长尾识别中数据增强的丰富度与偏差的权衡（EIG）

摘要：

长尾场景下，数据增强对尾类进行丰富度和数量的增强，然而数据增强背后的机理尚未探究清晰，我们从数据增强导致的丰富度增益和分布偏移角度，提出有效信息增益（EIG），通过 EIG 对增强的数据进行过滤能进一步提升性能（以数据为中心）。

引言：

长尾学习解决方案（1）（2）

（1）以模型为中心

解耦训练策略：特征提取器（特征学习）、分类器（分类器学习），分成两阶段分别修正

加权损失函数：给予尾类更大权重的损失，使得模型更多的关注尾类以修正偏见

（2）以数据为中心

在线数据增强：诸如 cutmix、mixup、fmix、sumix 等在线数据混合增强

知识引导增强（信息增强）：从头类迁移知识到尾类、从外部先验迁移知识到尾类

随着基础模型的快速发展，以数据为中心的普适于模型的策略越来越重要

以数据为中心潜在的问题（丰富度增益与分布偏移的权衡）

（1）尾类样本足以代表真实分布：

在线数据增强方法，使用现有的数据进行混合扩充，因为尾类已经覆盖真实分布，所以会使得在流形空间中该类对应的流形区域中的点云从稀疏变得密集，但区域几乎不变

（2）尾类样本不足以代表真实分布（更符合实际）：

头引导尾类或外部先验引导尾类的时候，在流形空间的视角实际上是局部分布向真实分布逼近的过程，但是这个过程充满随机性，增强样本可能使分布偏移真实，增强方法依赖于经验知识和参数调整，如何定量选择高质量的增强样本——丰富度增益与分布偏移的权衡

有效信息量增益 EIG：

我们提出了一种独立于模型依赖于增强数据的度量 EIG，增强数据的 EIG 在适当的范围内时，信息增强方法的性能最大化。我们进一步探讨了数据不平衡与最优 EIG 之间的关系，同时提出了一种选择具有特定 EIG 水平的增强数据的算法，我们在各种最先进的信息增强方法上实验了我们的工作。

相关工作：

信息增强 Information Augmentation

尾类样本不足以代表真实分布的时候，以模型为中心和以数据为中心的在线数据增强都无法使得模型学习到观察分布外的信息，信息增强通过引入额外的知识促进不平衡学习，具体的为头类迁移信息到尾类，外部先验迁移到尾类

以数据为中心的性能分析：

我们从数据流形的角度讨论与信息增强方法的性能相关的潜在机制，然后我们引入有效信息增益（EIG）的概念来量化信息增强方法引起的丰富度增益和分布偏移，我们的指标是直接从数据分布计算出来的，不依赖于模型。

边际效应的启示

（1）三种分布的定义

定义 1（观测分布 Observed Distribution）。由可用样本形成的分布（即，训练集）。

定义 2(真实分布 True Distribution)。当一个类有足够的样本以完全包含该类的所有特征时，由样本形成的分布。
定义 3（潜在分布 Underlying distribution)。真实分布与观测分布之间的分布。
（2）流形分布定律 The Law of Manifold Distribution：
自然数据（图片、语音、文本等多模态数据）的高维嵌入空间的特征映射到低维后，不同类有不同的形状，同时同一类会聚集，不同类会分离。
（3）边界效应 marginal effect：
类的特征多样性是有限的，也就是丰富度是有限的，也就是类的流形是有边界。
通过信息增强生成的样本如果在边界外，那么可能是自然数据不存在的样本，或其他类的样本，后者会导致尾类的流形与其他类的流形重叠。
有效信息增益 EIG

我们使用流形的体积来度量丰富度。类的特征矩阵 d（维度）x N（样本数)、使用特征矩阵得到协方差矩阵、使用协方差矩阵进而表示出流形体积，对于可能出现非全秩（导致体积解为 0）的情况，我们添加一个单位矩阵，为了计算的稳定性，我们转换成对数的形式进行计算。

图像空间中丰富度是固定的的（数据集固定，丰富度固定)，但在嵌入空间中，不同模型提取的丰富度不同（可能由模型参数初始化的随机导致)，实际上我们并不关心丰富度的绝对大小，而是侧重于随着分布偏移，丰富度的变化（分布变化-丰富度变化)，所以我们将 EIG 定义为比值

