

模式识别与机器学习大作业： 衣物颜色匹配

- 1) 每位同学须独立完成大作业，请在网络学堂提交电子版
- 2) 毕业班同学在 6 月 12 日 23:59:59, 非毕业班同学在 6 月 26 日 23:59:59 前提交大作业，不接受补交
- 3) 如有疑问请微信或邮件联系助教

吴文绪: wuwx21@mails.tsinghua.edu.cn

李嘉琦: lijq19@mails.tsinghua.edu.cn

颜钱明: yanqm18@mails.tsinghua.edu.cn

蔡昊晓: chx21@mails.tsinghua.edu.cn

目录

1 赛题描述	1
1.1 背景介绍	1
1.2 问题形式化	2
1.3 数据细节	2
2 作业安排与要求	3
2.1 代码报告 (60 分)	3
2.2 模型性能 (40 分)	4
2.2.1 绝对性能 (20 分)	4
2.2.2 性能排名 (20 分)	4
3 参考方案	5
3.1 用规则处理文字标签，用模型处理图片的方案	5
3.2 文字标签和图片均用模型处理的方案	6
4 FAQ	6
A 评测网站使用说明	7
B 公用计算服务器使用说明	9
B.1 中央主楼远程 Jupyter 服务器	9
B.2 中央主楼实体服务器	9

在本次作业中，我们需要处理的是一个更为折衷的颜色识别和区分问题——假设商家上传一个衣服所有款式的图片，和可供选择的一些描述颜色的文字标签，请你设计一套算法，将这些图片和这些文字标签进行匹配。

1.2 问题形式化

- **输入：**一个商品的所有图片： $\mathcal{I} = \{I_1, I_2, \dots, I_n\}$ 以及可选的描述颜色的字符标签 $\mathcal{T} = \{T_1, \dots, T_m\}$
- **输出：**图片和文字标签的匹配关系，即需要输出 $(I_1, T_{I_1}), (I_2, T_{I_2}), \dots, (I_n, T_{I_n})$ 。自然地，每张图的匹配标签需要在可选颜色字符标签集中做选取，即需要满足约束 $\forall i \in \{1, 2, \dots, n\}, I_i \in \mathcal{I}, T_{I_i} \in \mathcal{T}$

1.3 数据细节

本次作业的数据集由淘宝网上随机爬取的约 2.7 万个商品 (总计约 20 万张图片) 组成。每个商品有大于等于 2 种的不同颜色款式。随机划分出约 5 千个商品作为测试集，余下数据进行发放。可供下载的版本有全尺寸、中等尺寸 (短边放缩为 224 像素)、小尺寸 (短边缩放为 50 像素) 的版本¹。发放的数据集结构如下：

数据集文件夹结构

```
1  ./
2  |— 543437665996          # 商品id
3  |   |— 543437665996_0.jpg # 图片
4  |   |— 543437665996_1.jpg
5  |   |— 543437665996_2.jpg
6  |   |— ...
7  |   |— 543437665996_6.jpg
8  |   |— profile.json      # 可选颜色标签以及标签图片的匹配关系
9  |— 543447755424
10 |   |— 543447755424_0.jpg
11 |   |— 543447755424_1.jpg
12 |   |— ...
13 |   |— 543447755424_6.jpg
14 |   |— profile.json
15 |— 543470439046
16 |— ...
17 |— train_all.json        # 训练集各商品profile.json的合并
18 |— test_all.json         # 测试集各商品profile.json的合并
```

¹下载链接：<https://cloud.tsinghua.edu.cn/d/27849370d8774de3a2e2/> 其中全尺寸 (前缀 full) 数据集为分 4 卷压缩，中等尺寸 (前缀 medium) 和小尺寸 (前缀 thumbnail) 数据集各有一个压缩包。下载链接中还有这些压缩文件的 md5 可供下载后校验。解压后约使用空间：全尺寸 15G，中等尺寸 5G，小尺寸 1G。

profile.json 记录可选标签和匹配关系的格式

```
1 {
2   "optional_tags": [ # 可选标签
3     "绛红",
4     "米色"
5   ],
6   "imgs_tags": [    # 各图与标签的匹配关系
7                     # 测试集内此列表内的字典取值为{图片文件名: null}
8     {
9       "543437665996_0.jpg": "绛红"
10    },
11    {
12      "543437665996_1.jpg": "绛红"
13    },
14    ...,
15    {
16      "543437665996_6.jpg": "米色"
17    }
18  ]
19 }
```

