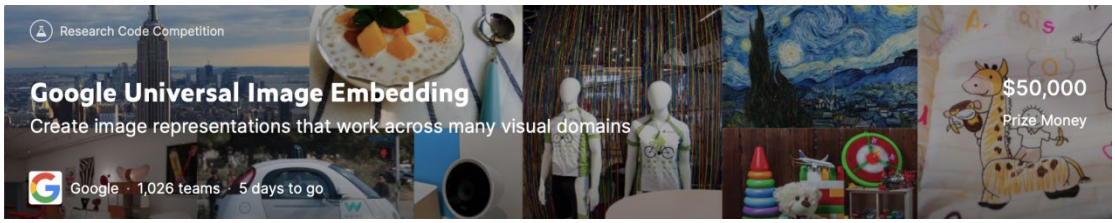
Google Universal Image Embedding

方案总结



1. 比赛简介

本次比赛为多领域图像嵌入比赛，要求参赛选手提交一个完整的模型，该模型能够将输入图像用一组纬度不超过 64 的向量表示。具体来说，搭建一个特征提取网络模型，该模型提取图像特征后用一组不超过64维度的特征向量表征该图像主要的实例级特征。（实例级就是说同样的一个物体，比如同一个人，同一只狗）

- 数据：该比赛不提供任何数据，选手自由收集任何在该比赛允许使用的数据。
- 评价指标：在测试时通过提交的模型提取图像表征向量，使用 KNN 计算出 mp@5，其中使用欧式距离对 index 中每一张图像相对 query 中每一张图像的距离进行度量。

2. 思路梗概

这个比赛可以用很多种方式完成，比如对比学习，实例级图像识别、检索等等这一类专注于实例级特征学习的方法。收集到许多实例级标注的数据后，我最终主要采用了实例级图像识别的方法。比赛主要集中于数据处理和模型结构设计。

3. 数据处理

(1) 数据收集过程中，我收集了比如 ImageNet 这类非实例级的数据集，在最后统一处理的时候这类数据都丢掉了，保留与实例级任务有关的数据集。

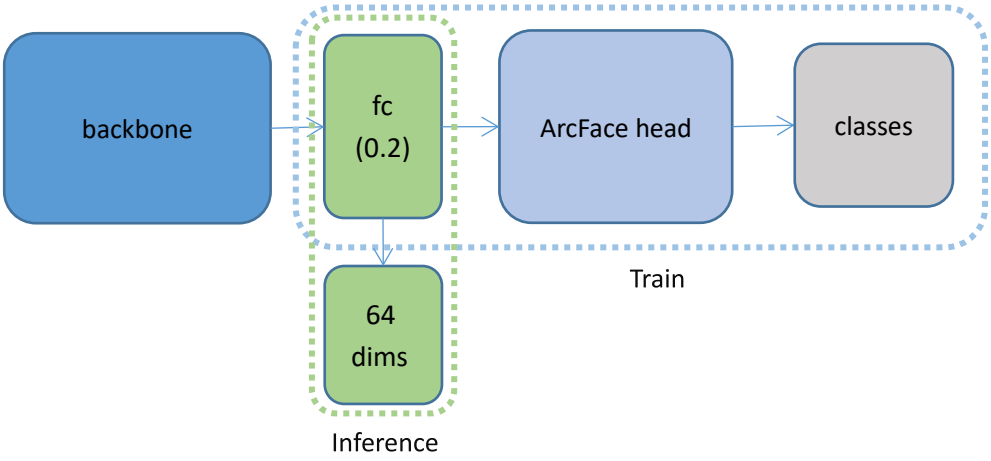
(2) 种类之间图像数量差距巨大，有的可能一两张，有的可能超过一千张，并且由于官方没有给训练集，所以我收集的数据的分布和测试集数据分布 99.9%概率是不同的，为了使模型能够对训练集里面各种类别的特征都能充分学习，对数据进行了采样以平衡类间权重，对所有数量大于 55 张的类别都只采样 55 张图像。

(3) 在图像进入模型开始训练之前，使用 clip 官方的 mean 和 std 进行标准化，减去 mean 除以 std，将图像像素值标准化为均值为 0，方差为 1，加速训练收敛。

4. 模型搭建

主干网络使用 open_clip[1]中的 vit-H，该模型使用对比学习的方式在 20 亿的图像文本对上进行了预训练，对于图像有着非常鲁棒的表征能力，与比赛的目标：多领域图像实例级特征表征 相契合。

为了不破坏主干网络的表征能力并加快训练速度，训练时将主干网络参数冻结，只训练模型后半部分：一个 dropout 概率为 0.2 的全连接层+一个 ArcFace[2]分类头。



推理时只需要用到 backbone 和 fc 模块以得到 64 维的表征向量，所以推理时稍微修改 forward 函数可以直接把 head 部分丢掉。

5. 整体流程

(1) 收集数据

从比赛讨论区大家分享的数据集链接选择下载需要的数据。

(2) 数据处理

去除与比赛任务不相关的数据、对图像数量较多的类别进行采样。

(3) 模型构建

使用 open_clip_vit_H 预训练模型。使用一层 fc 对主干网络提取的特征降维到 64dims 以满足比赛模型的需要，使用 ArcFace 作为分类头，增加类间距离，类内更加紧凑。使模型能够主动学习细节特征。

(4) TTA(Test Time Augmentation)

推理的时候使用 TTA 以增强模型稳定性。

6. 相关链接

[1] open_clip,

https://github.com/mlfoundations/open_clip

[2] Arcface,

https://blog.csdn.net/qq_45434461/article/details/115569426