准确率增加了大约 1.5%。同样，在 wide_resnet_28_10 分类器中添加 Cutout，在原始测试集上准确度提高了约 1%，在新测试集上提高了 2.2%。在另一个例子里，请注意，增加 ResNet 的宽度而不是深度可以为在新测试集上的性能带来更大的好处。

4.3 线性拟合模型

尽管图 2 中观察到的线性拟合排除了新测试集与原始测试集分布相同的可能

性，但新旧测试误差之间的线性关系仍然非常显著。对此有各种各样的合理解释。例如，假设原始测试集由两个子集组成。在「easy」子集上，分类器达到了 a_0 的精度。「hard」子集的难度是κ倍，因为这些例子的分类误差是κ倍。因此，该子集的精度为 1 − κ(1 − a_0)。如果这两个子集的相对频率是 p_1 和 p_2，可以得到以下总体准确率：

$$\mathrm{acc}_{\mathrm{orig}} = p_1 \cdot a_0 + p_2 \cdot (1 - \kappa(1 - a_0))$$

可以重写为 a_0 的简单线性函数：

$$\mathrm{acc}_{\mathrm{orig}} = \beta \cdot a_0 + \gamma$$

对于新的测试集，研究者也假设有由不同比例的两个相同分量组成的混合分布，相对频率现在是 q_1 和 q_2。然后，可以将新测试集上的准确率写为：

$$\begin{aligned} \mathrm{acc}_{\mathrm{new}} &= q_1 \cdot a_0 + q_2 \cdot (1 - \kappa(1 - a_0)) \\ &= \beta' \cdot a_0 + \gamma' \end{aligned}$$

此处像之前一样把项集合成一个简单的线性函数。

现在很容易看出，新的准确率实际上是原始准确率的线性函数：

$$\begin{aligned} \mathrm{acc}_{\mathrm{new}} &= \frac{\beta'}{\beta}(\beta a_0 + \gamma) - \frac{\beta'}{\beta}\gamma + \gamma' \\ &= \frac{\beta'}{\beta}\mathrm{acc}_{\mathrm{orig}} \ . \end{aligned}$$

研究人员注意到，这种混合模型并不是一种真实的解释，而是一个说明性的例子，说明原始和新的测试准确率之间的线性相关性是如何在数据集之间的小分布移位下自然产生的。实际上，两个测试集在不同的子集上具有不同准确率的更复杂的组成。尽管如此，该模型揭示了即使分类器的相对排序保持不变，分布移位也可能存在令人惊讶的敏感性。研究人员希望这种对分布偏移的敏感性能够在之后的研究中得到实验验证。

5. 解释差异

为了解释新旧准确率之间的巨大差距，研究人员探究了多种假设（详见原文）。

- 统计误差
- 近似重复移除的差异
- 超参数调整
- 检测高难度图像
- 在部分新测试集上进行训练
- 交叉验证

	vgg_15_BN_64	wide_resnet_28_10	shake_shake_64d_cutout
Split 1	93.87	96.16	97.16
Split 2	93.81	96.04	97.3
Split 3	94.01	96.37	97.38
Split 4	93.99	96.16	97.39
Split 5	93.5	96.5	97.4

表 2：交叉验证拆分的模型准确率。

6 讨论

过拟合：实验是否显示出过拟合？这是解释结果时的主要问题。简单来说，首先定义过拟合的两个概念：

- 训练集过拟合。过拟合的一个概念是训练准确率和测试准确率之间的差异。请注意，本研究的实验中的深度神经网络通常达到 100% 的训练准确率。所以这个过拟合的概念已经出现在已有数据集上了。
- 测试集过拟合。过拟合的另一个概念是测试准确率和潜在数据分布准确率之间的差距。通过使模型设计选择适应测试集，他们担心的是这将隐性地使模型适应测试集。测试准确率随后失去了对真正未见过数据准确性进行测量的有效性。

由于机器学习的整体目标是泛化到未见过的数据，研究者认为通过测试集适应性实现的第二种过拟合更重要。令人惊讶的是，他们的研究结果显示在 CIFAR-10 并没有这种过拟合的迹象。尽管在该数据集上具有多年的竞争适应性，但在真正的留出数据（held out data）上并没有停滞不前。事实上，在新测试集中，性能最好的模型比更成熟的基线有更大的优势。尽管这一趋势与通过适应性实现过拟合所暗示的相反。虽然最终的结果需要进一步的复制实验，但研究者认为他们的结果支持基于竞争的方法来提高准确率。

研究者注意到 Blum 和 Hardt 的 Ladder 算法分析可以支持这一项声明 [1]。事实上，他们表明向标准机器学习竞赛中加入一些小修改就能避免这种程度的过拟合，即通过激进的适应性导致过拟合。他们的结果表明即使没有这些修改，基于测试误差的模型调优也不会在标准数据集上产生过拟合现象。

分布转移（distribution shift）。尽管研究者的结果并不支持基于适应性的过拟合假设，但仍需要解释原始准确率和新准确率之间的显著性差异。他们认为这种差异是原始 CIFAR-10 数据集与新的测试集之间小的分布转移造成的。尽管研究者努力复制 CIFAR-10 数据集的创建过程，但它和原始数据集之间的差距还是很大，因此也就影响了所有模型。通常可以通过对数据生成

过程中的特定变换（如光照条件的改变），或用对抗样本进行攻击来研究数据分布的转移。本研究的实验更加温和而没有引起这些挑战。尽管如此，所有模型的准确率都下降了 4-15%，对应的误差率增大了 3 倍。这表明目前 CIFAR-10 分类器难以泛化到图像数据的自然变化。

论文：Do CIFAR-10 Classifiers Generalize to CIFAR-10?

Do CIFAR-10 Classifiers Generalize to CIFAR-10?

Benjamin Recht
UC Berkeley

Rebecca Roelofs
UC Berkeley

Ludwig Schmidt
MIT

Vaishaal Shankar
UC Berkeley

June 4, 2018

论文地址：https://arxiv.org/abs/1806.00451

摘要：目前大部分机器学习做的都是实验性的工作，主要集中在一些关键任务的改进上。然而，性能最好的模型所具有的令人印象深刻的准确率令人怀疑，因为多年来一直使用相同的测试集来选择这些模型。为了充分了解其中的过拟合风险，我们通过创建一组新的真正未见过的图像来测量 CIFAR-10 分类器的准确率。尽管确保了新的测试集尽可能接近原始数据分布，但我们发现，很多深度学习模型的准确率下降很大（4% 到 10%）。然而，具有较高原始准确率的较新模型显示出较小的下降和较好的整体性能，这表明这种下降可能不是由基于适应能力的过拟合造成的。相反，我们认为我们的结果表明了当前的准确率是脆弱的，并且容易受到数据分布中微小自然变化的影响。

写留言