无约束人脸识别系统设计

李宇飏
清华大学自动化系
liyuyang20@mails.tsinghua.edu.cn

Abstract

我们使用 CNN 模型与几种度量学习的损失函数完成一个无约束的人脸识别系统。我们首先通过已有的库简单处理人脸图像，标注人脸的 bounding box，然后通过训练基于 CNN 的模型学习将图像映射到特征空间。通过三种度量学习 loss，模型将学习把同一人的照片映射到一簇，将不同人的照片映射到可相互区分的不同区域。我们通过 t-SNE 验证这一效果。本大作业的代码见随作业提交的代码及 `README.md`，或见 `https://go.yuyangli.com/prml/face`（作业提交截止后开源）。

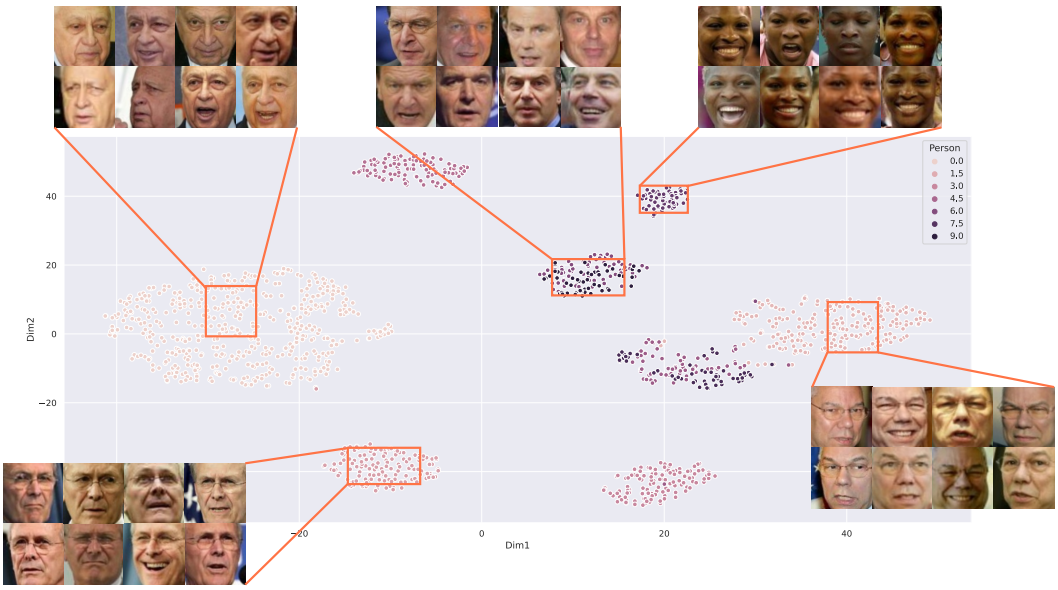


图 1: **使用本文所述方法学习的模型能够将图像映射到特征空间中的点，上图为部分样本点的 t-SNE 降维结果**。可以看到，同一人的图像聚成一簇，绝大多数不同人的图像相互可分。其中有 2 个簇无法在 t-SNE 的 2 维降维结果中区分，对应的两组人像（中上）也确实具有相似的面容。

1 介绍

人脸识别任务（Face Recognition）在人员管理、安防保障等领域发挥着愈加重要的作用。分辨人脸对于人类来说是个很自然和轻松的工作，但想要通过算法实现人脸识别

