

硕 士 研 究 生 读 书 报 告



题目 关于自动数据增强的方法

作者姓名 陈文儒

作者学号 21951428

指导教师 张志猛

学科专业 软件工程

所在学院 软件学院

提交日期 2020年10月1日

Methods of automatic data agument

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Zhimeng Zhang

By

Wenru Chen

Zhejiang University, P.R. China

2020

摘要

数据增强首先出现在扩充样本的情况。我们常常会碰到数据不足的情况，对于一个深层的网络模型来说，一点点的数据是完全不够的，这样会造成欠拟合或者过拟合。目前比较好的网络都是用大量数据喂出来的。而且大量数据集训练出来的效果也更好。数据增强包括旋转、翻转、颜色变换、噪声注入、cut、过滤器、混合图像、缩放等等。但无论怎么变化，这些都是人为的方法，其对模型的影响还是未知。因此有必要去发现自动化的数据增强，这种增强方式，应该是像我们模型一样，能够训练的，能最终达到提升模型精度的方法。

**关键词**：深度学习， 神经网络，数据增强，自动

Abstract

Data enhancement first appears in the case of expanding samples. We often encounter the situation of insufficient data. For a deep network model, a little data is not enough, which will cause under fitting or over fitting. At present, the better networks are fed with a lot of data. And a large number of data sets training results are also better. Data enhancement includes rotation, flipping, color transformation, noise injection, cut, filter, mixed image, scaling and so on. However, no matter how they are changed, these are all artificial methods, and their impact on the model is still unknown. Therefore, it is necessary to find automatic data enhancement. This enhancement method should be the same as our model, which can be trained and ultimately improve the accuracy of the model.

**Keywords：**deep learning, neural network, data augment, auto

1引言

数据增强首先出现在扩充样本的情况。我们常常会碰到数据不足的情况，对于一个深层的网络模型来说，一点点的数据是完全不够的，这样会造成欠拟合或者过拟合。目前比较好的网络都是用大量数据喂出来的。而且大量数据集训练出来的效果也更好。数据增强包括旋转、翻转、颜色变换、噪声注入、cut、过滤器、混合图像、缩放等等。但无论怎么变化，这些都是人为的方法，其对模型的影响还是未知。

此外数据增强不仅被运用到增加样本上，无监督学习也是其中的应用领域，例如对比学习（contrastive learning）中生成相似的样本。

数据增强技术是提高现代图像分类精度的有效手段。然而，当前的数据扩充实现是手工设计的。因此，有人提出，能否用自动化的方式，自动进行数据增强。所以，本文介绍了三篇有关数据增强的方式，目的是发现一种自动化的能提高模型效率的真实可靠的数据增强技术。

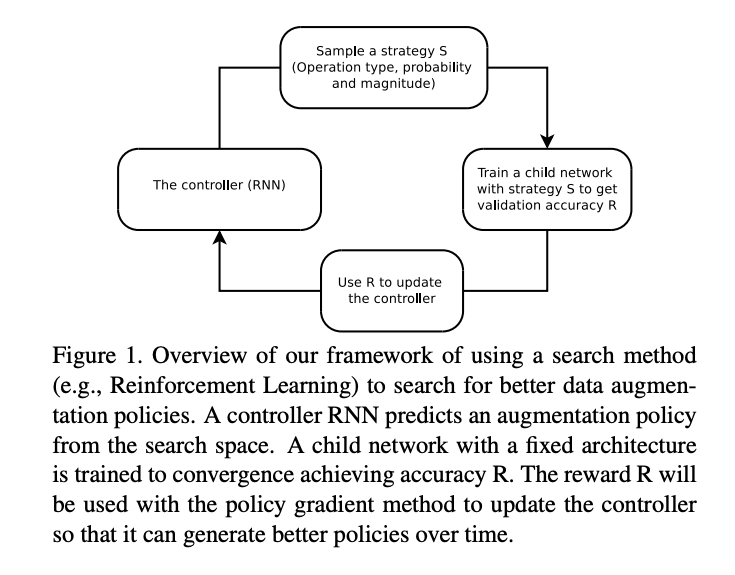
**2 开山之作——autoaugment**

在这篇论文中，作者描述了一个叫做AutoAugment的简单过程来自动搜索改进的数据扩充策略。在实现中，作者设计了一个搜索空间，其中一个策略由许多子策略组成，每个batch中的每个图像随机选择一个子策略。子策略由两个操作组成，每个操作是图像处理函数，如平移、旋转或剪切，以及应用这些函数的概率和大小。作者使用搜索算法来寻找最佳策略，使得神经网络对目标数据集产生最高的验证精度。作者的方法在CIFAR-10、CIFAR-100、SVHN和ImageNet上达到了最先进的精度（无需额外数据）。

