超越 Logits 的探索：基于嵌入的半监督分类层次动态标记（HDL）

摘要：

半监督学习任务生成伪标签场景下，基于模型预测生成伪标签在不同数据集间的适应性差，需要定制分类器（分类数量），每种分类器都要重新训练，但是特征提取部分泛化能力较强，同时研究表明模型的表示网络（特征提取器）可靠于分类网络（分类器），我们提出分层动态标记（HDL），依赖图像嵌入（特征）生成伪标签，同时还提出自适应算法选择 HDL 参数，此外，还将 HDL 与图像编码器（CLIP）（特征提取部分泛化能力更强）组合，我们的工作提高了模型的性能。

引言：

置信度学习

模型预测生成标签，分类器经过 softmax 函数转化成每个类的预测概率（置信度），每个预测选择置信度最高的作为当前预测的伪标签、选择置信度高于阈值的作为当前预测的伪标签（当前预测的各个类置信度都不高于阈值的预测直接不要）、每个预测根据置信度进行加权伪标签生成、等等策略，通过上面的方法，伪标签质量得到提高

预测网络生成伪标签

自动标注方法中的置信度学习通过生成高质量的伪标签，对模型进行训练，再生成高质量的伪标签，通过迭代：生成伪标签-训练模型 实现对模型的优化。但是这种通过置信度进行伪标签生成的方法依赖于分类器，而分类器存在两大缺陷：

1：分类器在不同数据集间的泛化能力差，定制分类器成本高

2：训练数据有偏或有噪声（有错误标签）时，分类器会表现偏差或对噪声的过度记忆，导致预测效果差，伪标签生成效果差

表示网络生成伪标签

研究表明：表示网络比预测网络更可靠，有偏发生在预测网络，无偏发生在表示网络，解耦训练策略的动机正是源于此，同时噪声检测领域研究表明，模型会过度记忆噪声样本

因此我们聚焦于基于嵌入的（特征）、以数据为中心的方法，具体的，在无监督学习中，往往使用聚类的方式区分类，如果嵌入 x 与 k 个近邻的嵌入属于同一类，则这种数据集是满足 K-NN 标签聚类性的，标签可聚类性意味着可以通过已知标签（已知标签会进行同类聚集），推断未标注标签的所属，要注意的是：满足 K-NN 标签聚类性的概率随着 k 的增加而降低。我们在类平衡和长尾场景下分别进行实验，利用表示网络提取嵌入后，执行 HDL 算法，将真实标签和 HDL 生成的伪标签一起训练模型，此外受益于 CLIP 强大的视觉表征（使用 clip 作为表示网络），使得设置极少量真实样本（较小的 k 值，意味着标签可聚类性的概率更高）也能实现未标注样本的聚类（这是因为表征能力很强，纵使少量样本也能表征聚类空间），判断未标注标签的所属：我们首先考虑一种树结构的搜索算法，然后在此基础上提出了 HDL（分层动态标记）算法。

工作主要贡献

1：我们提出了一个分层动态标记算法（HDL），它不依赖于模型预测，这种方法非常灵活，也可以与 CLIP 结合，作为通用的数据预处理模块。

2：我们提出了一种自适应选择 HDL 超参数 k 的方法，使 HDL 对经验选择的依赖性降低，从而增强了它的普适性。

3: 在类平衡和类不平衡的情况下, 我们的方法显著提高了半监督模型的性能（表 2 和表 4），验证了表示网络比分类器或预测器更可靠的概念。此外，我们的方法有可能改变半监督学习

中生成伪标签的范式。

相关工作

基于置信度学习的图像伪标签生成。（前面有详细介绍，这里不赘述，此块的工作主要依托预测网络，优化的方向诸如：对样本多次预测取平均值作为伪标签、预测标签分布与现有标签分布对齐、更优的置信度阈值选择算法等等）但是，我们的工作是基于特征，而不是预测

准备工作

任务和符号定义

见手推笔记

标签集群性

我们的工作旨在实现无需训练的自动图像注释，工作的基础是标签聚类性（假设满足标签聚类性），聚类性意味着两个接近的嵌入属于同一个真类的概率很高，对标签集群性的分析见手推笔记。