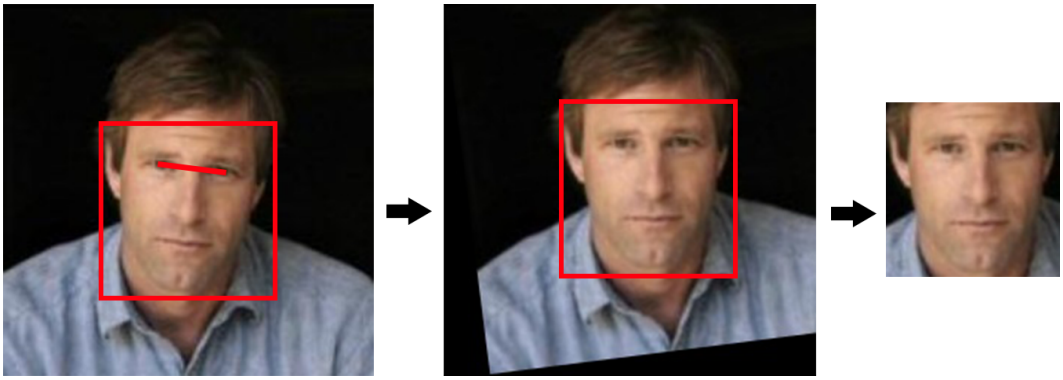


图 2: **人脸数据预处理**。包含人脸 BB 检测、眼部水平处理、人脸 BB 再检测。

却并不简单，主要难点包括但不限于 1) 图像质量参差不齐，亮度、对比度、背景等条件不同；2) 人脸在面部肌肉不同、面部配饰（如眼镜）不同等状态下外观差异较大；3) 人脸在图像中的位置、大小、投影角度、被遮蔽情况不同。

本次大作业设计实现“无约束人脸图像识别系统”。给定两张人脸图片，需要判断二者是否属于同一个人。

2 方法

我们首先对图像进行预处理，然后通过 CNN 提取经过预处理的图像的特征，采用度量学习的方法，通过优化模型参数 θ 与判定阈值 τ，使特征空间中同一人照片的特征距离较近，而不同人照片的特征距离较远。在对图像进行分类时，计算两张图片特征的距离，若小于 τ 则认为是同一人的照片，否则认识为不同人的照片。

2.1 数据处理

为了便于模型提取特征，我们依据人面部的器官标志点位对图像进行仿射变换，并剪裁出必要的图像区域，统一分辨率。

首先采用 `dlib` 在所有图片中检测人脸 bounding box，然后继续调用该库，根据该 bounding box 进一步标注人脸上的关键点坐标（landscape）。为了便于模型学习，我们根据两眼坐标 $(x_L, y_L), (x_R, y_R)$ 计算旋转 $\theta = \arctan \frac{y_R - y_L}{x_R - x_L}$，并计算对应的二维仿射矩阵 R。变换图像 $I \leftarrow RI$ 即可将两眼连线转正。此时，bounding box 也被旋转。我们重新检测人脸 bounding box。此 bounding box 内的图像经过 resize、normalize 等操作（见 章节. 2.4）后作为模型输入。如图 图. 2 所示。

2.2 模型设计

我们采用基于图像模型从图像中提取特征。对于图像数据 x，模型的参数集为 θ，它将图像映射到 $L-$ 维特征空间的流形上：$f : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^L$。我们希望学习该映射，使得流形上的距离 $d(x_1, x_2) = \|f_\theta(x_1) - f_\theta(x_2)\|_2$ 可以表征图像与身份的关系。具体地，同一个人的照片的该距离较小，不同人的较大，且可以通过某个阈值 τ 区分。

当今，视觉领域有众多优秀的视觉模型可以用来为图片提取特征，包括但不限于基于 CNN 的 VGG [1]、ResNet [2] 等，以及基于 Transformer 的 Vision Transformer 系列。但在人脸识别这一度量学习任务中，对度量空间与学习目标的设计远比特征提取重要。本次大作业不针对过多模型展开对比实验，仅简单对比 ResNet34 [2] 和 VGG11-BN [1]。该模型从图像中提取特征，得到特征向量：

$$\tilde{z}_i = f_\theta(x_i) \tag{1}$$

为了防止过拟合，我们使用 Dropout 与一个 FC 层处理特征向量：

$$z_i = \mathrm{FC}\left(\mathrm{Dropout}_{p=0.5}(\tilde{z}_i)\right) \in \mathbb{R}^L \tag{2}$$