**2.1 方法总览**

作者将寻找最佳扩充策略的问题表述为一个离散搜索问题，如图所示。算法由两部分组成：搜索算法和搜索空间。在高层次上，搜索算法（具体实现为控制器RNN）对数据增强策略S进行采样，该策略包含关于要使用的图像处理操作、在每个批中使用该操作的概率以及操作的大小的信息。该方法的关键在于，使用策略S来训练具有固定结构的神经网络，其验证的精度R将被发回以更新控制器。由于R是不可微的，控制器将采用策略梯度法进行更新。

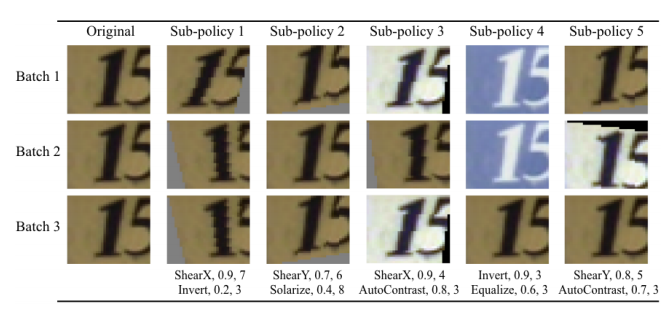
实际上采用了强化学习的方法。



**2.1 搜索方法**

在搜索空间中，一个策略由5个子策略组成，每个子策略由两个按顺序应用的图像操作组成。此外，每个操作还与两个超参数相关：1）应用操作的概率，2）操作的强度。

下图显示了一个在搜索空间中包含5个子策略的策略示例。第一个子策略指定ShearX的顺序应用，后跟Invert。应用ShearX的概率为0.9，当应用ShearX时，其幅值为7/10。然后应用概率为0.8的反转。反转操作不使用幅值信息。必须强调的是这些操作是按指定顺序应用的。



原文中搜索的操作有偏移X/Y、反转X/Y轴、旋转、自动对比、反转、均衡器、改变曝光、后处理、对比度、颜色、亮度、清晰度、剪切、样本组合。总共16中操作，于是就要在（16×10×11）^2个可能性的空间内寻找每个子策略成为一个搜索问题。然而，问题的目标是同时找到5个这样的子政策，以增加多样性。5个子策略的搜索空间大致为（16×10×11）^10≈2.9×10^32可能性。

**2.2 控制器RNN的训练**

控制器用一个奖励信号进行训练，这就是该策略在提高“子模型”（作为搜索过程的一部分训练的神经网络）泛化方面有多好。在实验中，作者预留了一个验证集来衡量子模型的泛化程度。子模型通过对训练集（不包含验证集）应用5个子策略生成的扩充数据进行训练。对于小批量中的每个示例，随机选择5个子策略中的一个来增强图像。然后在验证集上对子模型进行评估以衡量其准确性，并以此作为奖励信号来训练递归网络控制器。在每个数据集中，控制器采样约15000个策略。

**2.3 控制器结构**

控制器RNN是一个单层LSTM，每层有100个隐藏单元，以及与每个架构决策相关联的两个卷积单元（其中B通常为5）的2×5B 的softmax预测。控制器RNN的10B预测中的每一个都与一个概率相关联。子网络的联合概率是这些10B Softmax中所有概率的乘积。该联合概率用于更新控制器RNN的梯度。根据子网络的验证精度调整梯度以更新控制器RNN，使得控制器为坏子网络分配低概率，为好子网络分配高概率。本文采用了强化学习中的最近策略优化（PPO）。为了鼓励探索，作者还使用了权重为0.00001的熵惩罚。在实现中，基线函数是以前奖励的指数移动平均值，权重为0.95。控制器的权值在-0.1和0.1之间均匀初始化。

**3 改进的算法——fast-autoagument**

**3.1 总览**

本文是基于autoagument的改进，在速度上远远超过autoagument，精度上也不逊色。

其整体思想如下图所示，首先将一个训练集Dtrain 分割，将其分成k个包，在k个包里面分为两个另外的数据包。在一个分布式的集群上，分别训练K个模型，这样就能得到K个任务，以及结果。每个任务，都采样本身的数据，并随机做数据增强。训练一定轮数后，K个包最终能得到一个结果，这个结果将被收集起来，作为一个数据增强策略集合。最后，每个包里面的精度最高的前N个策略将被保留，最后再从这N个里面选出一定的数量作为最后的策略。