此外，在本次作业中我们还做一定的约束和简化：

1. 所给的可选标签的个数 m 一定小于等于图片数 n 。
2. 在 ground truth 中，任意一个可选标签下一定存在至少一张匹配的图片

2 作业安排与要求

本次作业满分 100 分，由代码报告、模型性能两部分组成。

2.1 代码报告 (60 分)

本部分主要考虑代码的正确性以及报告的规范性和完整性。

代码部分：需要提交完整的训练测试代码。代码中要求用相对路径以方便复现。另外需要附一个 README.md 说明：

1. 简述代码部分各部分文件作用
2. 说明代码运行环境和库的版本
3. 数据集在项目目录的放置位置
4. 提供复现你的最佳模型的训练过程命令以及对应地在测试集上测试的命令。按此命令训练出的模型，不应当与你评测网站上上传的最终结果差距过大。

完成作业过程中可以参考已有的代码，但要在代码和报告中相应部分给出引用说明。否则会影响代码查重结果和本部分得分。

报告部分：报告部分需要展示你对本问题的分析和解决方案。报告建议包括如下章节：问题分析与形式化、数据处理流程、算法原理、实现过程、实验结果、结果分析等。报告部分同样需要查重。

提交要求：本部分内容以 zip 压缩包的形式提交到网络学堂。提交的压缩包的命名格式要求为：“班号-名字-学号.zip”。压缩包内的目录和文件均采用英文命名防止因操作系统差异造成乱码；另外压缩包内不应该包含数据集，如果你对数据集做了静态的扩展或修改，请上传到清华云盘并在报告中附下载链接。最后压缩包内结构的要求为：

提交的作业目录结构要求

```

1  ./
2  |— codes                # 项目目录
3  |   |— README.md       # 代码运行说明
4  |   |— ...              # 代码，可包含多个文件、目录
5  |— report.pdf           # 作业报告

```

2.2 模型性能 (40 分)

本部分分数以算法方案在预先划分的测试集上的性能表现综合给出。所包含的指标有

1. 图片文字匹配准确率 (Accuracy): $\text{Acc} = \frac{\text{\#correct image-text pairs}}{\text{\#images}}$
2. 商品完全匹配率 (Exact Match): $\text{EM} = \frac{\text{\#products all matched}}{\text{\#products}}$

2.2.1 绝对性能 (20 分)

图片文字匹配准确率达到不同水平时分别得到如下分数

Acc	≤ 0.3	$0.3 \sim 0.7$	$0.7 \sim 0.75$	≥ 0.75
分数	0	$0 + \frac{8-0}{0.7-0.3} \times (\text{Acc} - 0.3)$	$8 + \frac{10-8}{0.75-0.7} \times (\text{Acc} - 0.7)$	10

商品全匹配率达到不同水平时分别得到如下分数

EM	≤ 0.3	$0.3 \sim 0.5$	$0.5 \sim 0.6$	≥ 0.6
分数	0	$0 + \frac{6-0}{0.5-0.3} \times (\text{EM} - 0.3)$	$6 + \frac{10-6}{0.6-0.5} \times (\text{EM} - 0.5)$	10

2.2.2 性能排名 (20 分)

我们开放了一个评测网站用于上传你的预测结果。具体使用方法见附录A。提交的预测结果的格式参考 `train_all.json` 或 `test_all.json`。我们会对保留 4 位小数后的两个指标分别维护一个排名榜。你的综合排名为 Acc 和 EM 两个榜中的最高排名。

