摘要

我们提出了一个新颖的框架，用于扩增机器学习的反例数据集。反例是分类错误的示例，对重新训练和改进模型具有重要意义。我们框架的关键组件包括一个反例生成器，它生成由模型和错误表错误分类的数据项。错误表是一种新颖的数据结构，用于存储与错误分类有关的信息。它可用于解释模型的漏洞，并用于有效地生成扩增的反例。我们将提出的框架与经典的扩增技术进行了比较——基于深度神经网络的自动驾驶中目标检测的案例研究，从而证明了该框架的有效性。

介绍

由机器学习算法(尤其是深度神经网络)生成的模型正被部署在值得高度关注的领域中，需要更高的准确性和保证力。然而使用深度学习来学习高精度模型受到大量数据需求的限制，甚至进一步受到劳动密集型标签的需求的限制。数据扩增通过使用保留标签的变换扩大训练集来克服数据的不足。传统的数据扩增方案涉及几何变换，这些变换会改变图像的几何形状(例如，旋转，缩放，裁切或翻转)，以及改变颜色通道的光度转换。这些技术的有效性最近已得到证明。像前面提到的方法一样，传统的扩增方案将数据添加到训练集中，希望提高模型的准确性，而无需考虑模型已经学习了哪些特征。最近，一种复杂的数据扩增技术被提出，该技术使用生成对抗网络(一种能够生成合成数据的特殊类型的神经网络)来充实训练集。也有诸如hard negative mining之类的扩增技术，它们以有针对性的否定示例充实训练集，目的是减少误报。

在这项工作中，我们提出了一种新的扩增方案，即以反例为指导的数据扩增。主要思想是仅使用新的分类错误的示例来扩增训练集，而不是使用原始训练集中的修改后的图像。提出的扩增方案包括以下步骤：1)生成被模型误分类的合成图像，即反例；2)将反例添加到训练集中；3)在扩增数据集上训练模型。可以重复这些步骤，直到达到所需的精度为止。请注意，我们的扩增方案取决于生成错误分类图像的能力。因此，我们开发了一种与采样器配合使用的图像生成器，以生成作为模型输入提供的图像。这些图像以一种能够自动添加地面真实值标签的方式生成。分类错误的图像构成了增加集并被添加到训练集。除图片外，图像生成器还提供有关分类错误的图像的信息，例如元素的位置，亮度，对比度等。此信息可用于查找反例中经常出现的特征。我们在称为“错误表”的数据结构中收集有关反例的信息。错误表对于提供有关反例的说明以及查找可能导致图像分类错误的重复模式非常有用。错误表分析还可以用于生成可能是反例的图像，从而有效地构建扩增集。

主要贡献如下：

以反例为指导的数据扩增方法，仅将错误分类的示例迭代添加到训练集中；

合成图像生成器，提供逼真的示例；

错误表存储有关反例的信息，它的分析提供了解释，促进生成反例图像。

我们通过分析不同的反例数据扩增采样方案，在卷积神经网络(CNN)上进行了对象检测实验，并将所提出的方法与经典数据扩增进行了比较。我们的实验展示了使用反例驱动的方法来对抗传统方法的好处。主要的改进是反例扩增集包含了模型无法从训练集中学习的信息，而经典扩增方案并未考虑这个方法。我们使用由图像生成器生成的合成数据集进行实验，这样可以确保所有处理的数据都来自同一分布。

总览

图 1 总结了提出的反例指导的扩增方案。该程序将修改空间 M(即图像生成器可能配置的空间)作为输入。基于领域知识将空间 M 构造为“语义修改”空间；即，每个修改都必须在使用机器学习模型的应用领域中具有含义。这使我们能够执行更多有意义的数据扩增而不仅仅是通过扰动输入向量(例如，在图像中对抗性选择和修改少量像素值)进行对抗性数据生成。

