告别数据集资源匮乏，谷歌与斯坦福大学用弱监督学习给训练集打标签

人工智能3 月 24 日

数据集就是机器学习行业的石油，强大的模型需要含有大量样本的数据集作为基础。而标记训练集中的数据样本是开发机器学习应用的最大瓶颈之一。

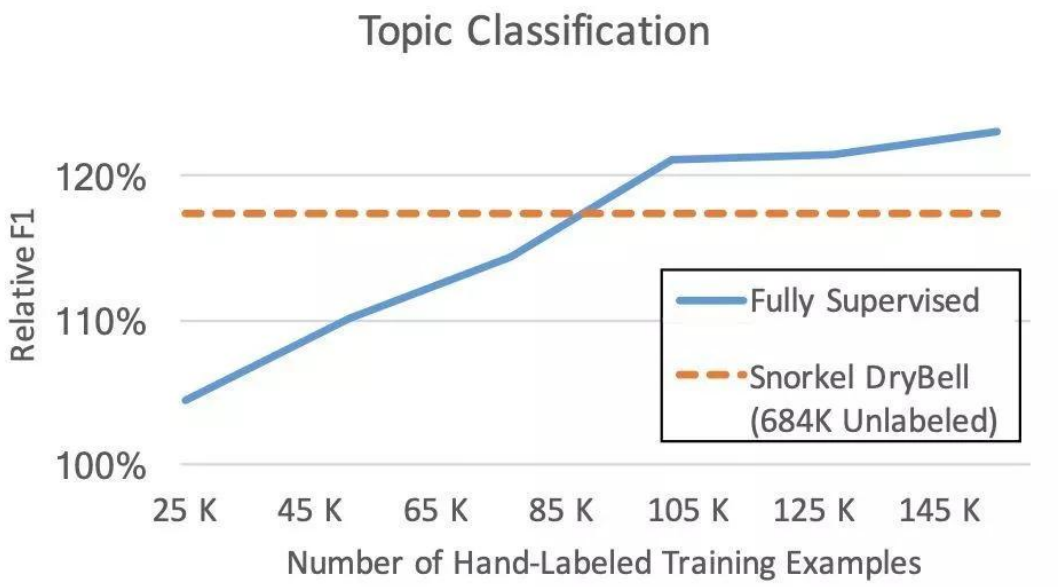
最近，谷歌与斯坦福大学、布朗大学一起，研究如何快速标记大型数据集，将整个组织的资源用作分类任务的弱监督资源，使机器学习的开发时间和成本降低一个数量级。

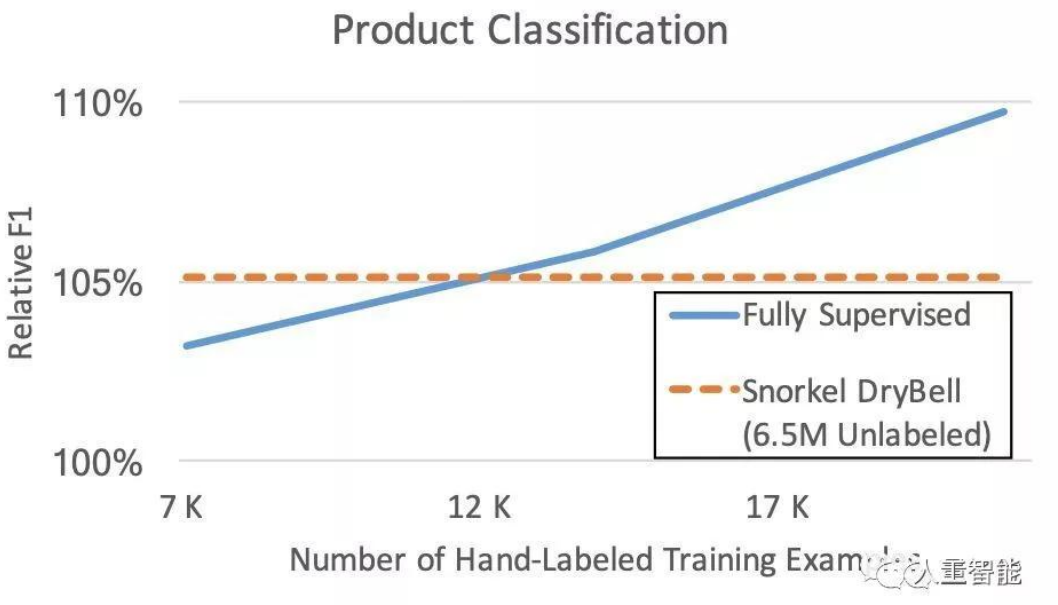
谷歌在论文中表示，这种方法能让工程师能够在不到 30 分钟的时间内对数百万个样本执行弱监督策略。

