

# 人脸识别实验报告

罗瑶

2017213866

luo-y17@mails.tsinghua.edu.cn

吴超月

2017213865

wucy17@mails.tsinghua.edu.cn

刘春亚

2017213876

liuchuny17@mails.tsinghua.edu.cn

## 摘要

本次实验基于 Inception-Resnet-v1 对人脸图片进行特征提取，根据提取的特征计算测试图片与索引数据库中所有图片的距离，将距离最近的一张或五张图片判断为与测试图片同属于一个人。在搜索算法中，我们通过人脸仿射变换和数据增广提高人脸识别的准确率。首先，我们分别利用两种 loss 函数在 vgg 数据集上训练 Inception-Resnet-v1，两种 loss 分别是 softmax、center loss 结合 softmax。其中利用 softmax loss 训练的模型在 LFW 上的准确度可达 99.4%，在误配率为 0.1% 时，验证率可达 97.6%。而该模型在 1vn 人脸识别的任务中 top 1 命中率达到 51.3%，top 5 命中率达到 65.95%。利用 center loss 结合 softmax loss 训练的模型在 LFW 上的准确度可达 99%，在误配率为 0.1% 时，验证率可达 97%。而该模型在 1vn 人脸识别的任务中 top 1 命中率达到 64.7%，top 5 命中率达到 75.6%。我们发现在两种模型的人脸验证能力相差不大时，利用 center loss 结合 softmax loss 训练的模型具有更强的 1vn 人脸识别能力。我们试图利用三种数据集在利用 center loss + softmax loss 训练的模型上进行 finetune 来达到效果提升：2000 张图片，2000 张图片的增广数据集，增广数据集混合 VGG 数据集。但是 fine tune 的效果一般，偶然得到一个略有提升的模型在 1vn 人脸识别的任务中 top 1 命中率达到 64.8%，top 5 命中率达到 75.05%，虽然 top 1 上升，但是 top 5 下降了。最后我们进一步利用数据增广和人脸的仿射变换对 HW\_1\_Face 图片进行处理后再输入我们训练的模型，最终我们的 top 1 命中率达到了 68.1%，top 5 命中率达到 78.1%。

## 关键词

人脸识别、卷积神经网络、特征学习

## 1. 简介

人脸识别在公安系统、安防领域以及金融支付领域有广泛应用。在业界，性能最好的人脸识别系统大多基于深度卷积神经网络(CNN)，比如 DeepID、DeepFace、FaceNet，根据对 CNN 的用法，可以分为两种，一种将 CNN 当做分类器去训练，并将中间的全连接层输出作为图像的特征，另一种根据自定义损失函数的优化目标，将 CNN 作为特征嵌入模型训练，网络直接输出需要的特征。本次实验，我们采用第一种方法进行特征提取。

我们采用 Google 提出的网络模型，Inception-Resnet-v1，做为基本分类器。我们分别采用 softmax 损失函数和 center loss 结合 softmax loss 的损失函数辅助 Inception-Resnet-v1 在 VGG 数据集上做训练，预处理训练数据时我们采用 MTCNN 网络进行人脸识别和对齐。我们在提供的 HW\_1\_Face 数据集上用下标为 2 的图片做测试集计算模型的 top 1 命中率和 top 5 命中率，同时我们也在业界标准数据集 LFW 测试我们的网络，实验证明模型在两者上的表现具有一致性。

我们分别比较了在两种 loss 函数下训练的模型在 LFW 和 HW\_1\_Face 数据集的准确率，我们发现两者在 LFW 的表现相似，但是采用 center loss 辅助训练的模型在 HW\_1\_Face 上具有更高的命中率。我们分析这是因为用 softmax 训练的特征仅具备可分性，而区分度较差的原因。相比之下，center loss 更能达到类内紧凑、类间分散的效果。

在利用 VGG 数据集训练得到一个效果已经很好的模型后，我们进一步对 HW\_1\_Face 数据集中下标为 1 的图片进行数据增广，然后利用增广后的数据集混合部分 VGG 数据集在 softmax loss + center loss 训练得到的模型上 finetune，模型 top 1 rate 略有提高，但是 top 5 rate 下降。

在 HW\_1\_Face 数据集中，部分图片无法利用 MTCNN 检测到人脸，影响命中率。我们利用数据增广对图片进行旋转、亮度、对比度等一系列的变换得到大量的增广图片。大部分数据增广得到的图片集包含能检测到人脸的图片，我们进而对原始图片进行替换等操作以提高命中率。在利用 MTCNN 检测到人脸后，我们通过计算人两眼的角度对图像进行仿射变换后再进行对齐切割，这也极大的提高了命中率。

## 2. 相关工作

### 2.1 人脸检测

人脸检测和对齐是很多基于人脸的应用不可缺少的一部分。然而由于视觉变化、姿势和光照对人脸的影响，使得人脸检测和对齐的这两个任务变得具有挑战性。

早期的人脸检测算法使用了模板匹配技术，即用一个人脸模板图像与被检测图像中的各个位置进行匹配，确定这个位置处是否有人脸。Viola 和 Jones 在 AdaBoost 算法的基础上，使用 Haar-like 小波特征和积分图方法进行人脸检测，他们设计了针对人脸检测更有效的特征，并对 AdaBoost 训练出的强分类器进行级联。然而，研究显示该算法在应对大量视觉影响的人脸图片检测时性能极具下降。Felzenszwalb 在 2008 年提出一种基于组件的检测算法 DMP，DPM 的方法采用的是 Fhog 进行特征的提取，将人脸分割成好几个部件进行特征的提取，但是由于该模型过于复杂，判断时计算复杂，很难满足实时性的要求。