最终，本部分分数将由综合排名线性给出，即得分为 $20 \times (1 - \text{排名百分位数})$

3 参考方案

抽象地说，本次任务是一个较为粗粒度的图像和文本两模态数据匹配的任务。所以按照两个模态数据的处理方法是基于规则还是基于模型，可以得到不同方案。

3.1 用规则处理文字标签，用模型处理图片的方案

类似淘宝的标准方案，首先定义一个标准颜色词库，然后据此设计文字标签和标准色库的匹配方案，利用标准色库将文字标签转为可以标准的“数字标签”。如下示例的文字标签和标准色库的匹配方案就是“标准色库的字词在文件标签中，即认为匹配”

```
1 def tag_to_label(tag):
2     for idx, standard_word in enumerate(STANDARD_LEXICON):
3         if standard_word in tag:
4             return idx
5     return len(STANDARD_LEXICON) # 未匹配中，所有未登录词处理为一个标签
6
7 labels = [tag_to_label(tag) for tag in tag_list]
```

转为标签后即可按照常规的分类任务，训练一个图像分类器。训练好图像分类器后，将算法部署到推理环节中，还需要思考怎么将分类器输出转为所需的“图文匹配关系”。

```
1 label_to_tag = {label:tag for tag in tag_list}
2 pred = torch.argmax(output, dim=-1) # 输出概率的argmax, 作为预测标签
3 result = [(i, label_to_tag[label]) for i, label in enumerate(pred)]
```

这套方案里，为了提升性能，你可能需要考虑的问题有：

1. 文本匹配规则的构建。如以上的简单匹配方法中，假设标准色库只有“红”，在可选标签是“红色、粉红色”之类的情況下，方案会完全失效。另一方面，可能存在一些不能匹配上你设定的标准色库的文字标签，需要处理。
2. 文本匹配规则和类别数的平衡。极端地说，你可以将训练集中的每个一文字标签单独对应一个数字标签，但这样得到的类别数会偏多，难以训练。
3. 类别不平衡问题。从词云分布不难看出，做如上的转换后，我们得到的类别间的样本数是不平衡的，你可能需要在训练模型时加以考虑。
4. 不同商品间，相似标签在描述不同现象。想象这样的两种有“蓝色、黑色”两个标签的商品。一种是纯蓝/黑色的短袖，另一种是白底有蓝色/黑色条纹的短袖。对于后者，很有可能是白色的类别预测概率最高，而蓝色/黑色预测概率都被放缩到较小且难以比较数量上。

3.2 文字标签和图片均用模型处理的方案

使用模型分别提取颜色字符标签和图片的特征向量，并做特征向量之间的匹配

```
1 img_features = [model_img(img) for img in img_list]
2 tag_features = [model_text(tag) for tag in tag_list]
3 match_scores = match(img_features, tag_features)
4 # 处理match_scores变为最终结果
```

这套方案里，你可能需要考虑的问题还有：

1. 文字标签的预处理。
2. 两模态特征数值大小差异，可能需要一定的归一化处理
3. 两模态特征提取网络训练速度的差异。可能需要分别调节两模态特征提取网络的学习率等优化参数
4. 两模态特征向量匹配的方案。如计算特征向量之间的内积作为相似度，又或把图文特征向量一并输入一个“匹配模型”做二分类，又或者先使用图片特征向量做聚为“可选标签个数”个类别，再将聚类类别和标签做一一地匹配。

4 FAQ

1. Q: 能否使用模型库中的模型，可否使用对应的预训练权重？

A: 都可以。此外，建议在报告中分析你使用模型的特点。

2. Q: 可否自行增删数据集

A: 可以。可以仿照助教爬取数据的流程或自己想办法 (如自己下载合作标注)。若采用了这类静态的扩展数据集的方法，需要在报告中分析扩增前后模型性能的变化，并标注出其他人的贡献 (如果你采用合作标注等方案)。但若你没有相关经验，不建议在本次任务中尝试，因为 1) 所给训练集有约 2 万款商品，15 万张图片，在助教预测的几种方案中数据量都是饱和的 2) 淘宝的反爬虫机制很强，初学者不容易处理。

A 评测网站使用说明

网站账号为你的学号, 初始密码详见外部表格。每人可用的提交次数 60 次。

另外为了大家有充足的时间撰写报告, 我们会在你作业提交日期前一天, 也即非毕业班同学的 6 月 25 日 23:59:59, 毕业班同学的 6 月 11 日 23:59:59 进行排名结算。排名分位数以结算时为准。

网站开放地址为<http://vsimu.au.tsinghua.edu.cn/prml2022/index>。第一次登录后会跳转到相应的登录页面 (在最终部署版本中, 还需填写 4 位区分大小写的验证码)