他们使用一种 Snorkel Drybell 系统，让开源 Snorkel 框架适应各种组织知识资源，生成 Web 规模机器学习模型的训练数据。

Snorkel 是由斯坦福大学在 2017 年开发的系统，它可以在弱监督条件下快速创建训练数据集，该项目已经在 GitHub 上开源。而 Snorkel Drybell 的目标是在工业规模上部署弱监督学习。

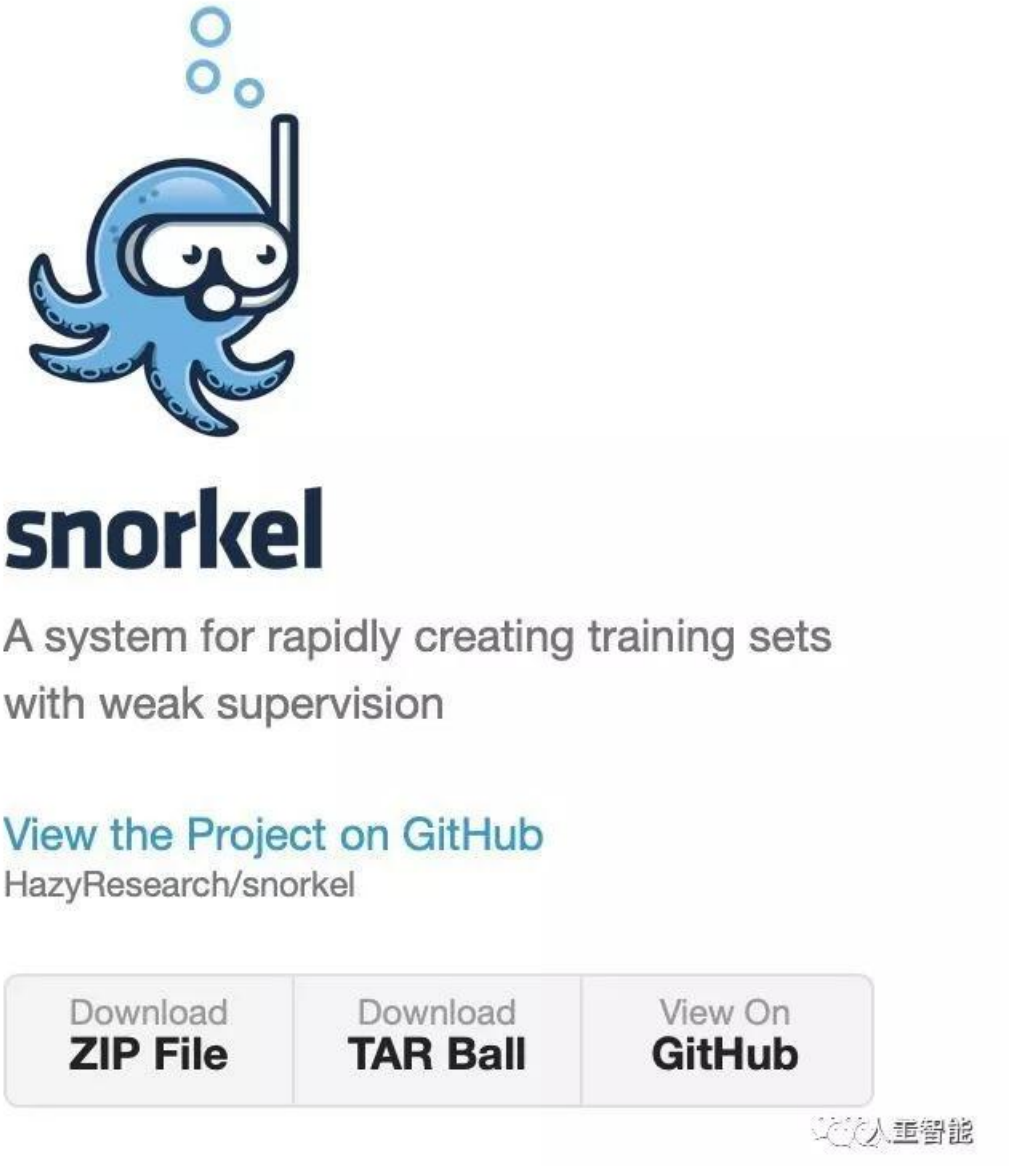
而且用这种方法开发的分类器质量与手工标记样本进行训练的分类器效果相当，把弱监督分类器的平均性能提高了 52%。