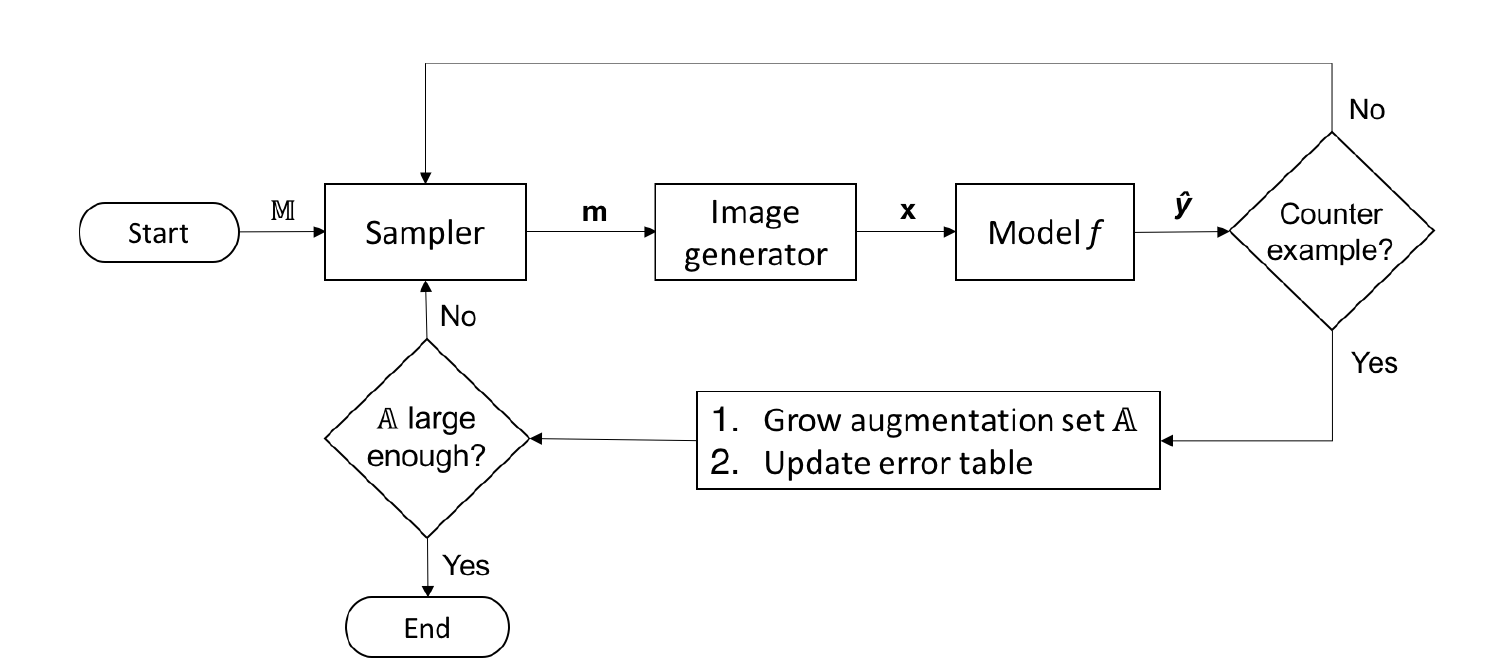


图 1:反例指导的扩增方案

在每个循环中，采样器都会从 M 中选择一个修改 m。采样是通过一种可以由预计算误差表(该数据结构存储有关模型误分类的图像信息的数据结构)所偏向的采样方法确定的。图像生成器将采样的修改呈现为图片 x。图像 x 被提供作为返回预测 y 的模型 f 的输入。我们检查 x 是否为反例也就是 y 是否错误。如果是，我们将 x 添加到扩增集 A 中，并将 x 的信息(例如 m，y)存储在误差表中，供采样器在下一次迭代中使用。重复循环，直到扩增集 A 足够大(或已充分覆盖 M)。

该方案返回一个扩增集和一个误差表，该扩增集将用于重新训练已处理的模型，该误差表的分析可以确定反例之间的共同特征，并帮助采样器选择候选反例。

图像生成器

我们的反例扩增方案的核心是一个图像生成器，可渲染逼真的道路场景的合成图像。由于反例是由合成数据生成器生成的，因此我们对生成的数据的基本事实标签有充分的了解。例如，在我们的案例中，当图像生成器将汽车放置在特定位置时，我们确切地知道了它的位置和大小，从而相应地确定了地面真实边界框。

抽样方法

采样器的目的是提供对修改空间的良好覆盖，并确定其具体化导致反例的样本。以下简要介绍一些我们集成到框架中的采样方法：

均匀的随机采样：均匀的随机采样可确保从 M 采样任何点的概率均等，从而保证了训练和测试过程中生成图像的良好混合。尽管这是一种既简单又有效的训练和测试技术，但它可能无法很好地覆盖修改空间。

低差异采样：低差异序列是 n 元组序列，比不相关的随机点更均匀地填充 nD 空间。低差异序列通过减少间隙和点的聚类来覆盖检测框，从而确保对样本空间的均匀覆盖。

交叉熵抽样：交叉熵方法是一种用来组合优化和重要性抽样的通用蒙特卡洛方法。这是一种迭代采样技术，我们从给定的概率分布中采样，然后通过最小化交叉熵来更新分布。

我们的实验使用 Halton 序列。在优化覆盖上有两个优点：首先，我们增加了快速发现反例的机会，其次，这组反例将具有很高的多样性。这意味着具体的图像看起来会有所不同，因此该模型将学习各种新功能。

