标注平台与训练平台

田兴 2023/4/7

目前我们做了一个数据标注平台, 起初的目的是加快 `语音信箱` 音频的分类标注速度.

我们没有直接用开源的代码, 主要考虑到自主开发标注平台可以具更多更丰富的功能.

标注平台的完善是一个长期的过程. 以下将会介绍一些标注平台, 训练平台可具备的特性.

标注平台

标注平台可具备的特性:

1. 在操作上, 提升标注速度.
2. 减少在任务分配时的数据传输, 数据管理更统一, 不会出现重复标注的情况. 减少沟通.
3. 通过外部API, 或已标注数据训练的模型, 为标注提供提示信息, 以帮助标注员更快地找到合适的标签.
4. 通过已标注数据训练的模型来判断哪些样本更难被区分, 将不容易区分的样本推荐出来标注. 提高数据标注的质量.
5. 将模型训练后, 会将一部分样本分类错误, 这有可能是标注时标错了 (或: 不同标注员的标准不同), 可将其推荐出来重新标注.
6. 当新增标注数据达到一定数量后, 自动触发模型训练任务, 并布署模型. 反复循环, 并在台平上展示目前模型所能达到的分数. 提升全流程的自动化.
7. 算法工程师可以在标注平台展示 `标注的标准`, 用算法预判等标注样本的信息后, 在标注时更精准地提示, 以提高不同标注员的标注的一致性.

训练平台

训练平台可具备的特性:

1. 规划常用算法库 (即: 预先做好配置). 可实现一键训练布署模型.
2. 提供对外API和测试界面, 训练师可在页面上测试模型, 从数据上查看模型的优缺点. 进而在场景配置时, 避免模型的缺点.
3. 为训练师提供有用的工具, 如ChatGPT接入, 或其它自研模型的体验界面.
4. AI能力展示. 对于做过的尝试,研究, 它可能无法在生产中使用, 但使用了多少数据, 用的什么模型, 方法, 达到的效果, 等, 可以在界面展示, 提供测试. 更能直观地理解各项工作的进展, 以及进一步深化的难度.
5. 也许, 训练师应该自行给不同的场景定义不同的标签体系, 训练平台可以描述各种模型的优缺点, 训练师可以自己标注数据并上传训练模型, 分析哪些标签的定义应该被优化.