什么是 Snorkel

Snorkel 是斯坦福大学在 2016 年为许多弱监督学习开发的一个通用框架，由这种方法生成的标签可用于训练任意模型。

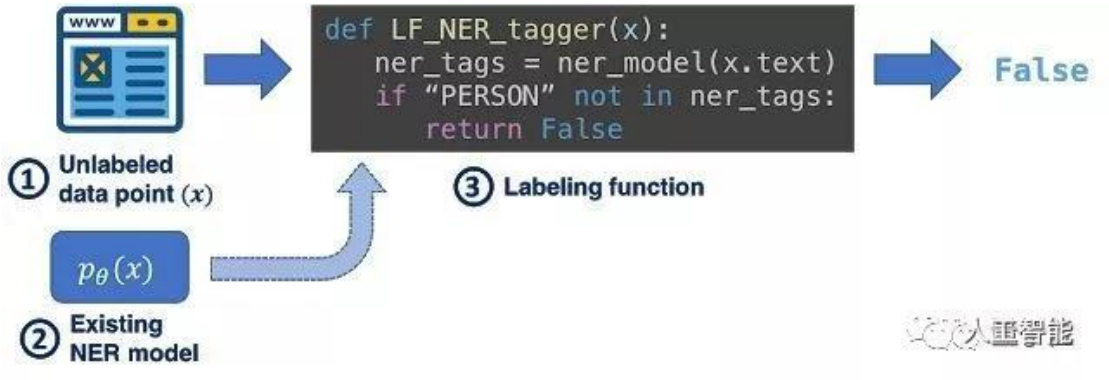


已经有人将 Snorkel 用于处理图像数据、自然语言监督、处理半结构化数据、自动生成训练集等具体用途。

原理

与手工标注训练数据不同，Snorkel DryBell 支持编写标记函数，以编程方式标记训练数据。

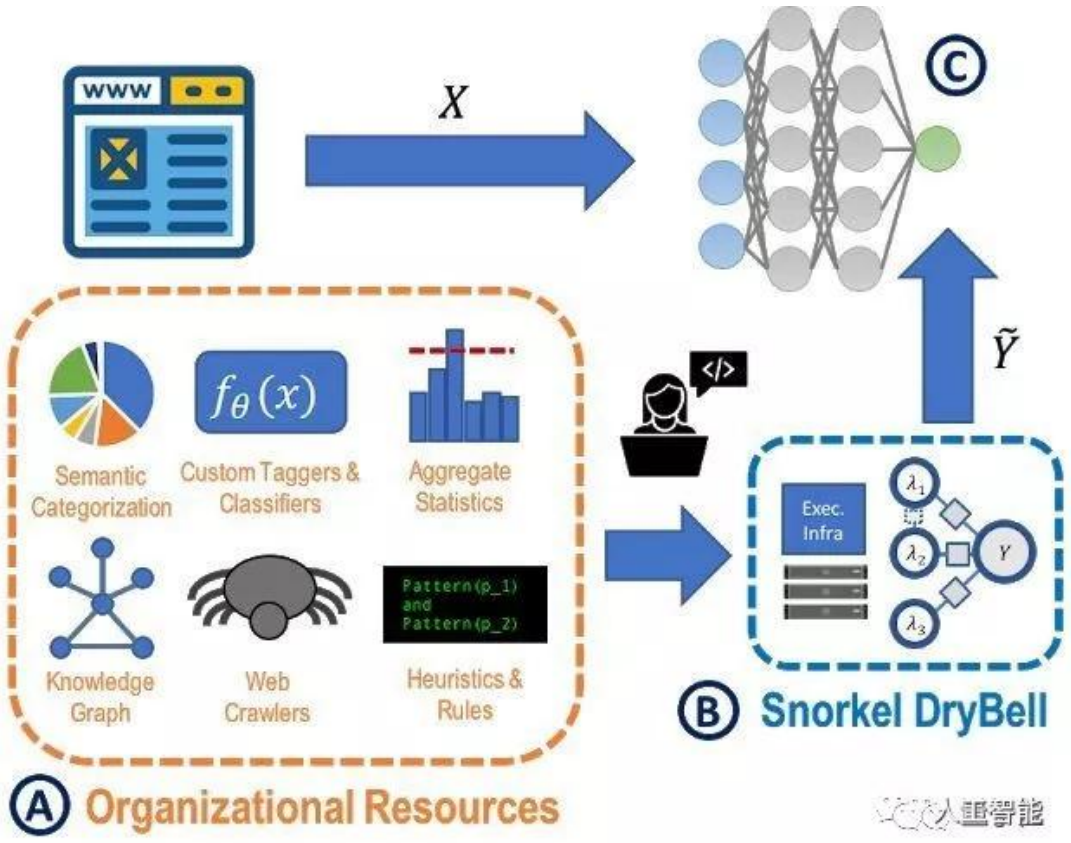
过去的方法中，标记函数只是以编程方式标记数据的脚本，它产生的标签是带有噪声的。



为了解决噪声等问题，Supert Drybell 使用生成建模技术，以一种可证明一致的方式自动估计标记函数的准确性和相关性，而无需任何基本事实作为训练标签。然后用这种方法对每个数据点的输出进行重新加权，并组合成一个概率标签。