错误表

扩增方案的每次迭代都会产生一个示例，其中包含指向学习模型局限性的信息。希望提取与反例相关的模式，并使用此信息有效地生成新的反例。因此，我们将错误表定义为数据结构，据列由跨生成的图像的重要特征组成。错误表分析可用于：

1.提供有关反例的说明，

2.生成反馈以采样新的反例。

在第一种情况下，通过在各个示例中找到共同的模式，我们向用户提供了反馈，例如“该模型未检测到在森林道路上驶离我们的白色汽车”；在第二种情况下，我们可以使采样器的修改偏向于更可能导致反例。

实验评估

在所有实验中，我们分析了 squeezeDet，这是一种用于自动驾驶的 CNN 实时物体检测器。所有模型都接受了 65 个时期的训练。

原始训练和测试集 X 和 T 分别包含由我们的图像生成器随机生成的 1500 和 750 张图片。初始精度 accfx(T)=(0.9847，0.9843)相对较高。但是，我们能够在准确度较低的训练集上生成与 T 一样大的反例集，表格中突出显示的条目显示最佳性能。报告的值是五个不同实验的平均值。

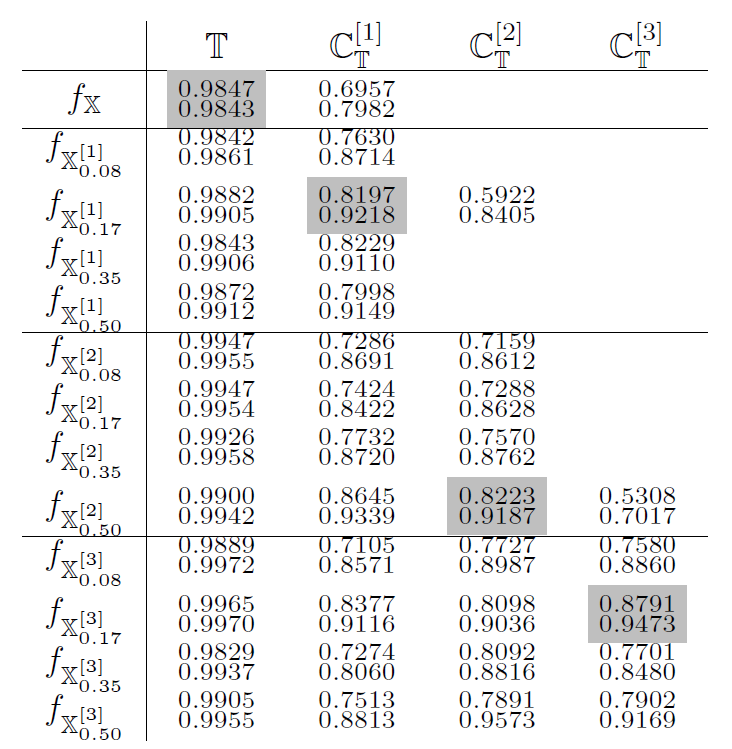


表 1：扩增循环

结论

在本文中，我们提出了一种通过反例扩增机器学习(ML)数据集的技术。我们生成的示例是由 M\* 模型错误分类的综合生成的数据项。由于这些项目是通过算法合成的，因此它们的地面真实标签也将自动生成。我们展示了如何使用错误表来有效地指导扩增过程。训练深度神经网络的结果表明，我们的扩增技术在图像分类方面比标准扩增方法表现更好。此外，当我们迭代扩增循环时，在计算上很难找到反例。我们还表明，错误表可以通过较小的数据扩增有效地提高准确性。

我们注意到，我们提出的方法还可以扩展到“系统级”分析和验证中使用反例，可以在周围环境的背景下分析整个系统的正确性。使用此类“语义反例”执行数据扩增是未来工作的有趣方向。

我们的方法可以看作是反例指导的归纳综合(CEGIS)的实例，它是程序综合的常见范例。我们的案例是 M\* 模型。CEGIS 本身是 Oracle 引导归纳综合的特例。未来，探索使用除产生反例的 Oracle 以外的其他 Oracle 来扩增数据集，并将我们的以反例为指导的数据扩增技术与其他以 oracle 为指导的数据扩增方法进行比较将是很有意思的。

最后，在这项工作中，我们决定完全依赖模拟的综合数据，以确保训练，测试和反例集来自同一数据源。将我们的扩增方法扩展到现实世界中的数据将会很有趣。例如，在驾驶过程中收集的道路场景图像。因此需要使用诸如领域适应或迁移学习之类的技术，它们可以使新生成的数据适应现实世界。

引用

Dreossi T, Ghosh S, Yue X, et al. Counterexample-guided data augmentation[J]. arXiv preprint arXiv:1805.06962, 2018.

致谢

本文由南京大学软件学院 2019 级硕士张朱佩田翻译转述。