第一次大作业

EasyNlp:

主要功能和流程：

Easynlp 主要为深度学习提供了算法框架，它具有以下特性：

易用且兼容开源：EasyNLP 支持常用的中文 NLP 数据和模型，方便用户评测中文 NLP 技术。除了提供易用简洁的 PAI 命令形式对前沿 NLP 算法进行调用以外，EasyNLP 还抽象了一定的自定义模块如 AppZoo 和 ModelZoo，降低 NLP 应用的门槛，同时 ModelZoo 里面常见的预训练模型和 PAI 自研的模型，包括知识预训练模型等。EasyNLP 可以无缝接入 huggingface/transformers 的模型，也兼容 EasyTransfer 模型，并且可以借助框架自带的分布式训练框架（基于 Torch-Accelerator）提升训练效率。

大模型小样本落地技术：EasyNLP 框架集成了多种经典的小样本学习算法，例如 PET、P-Tuning 等，实现基于大模型的小样本数据调优，从而解决大模型与小训练集不相匹配的问题。此外，PAI 团队结合经典小样本学习算法和对比学习的思路，提出了一种不增添任何新的参数与任何人工设置模版与标签词的方案 Contrastive Prompt Tuning，在 FewCLUE 小样本学习榜单取得第一名，相比 Finetune 有超过 10% 的提升。

大模型知识蒸馏技术：鉴于大模型参数大难以落地的问题，EasyNLP 提供知识蒸馏功能帮助蒸馏大模型从而得到高效的小模型来满足线上部署服务的需求。同时 EasyNLP 提供 MetaKD 算法，支持元知识蒸馏，提升学生模型的效果，在很多领域上甚至可以跟教师模型的效果持平。同时，EasyNLP 支持数据增强，通过预训练模型来增强目标领域的数据，可以有效的提升知识蒸馏的效果。

以使用该工具进行小样本学习实践为例。首先进行环境准备，设置环境变量并下载需要的数据集。然后需要根据不同的算法设置相应的参数。设置完成后开始训练即可，训练完成之后即可在设置好的路径得到相应的权重文件。

包含的主要功能模块：

CLUE Benchmark 提供 CLUE 评测代码，方便用户快速评测 CLUE 数据上的模型效果。

DataHub 为用户提供一个加载和处理各种数据的接口。

ModelZoo 为用户提供通用的预训练模型。

APPZoo 为用户提供不同的模块以便完成各种任务。

Pipelines 测试和部署 Easynlp 模型

Diffusion 提供 diffusion 模型以及训练接口

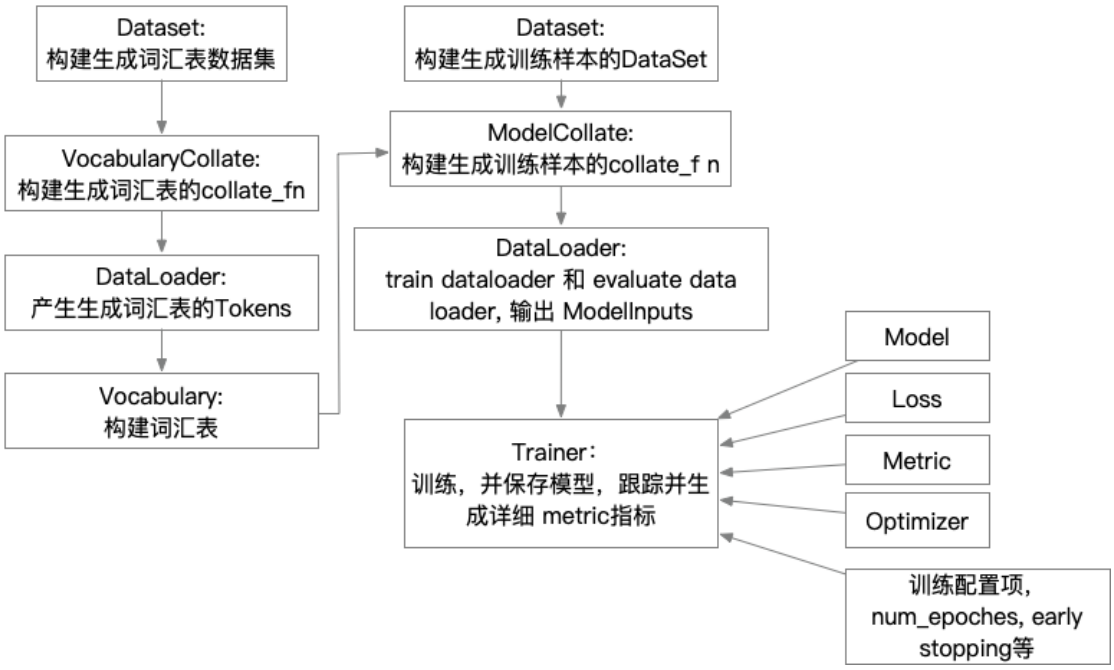
Fewshot_learning 提供小样本学习工具

选择分析的功能模块：

将对 Easynlp 以及 Easytext 这两个 NLP 开发工具进行对比分析。

Easytext:

主要功能和流程：



包含的主要功能模块:

Dataset 构建生成词汇表数据集以及构建生成训练样本的 Dataset

VocabularyCollate 构建生成词汇表的 collate_fn

DataLoader 产生词汇表的 Tokens

Vocabulary 构建词汇表

ModelCollate 构建生成训练样本的 collate_fn

Trainer 训练并保存模型，跟踪并生成详细 metric 指标。

选择分析的功能模块:

将对 Easynlp 以及 Easytext 这两个 NLP 开发工具进行对比分析。