使用多种知识来源作为弱监督

Snorkel Drybell先用多种知识来源作为弱监督，在基于MapReduce模板的pipeline中编写标记函数，每个标记函数都接受一个数据点生成的概率标签，并选择返回None（无标签）或输出标签。



这一步生成的标签带有大量噪声，甚至相互冲突，还行需要进一步的清洗才能用到最终的训练集中。

结合和重新利用现有资源对准确度建模

为了处理这些噪声标签，Snorkel DryBell 将标记函数的输出组合成对每个数据点的训练标签置信度加权。这一步的难点在于，必须在没有任何真实标签的情况下完成。

研究人员使用生成建模技术，仅使用未标记的数据来学习每个标记函数的准确性。通过标签函数输出之间的一致性矩阵来学习打标签是否准确。

在 Snorkel DryBell 中，研究人员还实现了建模方法一种更快、无采样的版本，并在 TensorFlow 中实现，以处理 Web 规模的数据。

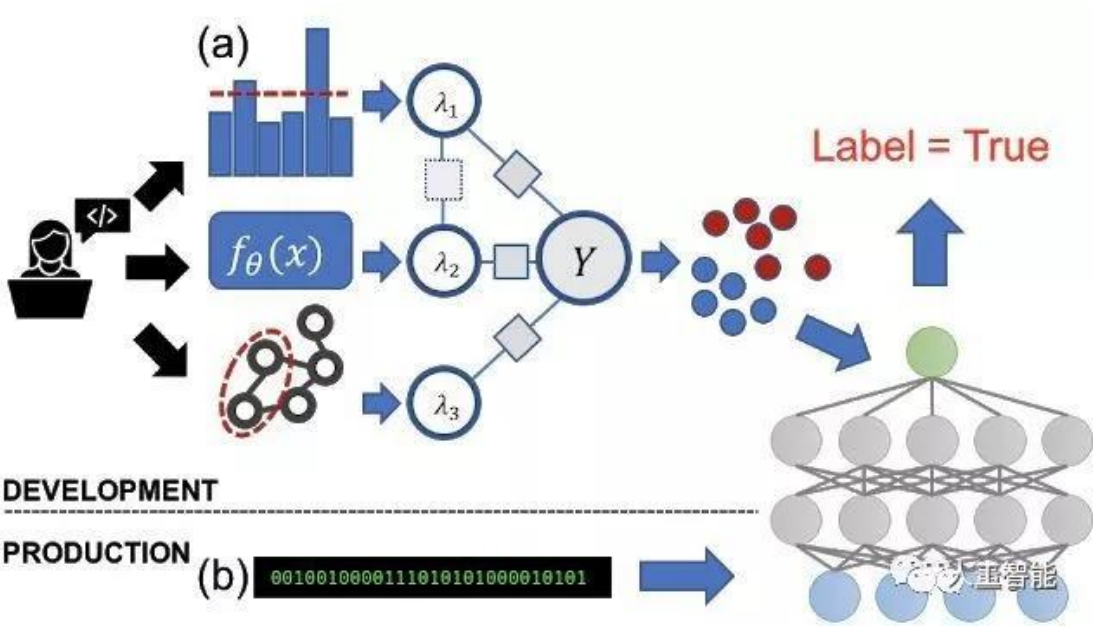
通过在 Snorkel DryBell 中使用此程序组合和建模标签函数的输出，能够生成高质量的训练标签。与两个分别有 1.2 万和 8 万个手工标记训练数据集比较，由 Snorkel DryBell 标记的数据集训练出的模型实现了一样的预测准确度。

将不可服务的知识迁移到可服务的模型

在许多情况下，可服务特征(可用于生产)和不可服务特征(太慢或太贵而无法用于生产)之间也有重要区别。这些不可服务的特征可能具有非常丰富的信号，但是有个问题是如何使用它们来训练，或者是帮助能在生产中部署的可服务模型呢？

在 Snorkel DryBell 中，用户发现可以在一个不可服务的特征集上编写标签函数，然后使用 Snorkel DryBell 输出的训练标签来训练在不同的、可服务的特征集上定义的模型。

这种跨特征转移将基准数据集的性能平均提高了 52%。



这种方法可以被看作是一种新型的迁移学习，但不是在不同的数据集之间转移模型，而是在不同的特征集之间转移领域知识。它可以使用速度太慢、私有或其他不适合部署的资源，在廉价、实时特征上训练可服务的模型。

资源地址

论文地址：

https://arxiv.org/abs/1812.00417

Snorkel 项目地址：

https://hazyresearch.github.io/snorkel/