这里有两个对 EIG 的约束，首先，当信息增强样本为 0 时，EIG 为 0，档信息增强样本全部聚焦于一个点时，实际上并没有提供有效信息给模型，反而加剧了运算成本，此时的 EIG 应该为负数，我们定义为 EIG 的下限值

对于增益后的总丰富度，使用原始样本和增强样本的体积和是错误的，应该使用两者的并集，原始样本和增强样本可能重叠一部分，导致体积计算不准确，此外使用并集的策略还能表征原始样本和增强样本的分布偏移程度

实验研究与分析：

四种数据增强方法的核心操作
1：图像空间的数据增强：
Remix：通过修改合成标签来改进 Mixup，以便当头部类和尾部类的样本混合时，最终标签 100%由尾部类贡献。
CMO：使用 CutMix 将尾类样本中的补丁粘贴到头类样本上，从而生成增强样本。
2：嵌入（特征）空间的数据增强：
OFA：将样本特征分解为类特定特征和类通用特征，并将尾类的特定特征与头类的通用特征相结合，以生成尾类的增强样本。
FDC：观察到相似的类具有相似的分布统计（方差)，因此将最相似的头类的方差转移到尾类以重新估计分布，从新分布中采样增强样本。
Summary and quiz：
实际上 EIG 只是定义了增强部分的体积和原始体积的比例，对于增强部分是否超出真实分布并没有做出约束，所以我们说 EIG 被确定在一定范围（不过低也不过高）
当 EIG 太低时，表明数据增强方法所带来的附加信息是很小，这样的数据增强意义不大。相反，当 EIG 太高时，它可能导致缺乏实际意义的增强样本（增强样本不在真实分布内）

为了验证这一假设，我们在常用的长尾数据集 CIFAR-10-LT、CIFAR-100-LT 和 ImageNet-LT 上进行了不同不平衡因子的实验。对于 CIFAR-10-LT 和 CIFAR-100-LT，我们使用 SGD 优化器训练 ResNet 32，动量为 0.9，权重衰减为 0.0002。在 ImageNet-LT 上，我们使用 ResNeXt-50 作为所有方法的主干。模型使用 SGD 优化器进行训练，批量大小为 256，动量为 0.9，权重衰减因子为 0.0005，学习率为 0.1（线性 LR 衰减）。

生成包含不同 EIG 值的增强数据

EIGtail：多个尾类的 EIG 相加/尾类数量

重复 100 组 Remix，得到 100 组增强数据，然后进行 100 组 EIGtial 的计算，从高到低随机抽取 50 个，发现这个 50 组的 EIGtail 差距不大，为了获得具有广泛有效信息增益的增强数据集，我们提出了算法 1：对于一组 Remix，先直接生成大量增广样本（形成多个所谓的子集，部分这些子集后面可以构成目标增强数据集），使用 k-means 聚类 500 个子集，假设一个子集 10 个样本，为了补全尾类的 450 个样本，我们需要 45 个子集，然后选择最大化 EIGtail 的 45 个子集，然后选择最小化 EIGtail 的 45 个子集，通过动态选择不同的子集，我们可以创建具有不同 EIG 值的增强数据集，然后通过这种方法生成了 50 组高 EIGtail 的增强数据和 50 组低 EIGtail 的增强数据，并从高到低随机抽取 50 个。

增强数据需要适当的 EIG

在六个数据集：CIFAR-10-LT（IF=10，50，100）和 CIFAR-100 LT（IF=10，50，100）

应用四种信息增强策略：Remix、CMO、OFA、FDC

一种信息增强策略有从高到低 EIGtail 值的 50 组增强数据，一共 50x6x4=1200 个实验

当增强数据的 EIGtail 落在适当的范围内时，信息增强方法的性能达到其最佳值，超过了原始方法报告的结果。这一结果证实了我们的假设，即有效的增强数据需要在丰富度增益和它引入的分布偏移之间取得平衡。此外，随着数据集的不平衡水平增加（观察分布更放缩于真实分布），数据增强方法需要生成提供更大 EIG 的增强数据（更大的增强体积）。我们的研究结果表明，引导数据增强方法生成更合适的样本可以进一步提高长尾识别模型的性能。

图 4 和图 5 表示对于三个 IF（3）下的 CIFAR-10-LT 和 CIFAR-100-LT（2），四种信息增强方法（4），一共 3X2X4=24 张图，用算法 1 得到的差异递减 EIGtail 的 50 个增强数据，分别合并原始数据训练模型的 Top1 精度（离散点），有些之前的文章做过所以有基线，曲线为离散点的拟合曲线，凸函数意味着有极值。

