说明

yuchengye

日期：December 8, 2021

目录

1 说明

1.1 数据

数据包括 openai-100k, math 和 chemistry

- root_dir

  images openai-100k 的图片

  labels openai-100k 的标注

  adaptation_data math 和 chemistry

  math_images_resized math 图片

  chemistry_images_resized chemistry 图片

  math_labels math 标注

  – train

  – val

  – test

  chemistry_labels chemistry 标注

  – train

  – val

  – test

  pre_vocab.txt 用来给 label 分词的，最终建立的 vocab 要更大一些。

还有一些为了减少加载数据时间的.cache 文件

openai-100k 的图片的尺寸本来就只有 19 种，并不多。

math 和 chemistry 图片的尺寸基本上每一张都不一样，这样一个 batch 里的图片尺寸可能都不一样，导致过滤完只剩第一张图片，影响训练效果。因此 kmeans 聚出了 32 个尺寸，将图片分别 resize 到对应尺寸，可以保证宽高比变化很小，文字畸变很少。

1.2 train.py

假设数据解压到了 `/home/featurize/data` ,则修改 `train.py` 27 行的 `root_dir` 为 `"/home/featurize/data"`

装依赖

```
pip install -r requirements.txt
```

训练

```
python3 train.py --cache --fp16 --batchsz 64 --mean_teacher --train_on math
```

如果带–cache 运行提示找不到.cache 文件或者提示找不到图片，那就去掉–cache 再运行一次就可以了。

参数含义：

–cache 使用.cache 文件加载数据。如果不指定此参数，会从头加载一遍数据，并生成新的.cache 文件，下次训练可以用。

–fp16 混合精度训练。提高训练速度，减少显存占用。建议开启。

–batchsz batch size

–mean_teacher mean teacher 方法，推荐开启 (必须打开)，多数情况下提升显著。具体是什么，可以点此查看原文或者搜索。

–train_on 在 [openai（默认）| math | chemistry] 上训练

另外还有参数：

–Adam 使用 Adam 优化器。不指定的话用的是 SGD。不推荐使用 Adam，虽然收敛快，但不一定有 SGD 收敛的好，而且会占用相当一部分额外的显存，速度也必将慢。如果优化器是 SGD，学习率衰减器使用的是 OneCycle，会先从一个较低学习率线性增长到 initlr，然后开始做类似指数递减。如果优化器是 Adam，学习率衰减器用的是指数递减。

–initlr 初始学习率。默认 1e-3。如果使用 Adam，可以设小一些，比如 1e-4。

–split_batch 不推荐使用，很慢。

1.3 proc.py

主要是 `kmeans_for_img_size(img_root, caching, resize)` 函数，根据 img_root 下的图片的尺寸聚类出 32 个尺寸，并将图片 resize 到对应尺寸

压缩包中的图片已经经此处理。

1.4 data.py

加载数据

1.5 non_local_emb_gaussian.py

一种带全局注意力的卷积操作。原文 以及 pytorch 实现

1.6 transformer.py

模型 CNN+transformer。CNN 是 efficientnet_b0。在其中加了两层 non-local block，并修改了最后全连接层的 out_features。

1.7 test.py

在 test 集上测试
在 28 行设置 root_dir 为数据解压到的位置，和 train.py 是一样的
在 116 行设置测试的权重文件。

```
python3 test.py --cache --fp16 --mean_teacher --test_on math
```

有可能测试集还没有生成.cache 文件或者提示找不到图片，去掉–cache 再运行一次就可以了。