使用嵌入的自动分类

我们首先介绍利用标签聚类性进行注释的最简单的想法，然后逐步完善它，提出一个更通用和上级的算法。为了演示的清晰性，本节中使用 CLIP 从数据集中提取嵌入;然而，实验部分将包含各种图像编码器以进行更广泛的评估。

基于（k，δk）标签聚类性的图像标记直觉（KNN-DV）

见手推笔记

未标记的嵌入可以促进标记（KNN）

见手推笔记

分层动态标记（HDL）

见手推笔记

k 的自适应选择和分析

见手推笔记

实验

实验目标和方案设计

工作的动机：图像嵌入比模型预测更可靠。因此，我们提出了基于图像嵌入的 HDL 算法。本文的目标是证明 HDL 优于基于置信学习（预测网络）的标记方法。

半监督训练模型，同时预测未知样本的伪标签，以实现交替优化。这就引出了一个问题：如果基于嵌入的 HDL 比模型预测更可靠，那么在半监督学习收敛之后，使用 HDL 重新标记未知样本并继续训练半监督模型将进一步提高性能。如果性能无法提高，那就证明我们的方法并没有带来任何额外的好处。

基于这种思想，我们从收敛的半监督模型中提取图像嵌入，并在具有类平衡（见 5.4 节）和长尾分布（见 5.5 节）的场景中进行了上述实验。

数据集和详细设置

（1）数据集：

平衡数据集包括 CIFAR-10、CIFAR-100 和 STL-10。类别不平衡数据集是具有各种不平衡因素的 CIFAR-10-LT。

CIFAR-10：遵循半监督学习中的常见做法，将数据划分如下：对于每个类，我们随机选择 4 个，25 个和 400 个样本作为标记数据，并将训练集中的剩余样本作为未标记数据。

CIFAR-100：我们随机选择 4、25 和 100 个样本作为每个类别的标记数据，并将剩余样本视为未标记数据。
STL-10：包括 10 个类，500 个训练样本，800 个测试样本，以及另外 10，000 个未标记样本。对于每个类，我们从训练集中随机选择 1，2 和 100 个样本作为标记数据。
CIFAR-10-LT（IF = 50，100，200）：标记率始终设置为 10%，最频繁类（完整 5000 个样本）包含 500 个标记样本和 4500 个未标记样本。
（2）替换细节：
我们分别在 CIFAR-10、CIFAR-10-LT、CIFAR-100 和 STL 10 上使用了 Wide ResNet-28-2、Wide ResNet-28-2、Wide ResNet-28-8 和 Wide ResNet-37-2 作为骨干网络，MixMatch、ReMixMatch、FreeMatch 和 DualMatch 作为我们的基准，具体来说，在四个半监督模型完成训练后，我们利用 HDL 重新标记未标记的数据，并继续训练额外的数据。
（3）比较方法：
与 HDL 类似，我们使用 kNN-DV（第 4.2 节）增强了 MixMatch [2]，ReMixMatch [1]，FreeMatch 251 [33]和 DualMatch [32]，并将其与 HDL 进行了比较。此外，我们将我们的方法与常见的半监督模型进行了对比，包括 E-Model [15]，Pseudo-Labeling [16]，Mean Teacher [31]，UDA [35]，FixMatch [30]，Dash [37]，CCL [6]和 RA [7]。

超参数 k 的设置

使用待改进的训练好的半监督模型（即基于置信度学习的预测网络半监督模型训练）提取图像嵌入，然后基于第 4.4 节中提出的方法选择 k。

类平衡数据集的结果

1：半监督学习的两种方法
一致性正则化模型：
模型对同一数据点（未标注样本）在不同扰动（数据增强、高斯噪声）下的预测结果应该是一致的（语义不变）。通过这种一致性约束（最小化差异损失），训练模型
伪标记模型（置信度学习）：
标注样本训练模型，模型推理未标注样本形成伪标签，伪标签参与模型训练
2：实验结果

无论是一致性正则化模型还是伪标记模型，它们的共同目标都是优化模型，使其对相似样本的表示更加一致和相似，从而提高性能。观察表 2 和表 3 中的 kNN-DV 和 HDL 的结果，它们显著提高了现有半监督学习方法在不同数据集上的性能。实验结果表明（在使用不同数量的标注样本进行训练情况下，计算不同方法训练的半监督模型对未知标签的预测错误率），充分训练的半监督模型（先是置信度学习训练半监督模型，然后提取未标注样本的特征，从特征角度再预测一次未标注样本，然后训练半监督模型形成充分训练的半监督模型）生成的图像嵌入比单纯的置信度预测更可靠（在置信度学习后的半监督模型，使用 HDL 仍有提升，所以 HDL 是有效的）。因此，我们强调了关注基于图像嵌入的伪标签半监督模型的重要性。

