图像分类任务的多指标样本质量评估框架

项目方案汇报

李想

1 研究背景与设计目标

在图像分类任务中，训练数据的质量对模型泛化性能具有决定性影响。然而现有的数据选择方法往往仅关注已有训练集，不具备对未见样本的评分能力，或者需要依赖复杂的训练动态，难以高效应用。本项目旨在设计一个具备以下能力的统一框架：

- **可解释性强的样本评分机制**：对每个样本生成明确的质量得分，得分反映其语义一致性、多样性贡献、边界信息价值等。
- **支持未见样本评分**：不依赖训练动态，可在训练完成后对新样本快速评分。
- **具备数据选择能力**：基于评分进行可控比例的数据选择，并支持进一步利用 Selection Optimization 优化选择结果。
- **从代理模型训练动态中学习评分指标权重**：仅需一次代理模型训练动态，即可获得长期可复用的评分权重。

本框架以静态多指标评分为基础，以训练动态信息学习权重为增强机制，以 Selection Optimization 为最终优化模块，形成一个完整而统一的样本评估体系。

2 方法总体架构

本项目的方法整体由三大模块构成：

1. **静态多指标样本评分（Static Scoring）**
2. **从训练动态学习指标权重（Weight Learning via Dynamics）**

3. **基于权重评分的选择优化（Selection Optimization）**

给定样本 x_i，其最终评分定义为

$$\text{Score}(x_i) = \tilde{w}_1 \cdot \text{SA}_i + \tilde{w}_2 \cdot \text{Div}_i + \tilde{w}_3 \cdot \text{DDS}_i,$$

其中 $(\tilde{w}_1, \tilde{w}_2, \tilde{w}_3)$ 为从训练动态学习得到的权重。

该评分同时用于：

- **（1）训练后对未见样本的质量评估**
- **（2）在 Selection Optimization 中作为优化目标的一部分**

3 静态多维度样本评分体系

静态评分体系完全不依赖代理模型训练动态，因此天然支持对未见样本的高效评分。为每个样本 x_i 定义如下三类静态评分特征：

3.1 语义对齐度（Semantic Alignment, SA）

使用 CLIP 计算图像与类别文本的语义相似度：

$$\text{SA}_i = \cos(\text{CLIP}_{img}(x_i),\ \text{CLIP}_{txt}(y_i)).$$

SA 用于衡量样本是否语义干净、是否符合类别定义，是一个强鲁棒的静态指标，可以利用开源 CLIP 模型的海量分布外知识。

3.2 多样性覆盖度（Diversity, Div）

本模块衡量样本在类内空间中的“覆盖程度”。可采用如下公式（基于 ε-sample-cover）：

$$\text{Div}_i = \exp\left(-\frac{\text{density}(x_i)}{\varepsilon}\right),$$

其中 density 基于 CLIP 特征的局部密度估计（如 kNN 或核密度估计）。

该指标体现样本是否覆盖了训练集中稀疏区域，从而对模型提供额外信息。

3.3 类内难度方向得分（Difficulty Direction Score, DDS）

对每个类别做 PCA，得到主轴 u_k 及方差 λ_k。低方差方向代表类内“罕见但重要”的变化维度。定义：

$$\mathrm{DDS}_i = \sum_{k \in \text{low-variance}} |\langle f(x_i) - \mu_c,\ u_k \rangle| .$$

该指标捕捉样本在关键变化方向上的贡献，衡量其边界信息量和泛化价值。

4 从训练动态中学习评分权重

静态评分指标的重要性在不同数据集和任务中不尽相同。本项目通过一次代理模型训练动态信息，学习各指标的权重，使评分体系更具数据自适应性。

代理模型（如 ResNet-18）训练 E 个 epoch，过程中记录每个样本 x_i 的训练动态，包括损失、预测正确性和 logits。

从训练动态中构造三个经过归一化的样本级指标：

4.1 Early Loss 指标

定义训练前 E_e 个 epoch 的鲁棒损失平均：

$$L_i^{early} = \frac{1}{E_e} \sum_{t=1}^{E_e} \log(1 + \ell_t(i)).$$

归一化得：

$$\mathrm{EarlyLossScore}_i \in [0, 1].$$

反映样本的整体学习难度。

4.2 Margin 指标

定义每轮 margin：

$$m_t(i) = z_{t,y_i}(i) - \max_{c \neq y_i} z_{t,c}(i).$$

并分三类打分：

$$s_t(i) = \begin{cases} -1, & c_t(i) = 0, \\ +1, & c_t(i) = 1 \text{ 且 } m_t(i) \leq \delta, \\ 0, & c_t(i) = 1 \text{ 且 } m_t(i) > \delta. \end{cases}$$

最终得到：

$$\text{MarginScore}_i = \frac{\frac{1}{E}\sum_{t=1}^{E} s_t(i) + 1}{2}.$$

该指标专门反映样本在训练中停留于边界附近的时间比例。

4.3 Forgetting 指标

定义整体正确率：

$$r_i = \frac{1}{E}\sum_{t=1}^{E} c_t(i),$$

统计后半段遗忘次数：

$$F_i = \#\{t > E/2 : c_t(i) = 1,\ c_{t+1}(i) = 0\}.$$

归一化 F_i 后，将 $(r_i, \tilde{F}_i)$ 映射为最终分数 $\text{ForgettingScore}_i \in [0, 1]$。该模块用于区分四类样本：简单样本、有价值难例、噪声样本、不稳定样本。

4.4 综合效用标签 u_i

三个指标归一化后取算术平均：

$$u_i = \frac{1}{3}\big(\text{EarlyLossScore}_i + \text{MarginScore}_i + \text{ForgettingScore}_i\big).$$

u_i 是一个连续的训练动态效用标签，刻画样本对模型训练的贡献。

4.5 使用线性回归学习指标权重

对每个样本，我们已有静态特征 $f_i = (\text{SA}_i, \text{Div}_i, \text{DDS}_i)$，目标值 u_i。

定义 Ridge 回归损失：

$$\mathcal{L}(w, b) = \frac{1}{N}\sum_{i=1}^{N}(w^\top f_i + b - u_i)^2 + \lambda\|w\|_2^2.$$

训练得到 w 后，对权重进行非负截断并归一化：

$$\tilde{w}_k = \frac{\max(w_k, 0)}{\sum_{j=1}^{3}\max(w_j, 0)}.$$

从而得到最终评分权重。

5 Selection Optimization

在获得最终评分 Score(x_i) 后，可以使用 Selection Optimization 进一步优化选样策略。

定义可学习选择变量 d_i：

$$s_i = \sigma(d_i),$$

其中 σ 为 sigmoid。

优化目标：

$$L = -\sum_i s_i \cdot \text{Score}(x_i) + \beta L_s,$$

其中 L_s 为比例约束损失，保证最终选样比例与设定值一致。使用直通估计器（STE）处理 0/1 选择的不可导性。

该步骤用以提升数据选择性能。

6 总结

本项目提出了一个统一的图像分类样本评估框架，实现了以下目标：

- 基于 CLIP 的可解释、多维静态评分体系；
- 通过一次代理模型训练动态学习指标权重；
- 对训练集与未见样本均可直接给出质量评分；
- 可作为数据选择方法使用，并支持 Selection Optimization。

该框架既具理论意义，又具工程可行性，可用于数据集优化、模型训练高效化等多个方向。