模式识别与机器学习2022春大作业评测 -- 用户登录



学号
密码
登录 重置

图 3: 评测网站登录界面

输入账号密码后进入主界面。你可以在左侧修改账户设置, 右侧提交待评测的预测结果, 和选择最终参与排名的结果。下侧可以看到本次作业的排行榜。



模式识别与机器学习2022春大作业评测

当前用户: 2021210977
账户设置修改

预测提交: 剩余提交机会 60 次

选择文件 未选择任何文件
提交 校验

名称
新密码
重置新密码
修改账户设置 重置 登出

标记为最终

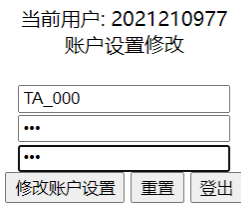
排名	名称	Acc	EM
1	TA_002	0.9872	0.9328
2	TA_001	0.8673	0.6155

图 4: 评测网站主界面

- 账户设置修改部分

这部分可以修改账户的名称和密码。账号的名称和账号不同, 名称是用于排行榜显示, 账号也即你的学号用于登录。为了防止在排行榜上直接显示学号可能会带来的麻烦, 我们提供了名称修改接口。你的初始名称同账号, 可修改为由字母数字下划线组成的字符串。

修改密码部分, 你修改的密码强度必须足够才能通过。具体的要求为长度至少 8 位, 包含数字, 字母, 以及 +-* / 四个字符中的至少一个。这是为了响应学校信息部门的安全要求, 网络安全造福你我他, 希望大家理解。



当前用户: 2021210977
账户设置修改

TA_000
...
...

修改账户设置 重置 登出

图 5: 账户设置修改示意

- 预测提交部分

提交部分，你可以看到你剩余的提交机会。在下方选择文件窗口选择你的预测结果并上传。点选“校验”会校验上传结果的格式，不消耗提交机会，方便大家在项目初期校验自己程序的输出格式。点选“提交”会正式提交结果，消耗一次提交机会，系统会返回这次预测的两个评测指标。你会在下方的复选框中看到你的所有提交的概况，包含提交的时间，提交结果的 Acc 和 EM。

预测提交: 剩余提交机会 59 次

未选择任何文件

0511-100642_Acc_0.8839_EM_0.6243

图 6: 预测提交和返回结果

本次作业中我们有两个评价指标，两个指标有一定的相关关系。但可能存在的一种情况是：你的两次结果中，某一个结果相对另一个结果，Acc 更高，EM 却更低。因为你的综合排名是两个指标榜上排名的最高值，在这种情况下，你可能需要根据其他人的情况，选择对你的排名更有利的结果。因此，我们开放了接口，让你决定你最后参与排行的预测。标记或修改最终结果后，你排名榜的位置应该会有相应的更新

模式识别与机器学习2022春大作业评测

当前用户: TA_000
账户设置修改

预测提交: 剩余提交机会 59 次

未选择任何文件

0511-100642_Acc_0.8839_EM_0.6243

排名	名称	Acc	EM
1	TA_002	0.9872	0.9328
2	TA_000	0.8839	0.6243
3	TA_001	0.8673	0.6155

图 7: 最终结果标记和排行榜情况

- 排行榜部分

评测网站默认的排行指标是 Acc。你可以通过点击表头的 Acc 或 EM，修改排行榜上显示的排名所依据的指标，以掌握自己结果的综合情况。

B 公用计算服务器使用说明

需要使用公用服务器的同学请进计算资源群，一些公用资源的调配事宜，以及需要在服务器上安装特定版本的库可以在此交流。



图 8: 计算资源群二维码 (若二维码过期请联系助教入群)

B.1 中央主楼远程 Jupyter 服务器

服务器地址:

http://166.111.72.3/jupyter_ai5

http://166.111.72.3/jupyter_ai105

http://166.111.72.3/jupyter_ai201

http://166.111.72.3/jupyter_ai203

密码:zdhxai2020

服务器上已经有原始分辨率版本的数据集，无需每个用户再次上传。登录服务器后，请先建立自己的文件夹，在自己的用户文件夹下编写代码。

B.2 中央主楼实体服务器

待疫情情况好转后可能开放，具体事宜另行通知。