2.3 训练目标

我们为上述的模型设计训练 loss function，并通过优化器寻找将其最小化的模型参数 θ：

$$\theta^* = \arg\min_\theta \mathcal{L} \tag{3}$$

$$\mathcal{L} = \lambda_m L_m + \lambda_n \mathcal{L}_n + \lambda_r \mathcal{L}_r \tag{4}$$

其中，Regularization 项 $\mathcal{L}_r$ 计算模型参数的 L2 Norm，减小过拟合风险：

$$\mathcal{L}_r = \|\theta\|_2 \tag{5}$$

流形约束项 $\mathcal{L}_n$ 鼓励 feature 在 R^L 上的单位球上：

$$\mathcal{L}_n = (\|f(x)\|_2 - 1)^2 \tag{6}$$

2.3.1 Triplet Loss

在人脸数据中，同一人的两张照片被认为是正样本对 (x, x^+)，不同人的两张照片被认为是负样本对 (x, x^-)。从一个样本 x 出发的一正一负两个样本对可以组成 triplet（三元组）(x, x_+, x_-)。Triplet Loss [3]（三元组损失）的思想是让特征空间中，负样本对的距离应当比正样本对远，且超过一个 margin 值 δ，从而可区分：

$$\mathcal{L}_t = \max(d(x, x_+) - d(x, x_-) + \delta, 0) \tag{7}$$

选择合适的样本对 (x, x^+, x^-) 至关重要。从数据集中随机选择三元组是最简单的方案，并行实现简单；但随着训练进行，很多随机选出的 triplet 已经满足 margin，并不会更新网络参数，导致大量的运算被浪费，降低效率；同时，这容易使网络陷入 local minima，即在海量的正负样本对中，极少数被 sample 出来的对满足上述距离关系，但整个特征空间依旧不够好。

研究人员提出两种选取样本对的方法 [3]：(1) 从数据集中找到距离最远的正样本对、最近的负样本对，组成三元组；但由于数据集过大，这个操作的计算量过大，只能 offline 地每隔 n 步选取。(2) 从当前 mini-batch 种找到距离最远（最难）的正样本对、最近（最难）的负样本对，组成三元组；这个方法的计算量相对较小，可以 online 地选取样本对。但在实践中，如果从训练开始就选择最难的负样本对，训练会非常不稳定。最后，我们对比两种 Triplet Loss：

Weak Triplet Loss 直接在数据中采样正负样本对，根据 Eq. (7) 计算，而不根据难度特地挑选；

Triplet Loss 在 minibatch 中计算所有负样本对 $\left\{x^{(i)}, x_-^{(i)}\right\}_{i=1}^{N_-}$ 的距离 $d_-^{(i)} = d(x^{(i)}, x_-^{(i)})$，并以 $\mathrm{Softmin}\left(d_-^{(i)}\right)$ 加权求和，作为 Eq. (7) 中的 $d(x, x_-)$ 项。

2.3.2 Lifted Structure Loss

Triplet loss 只利用训练数据中的难样本对的信息, 这样的计算效率并不高; 此外, 单个负样本对很难代表全数据集的统计信息，而对于一个样本，若能利用与之有关的多个负样本对一起计算梯度，则梯度更接近使用整个数据集计算该样本的负样本对的梯度。

Lifted Structure Loss [4] 使用 mini-batch 内所有的样本对计算，进一步提高了训练效率。其思想是在一 mini-batch 数据中，除了正样本对外，任意两张图像配对均可视为负样本对，因此应当最小化：

$$J = \frac{1}{2|\hat{\mathcal{P}}|} \sum_{(i,j)\in\hat{\mathcal{P}}} \max(0, J_{ij})^2 \tag{8}$$

$$J_{ij} = \max\left(\max_{(i,k)\in\hat{\mathcal{N}}} \alpha - d(x_i, x_k), \max_{(j,l)\in\hat{\mathcal{N}}} \alpha - d(x_j, x_l)\right) + d(x_i, x_j) \tag{9}$$