图示

描述已自动生成

**3.2 搜索空间**

设O是在输入图像空间X上定义的一组数据增强操作（图像变换）。每个操作O有两个参数：调用概率p和决定操作可变性的幅度λ。有些操作（例如反转、翻转）不使用幅值。设S是子策略集，其中子策略τ∈S由Nτ个连续操作组成，其中每个操作按顺序应用于输入图像，概率p如下：

图片包含 图表

描述已自动生成

**3.2 搜索策略**

在本文中，作者将增广策略的搜索看作一对训练数据集之间的密度匹配。设D是X×Y上的一个概率分布，并假设数据集D是从该分布中取样的。对于给定的分类模型M（·|θ）:X→Y，模型M（·|θ）在数据集D上的期望精度和期望损失分别用R（θ| D）和L（θ| D）表示。

**3.3 算法表示**

Fast-autoagument搜索期望的增广策略，将贝叶斯优化应用于分布式训练集分割。也就是说，整个搜索过程包括两个步骤：（1）使用默认的扩充规则对K-fold训练集数据进行模型参数训练；（2）使用HyperOpt搜索最优的扩充策略进行探索和开发。下面的伪代码程序大多是可并行化的，这使得所提出的方法在实际应用中更为有效。

Shuffle（第1行）：使用了sklearn中的StratifiedShuffleSplit方法分割训练集，同时保留每个类的样本百分比（分层洗牌）。

训练（第四行）：每个训练分段上的训练模型。如果计算资源足够，那么可以在多台机器上并行运行，以减少总运行时间。

Explore and Exploit（第6行）：使用来自Ray的HyperOpt库，搜索数量为B，最大并发计算为20。与AutoAugment不同，本文没有离散化搜索空间，因为本文的搜索算法可以处理连续值。作者探讨了概率p和幅值λ的一种可能的运算。概率和幅度的值从开始时的[0，1]均匀采样，然后HyperOpt调整这些值以优化目标L。

合并（第7-9行）：为每个拆分选择前N个最佳策略，然后将从所有拆分中获得的策略组合起来。这组最终策略用于重新培训。

文本, 信件

描述已自动生成

**3 回归到朴实——rand-agument**

Autoagument的作者（第二节），发现了，其实数据增强在哪个程度上能学习到更好的效果，用强化学习来做是耗时耗力的。而实际上，这种随机的学习方法是在很大程度上就是在做无用功。于是作者提出了一种非常非常简单的方法。完全随机选取图像的操作。

这里先放出来全部代码：

文本

描述已自动生成

是的没看错，这就是全部代码，去掉数据操作的枚举，代码只有两行。那么为什么这么随机能直接达到目标呢？

作者提过，目前的工作是在一个非常巨大的搜索空间进行学习的。大规模采用这些方法的一个障碍是它们需要一个单独且昂贵的搜索阶段。克服搜索阶段开销的常见方法是使用较小的代理任务。

然而，目前还不清楚代理任务上找到的优化超参数是否对实际任务也是最优的。在本文中，作者重新思考设计自动化增强策略的过程。作者发现，虽然先前的工作需要独立地搜索每个操作的幅度和概率，但是只搜索一个共同控制所有操作的单个失真幅度就足够了。因此，他们提出了一个简化的搜索空间，大大减少了自动扩充的计算开销，并允许删除一个单独的代理任务。

实际上本质上，还是要让操作足够随机，足够随机后产生的样本的分布才会和原本的样本相同，而从个体上来说不同。

**5 小结**

本文描述了近年来较为火热的自动数据增强技术。由于近年来深度神经网络的发展和演变，对数据要求越来越高，机器需要的数据要求也越来越高，而自动增强技术正好能满足这个需求，他能自动增广原始数据集，能达到原始数据1倍甚至好几倍的数据。

其实本文的阅读是为了最近火热的contrastive learning服务的，下面contrastive learning简称cl。cl的基本思想是，用一对相似的图像和其他不相似的图像通过对比学习，来达到无监督自学习的过程。

具体的做法是通过数据增强技术，生成对应相似的图像，而其他不适本身或者自身生成的图像，则认为不相似，通过最后判别网络模型提取出来的特征是否相似，最终得到一个cl的loss，通过这个loss来实现模型的学习。

而autoagument就是用这种自动增强的技术来生成cl中相似的图片。

总之自动数据增强还有很多应用，应该不断学习，达到最后的目的，为最后的落地应用服务。

参考文献

[1] Cubuk E D, Zoph B, Mane D, et al. Autoaugment: Learning augmentation policies from data[J]. arXiv preprint arXiv:1805.09501, 2018.

[2] Lim S, Kim I, Kim T, et al. Fast autoaugment[C]//Advances in Neural Information Processing Systems. 2019: 6665-6675.

[3] Cubuk E D, Zoph B, Shlens J, et al. Randaugment: Practical automated data augmentation with a reduced search space[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 702-703.