数据不平衡与最优 EIG 的关系

CIFAR10 构建了 10 个长尾数据集（IF=10、20、…、100），使用 CMO 在每个长尾数据集上生成 50 组具有不同 EIGTail 水平的增强样本，对离散点进行拟合，取凸函数的最高点对应的 EIG，然后得到 10 个长尾数据集的最高精度对应的 EIG（最优 EIG），得到规律，随着 IF 的增大，最优 EIG 增大，最后趋于稳定，在流形空间的表征，就是随着长尾现象的加剧，观察分布的体积会变小，我们增强数据的体积在真实分布中的机会变多（无脑体积扩大的情况下，有效 EIG 的机会变大），但是会随着长尾有一个极端体积，之后长尾加剧就几乎不再缩小体积

选择具有指定 EIG 的增强数据

已知最优 EIG（一般 EIG 是 EIGtail，因为只有尾类要进行增强，才会有 EIG 的讨论），或者最优 EIG 区间，我们通过动态选择增强子集来实现特定 EIG 的增强数据选择，本质是动态调整当前增强数据的子集，实现当前增强数据 EIG 与目标 EIG 的差值最小化

增强对 ImageNet-LT 的影响

与之前的类似，结果：总体提升、尾类大幅提升

总结：

大多数现实世界的应用涉及长尾分布的数据，导致有偏见的模型。为了确保在各种场景下的鲁棒模型性能，以模型为中心的方法，如重新加权损失函数，迁移学习和集成学习已经被广泛提出。然而，随着基础模型（大模型）开发的加速，以模型为中心的方法的有效性逐渐减弱。回到问题的本质，直接改进数据是解决长尾学习挑战的根本方法。这项工作的核心贡献是引入了有效信息增益（EIG），这是一个可以指导数据增强方法设计和应用的指标。虽然信息增益和分布偏移对数据增强性能的影响被广泛认可，但本研究首次量化了这两个因素，并探讨了它们之间的适当平衡如何影响模型性能。在未来，我们的目标是促进以数据为中心和以模型为中心的方法，以推进长尾学习领域。

缺陷和可能方法：

1：实际上无法直接得到最优 EIG 或最优 EIG 区间，而是通过先验的实验结果，对比出最优 EIG 或最优 EIG 区间，再利用这些先验结果，对重复类似试验进行指导，得到最优 EIG 对应的增强数据

未来的工作：从流形的角度全面的分析观察分布、真实分布、补全分布（增强后总数据集的观察分布）、丰富度度量（体积、面积、曲率、本征维度）、分布偏移，如果得到一个真实分布（或模拟真实分布）的边界（是否可以通过近似类的高斯分布模拟，近似类可以在头类近似尾类、外部数据集先验（同类或近似类）、联邦共享同类实现），同时得到当前体积扩充方向（知道边界后，只有知道体积漫延方向才能规划，几何方向扩充-师兄的 IJCV），那么将可以定量 EIG，从而更有效的对信息增强策略进行指导!!!

2：实际上本文对分布偏移的解释比较模糊，不同信息增强策略，有不同的分布扩充策略，如果定性分析对不同的信息增强策略的分布偏移方向对可解释机器学习至关重要(不一定像师兄的 IJCV 那样，沿着先验的几何方向扩充，可能是非线性的，如果解释背后的机制和机理？)

未来的工作：延续师兄的系统性工作，我希望从人类本身的生理结构中得到启发，将机器学习背后的机理和机制展开

AI for Science：

DNN 深度神经网络中的偏见性问题，实际上能够利用人脑视觉作为启发的线索，人类视觉系统中，神经元收到同一类别不同物理特征的视觉刺激时，形成物体神经流形（Science），而视觉皮层通过逐层加工，使得复杂的物体神经流形解开缠绕，使不同的物体神经流形在网络深层实现清晰的分类，实现物体识别，而神经流形的几何复杂度会影响解缠绕的难度和识别的效果，DNN 对图像的响应与人类视觉相似，服从流形分布，此外在结构上模仿了这种多层的解耦机制，通常由表征网络和分类器组成。为了和人类视觉系统进行类比，我们将数据流形经过表征网络映射后形成的点云流形称为感知流形。未来的研究尝试通过分析类别感知流形的几何复杂性，探索 DNN 偏见的来源。