但实践表明这个函数不够光滑，不便于训练。参考 [4]，我们优化它的一个平滑上界，即令：

$$J_{ij} = \log\left[\sum_{(i,k)\in\hat{\mathcal{N}}} \exp\left(\delta - d(x_i, x_k)\right) + \sum_{(j,l)\in\hat{\mathcal{N}}} \exp\left(\delta - d(x_j, x_l)\right)\right] + d(x_i, x_j) \tag{10}$$

图像模型	损失函数	T Acc. / % ↑	FN / % ↓	FP / % ↓
ResNet34 [2]	Weak Triplet	80.65	7.83	11.52
	Triplet	86.22	7.51	8.27
	LiftedStructure	**87.36**	**7.50**	**6.70**
VGG11-BN [1]	Triplet	**86.09**	5.52	**8.38**
	LiftedStructure	85.64	**5.04**	9.32

表 1: **Baseline 训练结果（训练集/验证集）**。 T Acc. 为 Triplet 准确度，FN 为假阳率，即认为 x, x_+ 为不同人的照片；FP 为纳伪率，即认为 x, x_- 为同一人的照片。加粗代表该组最优。由于不具有测试集标签，无法报告测试集性能。

2.4 数据增强

为了提高模型的鲁棒性，我们在训练中对图像数据施加一系列变换：1) 由于 CNN 一般不具有镜像不变性，我们将图像横向翻转。特别地，将 x 横向翻转，可以作为 Triplet 中的 x_+ 项；2) 随机的仿射变换，包含小范围的平移、旋转、剪切变换；3）Gaussian 加性噪声；4) Gaussian 模糊；5) 我们对三维的色彩进行变换，包括亮度、对比度的波动、灰度处理。

最后，图片的色彩维度被 normalize 到 $\mu = 0.5, \sigma = 0.5$。除了色彩维度归一化为确定性操作，其他的变换均以各自的概率 p 被施加在图像上。

2.5 人脸识别

在完成训练后，对于待对比人脸 x_1, x_2，选择合适的 τ，若 $d(x_1, x_2) < \tau$，则认为两张人脸为同一人的，否则为不同人的。显然，τ 与训练中，$f_\theta(x)$ 与 δ 的尺度均有关。我们使用验证集数据 $\mathcal{D}_V = \left\{x_0^{(i)}, x_1^{(i)}; y_i\right\}_{i=1}^{V}$ 找到最佳的 τ：

$$\tau_i = \arg\max_{\tau} \sum_{i=1}^{V} y_i \mathbb{1}_{d\left(f(x_0^{(i)}), f(x_1^{(i)})\right) < \tau} \tag{11}$$

3 实验分析

对于下文的实验，我们设置随机种子 `seed=42`，将给定的数据集随机选取约 500 对正样本、500 对负样本作为验证集，其余数据作为训练集。以 `batch_size = 256`，$H = W = 128, \sigma = 0.05, \alpha = 10°, p = 0.5$ 训练 200 epochs。在单张 NVIDIA RTX 3090 Ti 上，完成一次 baseline 训练约需要 40 min。

3.1 Baseline 实验

对于 baseline 模型（Eqs. (2) and (4)），我们给出使用 2 种模型在不同 loss 下的训练结果，见 表. 1 所示。

图 X 中展示了验证集中的 y 张人脸，及其在特征空间的最近、最远的 z 张人脸。可以观察到，我们的方法有能力学习一个良好的特征空间，即使照片中人像的衣着、面部

损失函数	T Acc. / % ↑	FN / % ↓	FP / % ↓
Full Method	**87.36**	**7.50**	**6.70**
w/o aug	79.09 (**- 8.27**)	10.28 (**+ 2.78**)	14.73 (**+ 8.03**)
w/o warm-up	86.28 (- 1.08)	7.44 (- 0.06)	7.03 (+ 0.33)