具体来说，在 CIFAR-10 上有 40 个标签，HDL 将 MixMatch 和 ReMix Match 的性能分别提高了 1.52%和 1.18%。在具有 250 个标签的 CIFAR-10 上，HDL 在 MixMatch、ReMixMatch、FreeMatch 和 DualMatch 上的性能分别提高了 1.35%、1.12%、0.97%和 0.93%，增强的 FreeMatch 实现了最先进的性能。此外，在具有 4000 个标签的 CIFAR-10 数据集上，HDL 将 MixMatch 的性能提高了 1.15%， DualMatch + HDL 实现了最佳性能。对于 CIFAR-100 和 STL-10，HDL 一致地全面增强了现有方法的性能。

重要的是，我们的方法不需要对现有方法进行任何更改，只需要使用训练好的半监督模型进行特征提取，重新标记和继续训练原始网络。

类不平衡数据集的结果

与类平衡的场景相比，HDL 在长尾数据集上表现出更明显的性能。因为长尾场景中基于图像嵌入的方法可以减少模型偏差（研究表明，有偏数据集下，特征提取网络几乎无偏，偏差主要发生在分类器，传统半监督模型有偏，伪标签自然也是有偏的，基于无偏的特征进行重伪标签生成，再次训练分类器，纠正分类器的偏差）。

kNN-DV 和 HDL 的附加分析

kNN-DV 不考虑未标记样本，潜在地提高了对图像嵌入集的可聚类性的要求，以使其有效。我们将 kNN-DV 和 HDL 的超参数 k 设置为 3，以量化和分析 kNN-DV 的上述缺点。在我们对 CIFAR-10、CIFAR-100 和 STL-10 数据集的研究中，我们发现，在所有三个数据集上，kNN-DV 的聚类概率始终低于 HDL（如图 2 左侧所示）。为了增强聚类能力，我们不得不减少最近邻居的数量（减小 k）。然而，减少最近邻居的数量（减小 k）会损害投票算法的可靠性。因此， HDL 理论上在性能上超过了 kNN-DV，实验结果也同样证明。

用于数据处理模块的 HDL（模块、clip+模块）

使用 HDL 作为通用的数据处理模块时（提取未标注标签的特征，在特征空间使用 HDL 生成伪标签），可以使用 CLIP 来提取图像嵌入。因为 CLIP 是一种通用的图像编码器，我们可以假设所提取的嵌入集满足聚类性的概率在不同的图像数据集中大致相同，下面进行实验证明。

实验证明：我们在 mini-ImageNet、Clothing-1 M、CIFAR-100 和 Caltech 101 数据集上评估了不同 k 值下标签聚类性的概率 μk。随后，我们计算并绘制了 μk 的均值和方差，如图 2（右图）所示。实验结果表明，在所有四个数据集上，μk 的方差非常小，这验证了我们的假设。

同时意味着在实践中，使用 clip 提取不同数据集的特征执行 HDL 算法时，没有必要每次都对 μk 进行统计评估（不需要进行不同数据集和不同 k 值的 μk 评估），相反，可以使用可靠的经验值作为替代（clip 作为 backbond 时，任何数据集，直接固定不同 k 对应的 μk）。

总结

这项工作源于嵌入比置信度预测更可靠的观点。它引入了基于嵌入的分层动态标记算法，并显著增强了现有的半监督模型。这种进步具有促进伪标签生成范式改进的潜力。

缺点和可能的方法

1：关于标签集群性背后的机制仍有探索的空间

实际上是不同 backbond 表征能力的量化，能否在特征空间或者其他空间进行度量，尤其是使用人脑视觉系统类比分析 backbond 结构，进行表征能力的量化

2：迭代到最终的 HDL 算法，仍然有可能陷入循环搜索

KNN 到 HDL 的过程，HDL 实际上最大可能的避免了发生循环搜索（自我最优、整体最优），但是应对循环搜索，是否可以有一个标注，当发生循环时（跳回走过的路时），马上更换中心实例，再尽可能避免发生循环