表 2: **消融实验结果**。实验采用 ResNet18 + Lifted Structure Loss，metric 同 表. 1。

配饰、光照、朝向、遮挡情况、人脸位置等有差异，特征空间中的距离依然可以代表照片的身份。

3.2 消融实验

我们以 ResNet18 + Lifted Structure Loss 为基准，分别针对 1) 数据增强（aug）、2) margin warm-up 开展消融实验。

结果表明，数据增强可以提升性能。这是因为数据增强增加了输入数据中与人脸识别语义关系较弱的因素（如面部朝向，照片亮度、对比度）等的随机性，使得模型能够学习到更重要的参考因素，并防止过拟合。尽管 warm-up 对性能没有太大影响，在实验中它能够提高训练稳定性。

3.3 特征空间降维

数据降维是将高维的数据空间转换到低维的子空间，从而缓解维度爆炸等问题。常见的降维手段包括但不限于 PCA、IsoMap、Stochastic Neighbor Embedding (SNE) 等。

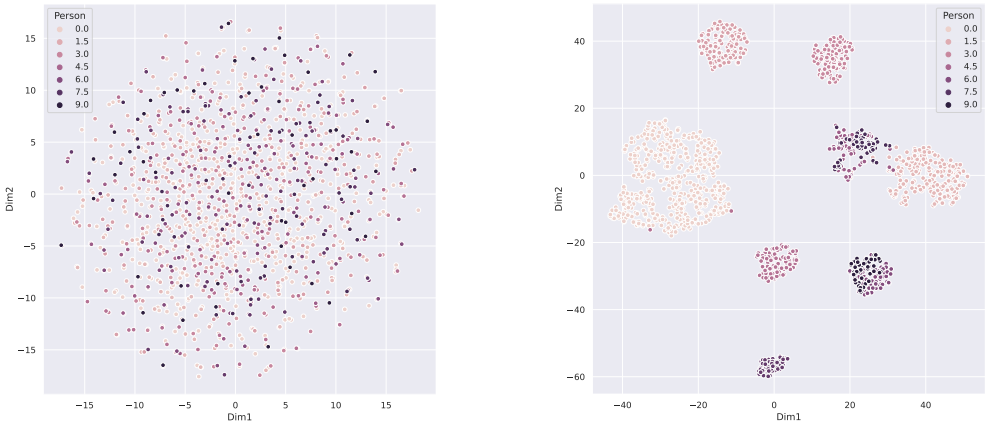
SNE 是一类非监督的降维方法，其核心思想是通过仿射变换，将数据映射到某个概率分布上。具体地，分别在高维和低维建立两个数据间的概率分布。在高维，对于每个样本，使其他样本的概率同其与该样本的相似程度正相关。在低维，构建一个概率分布使之与高维的分布尽可能接近。

但由于 SNE 存在“拥挤问题”，且迭代计算较困难，于是研究人员提出了优化版本 t-SNE [5] 。它是采用 t-分布的 SNE 在低维代替 Gaussian 分布的方法，简化了梯度计算，并通过 t-分布的长尾特性解决“拥挤问题”。

我们从训练集中抽取 10 个照片较多者的所有照片，使用 t-SNE 将特征空间里的投影 z_i 降维。如图 图. 3，可以看到在训练模型前，降维结果中各个人的样本几乎随机散布，无法分辨；训练后，同一个人的照片基本聚成一簇，较好分辨。本文的 teaser（图. 1）给出了具体的聚类结果示意图。

3.4 人脸匹配与近似人脸搜索

在完成训练后，模型还可以用来搜索近似人脸。给定训练数据库，若模型训练得当，它应当可以根据一张没有见过的人脸图片 $\tilde{x}$，在数据库里搜索与之最接近的人脸：



(a) **随机初始化的 t-SNE 结果**。 (b) **完成学习后的 t-SNE 结果**。

图 3: **学习前、后特征空间样本点的 t-SNE 结果**。可以看到，学习前，样本点分布随机、无法相互分辨。完成学些后，样本点能明显分开，有极少数样本点在降维结果中无法分开。

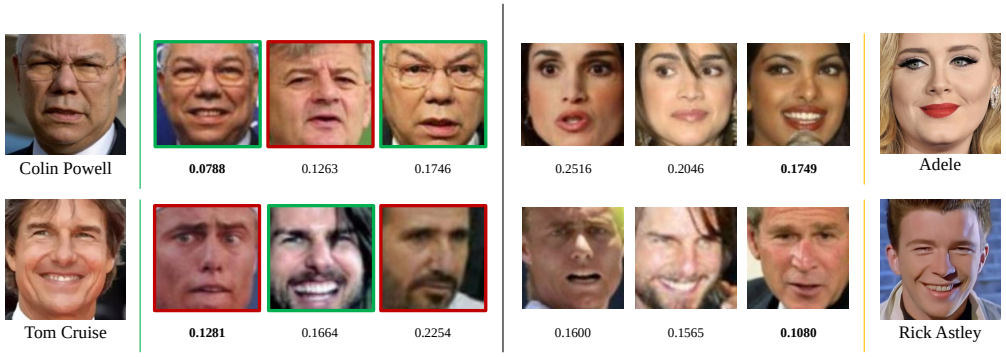


图 4: **人脸匹配（左）与相似人脸搜索（右）结果**。接近目标图像者在高维具有越接近的距离（见图下方数字）。正确、错误的匹配分别用红、绿色框标记。

$$x^* = \arg\min_{x \in \mathcal{D}} d(x, \tilde{x}) \tag{12}$$

人物匹配 我们从互联网上选取 2 张人脸照片，其人物出现在训练集中，但照片不包含在训练集中，通过 Eq. (12) 搜索该人物的人脸。

相似人脸搜索 我们从互联网上选取 2 张人脸照片，其人物不包含在训练集中，通过 Eq. (12) 搜索最匹配的人脸。

结果如图 图. 4 所示。可以看到，在给出的案例中，模型的搜索结果与原图具有相同的性别、相似的面容。但是人脸搜索结果依然有较大的错误，说明仅仅通过本文给出的方法，使用正负样本训练模型，还无法使特征空间足够精准与完善。

4 总结

在本次大作业中，我们实现了一个人脸识别系统。其核心是一个基于 CNN 的模型，将 RGB 图像映射到特征空间中；在该空间中，同一人的投影应当聚成一簇，不同人的投影簇应当可以相互分离。我们通过度量学习的手段学习这样的映射，并通过实验探究了不同损失函数设计的性能。

实验表明，损失函数、特征提取模型，以及样本间隔（margin）等超参数对于人脸识别的任务都有影响。VGG11-BN 相比 ResNet 34 要大一些，但性能并没有明显优于后者；而在实验中，使用 ResNet101 等更大、更复杂的网络，训练效果甚至反而更差，且训练过程更不稳定，说明网络的参数量与性能并不是正相关的。此外，几种 loss 在同一特征提取模型的实验中表现没有特别明显的优势 loss，但是 Weak Triplet Loss 相比 Triplet Loss 收敛更慢等现象验证了选取较难的样本对可以帮助网络学习。

此外，学习好的映射不仅可以用来进行人脸判断，还可以支持人脸检索等任务。

4.1 局限

由于时间等关系，本次大作业仍有诸多局限，包括但不限于：1) 没有考虑数据集中的 bias，如男女比例、正负样本比例；2) 使用了较为基础的度量学习方法，而没有参考较先进的对比学习方法，如 InfoNCE 等；3) 没有较好的 threshold 选择方法，较为先进的方法一般通过统计正负样本对的距离分布，并通过二者的均值、方差确定 threshold。

References

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 5

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3, 5

[3] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 3, 4

[4] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[5] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6