Cascade CNN 可以认为是传统技术和深度网络相结合的一个代表，和 VJ 人脸检测器一样，其包含了多个分类器，这些分类器采用级联结构进行组织，然而不同的地方在于，Cascade CNN 采用卷积网络作为每一级的分类器。由于大多数以前的脸部检测和脸部对齐方法忽略了这两者之间的固有相关性，Zhang 等人提出新的级联架构 (MTCNN) 来整合多任务卷积神经网络学习的问题，它将人脸区域检测和人脸关键点检测放在一起，同 Cascade CNN 一样也是基于 cascade 的框架，但是整体思路更加巧妙合理，利用检测和校准之间固有的相关性在深度级联的多任务框架下来提升它们的性能。

## 2.2 人脸识别

在网络架构方面，从最开始吴翔提出的浅层网络 lightCNN，以及到现在的 ResNet, Inception-ResNet, ResNeXt 和 SeNET 等，由于这些网络模型都是针对图像识别而设计的网络，因此不同模型之间的性能相差不大，目前常用的网络模型为 Inception-ResNet。

在目标函数方面，广泛使用的损失函数有 softmax loss, contrastive loss, center loss, normface, triple loss, large margin loss, coco loss, InsightFace 等等。由于 softmax loss 只会使得类间特征分离，并不会使属于同一类的特征积聚，这样的特征对于人脸识别来说不够有效，因为对人脸识别来说，特征不仅要求可分，还应该是可判别的。Wen 等人为了最小化类内差距，在 softmax loss 上添加一项 center loss（即每个样本和它对应的类别的特征向量的中心的距离）。为了最大化类间差异和最小化类内差异，文章使用 softmax loss 和 center loss 联合训练。在 FaceNet 中，作者提出了基于度量学习的损失函数 triplet loss，选择最佳的三元组，通过训练，使得类间的距离大于类内的距离。

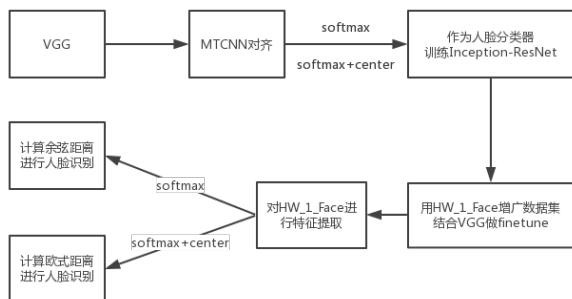
损失函数从欧式距离向余弦角度距离发展，Sphereface 提出了一种角度的 softmax 学习到角度判别性的特征，归一化 softmax 前的全连接层权重，增加角度裕量，使得类间距离的最小值大于类内距离的最大值。同时保留较大的裕量。L-Softmax 利用余弦距离，通过调节类间的间隔，使得类内距离减少，类间距离增大，有效的防止了过拟合。

## 3. 方法

### 3.1 人脸识别框架

本次实验的总体流程图如图 1，在之后的部分我们会分别介绍每一步的方法：

图 1. 总体流程图



## 3.2 数据集与数据预处理

### 3.2.1 数据集

本次实验所使用的大型人脸数据集主要包括两部分：训练数据集 VGGFace<sup>1</sup> 和验证数据集 LFW<sup>2</sup>。VGGFace 2 包含 9131 个对象的 331 万幅图像，每个对象平均有 362.6 幅图像。图像

来自于谷歌图片搜索，并在姿势、年龄、光照、种族和职业等方面有很大差异。该数据集有以下几个特点：

- (1) 对象的个数很多，每个对象包含的图像也很多；
- (2) 涵盖的姿势、年龄和种族的范围很大；
- (3) 通过自动和手动过滤来最小化标签噪音。

LFW 是针对研究非受限情况下的人脸识别问题的人脸图片数据库，包括超过 13000 张从网上搜集来的图片，共包括 1680 人，每人在数据集中有两张或更多不同的照片。该数据集被广泛应用于评价人脸识别与认证任务相关算法的性能。

同时，我们在课程提供的 HW\_1\_Face 数据集上测试人脸识别模型的 top 1 命中率和 top 5 命中率。

### 3.2.2 数据预处理

通常来说，人脸识别问题一般包括人脸检测、人脸对齐和人脸识别与比对三个步骤。人脸检测即在图像中定位出人脸的位置、大小和姿态；人脸对齐是指定位出面部关键特征点，如眼睛、鼻尖、嘴角点、眉毛以及人脸各部件轮廓点等。大多数人脸识别任务都需要对人脸图像数据进行预处理来提高准确率。

传统的人脸对齐方法包括基于人脸形状建模和基于人脸表现建模。所谓人脸形状建模从局部特征中搜索关键点，包括可变形模板、点分布模型、图模型、级联形状回归模型等，这类方法对噪声比较敏感；基于人脸表现建模又分为两类，一是对全局纹理建模，这里全局是指整张脸，主要有 AAM 等；另一种是对局部纹理进行建模，有颜色模型、投影模型等。除此之外，人脸对齐方法还包括基于深度网络的相关算法，比如卷积神经网络 (CNN)、深度自编码器 (DAE) 和受限玻尔兹曼机 (RBM) 等。

本次实验所使用的人脸图片为单人粗粒度的人脸切割，通过调研我们发现，Zhang [1] 等人于 2016 年发表的论文中提出的基于多任务级联卷积神经网络的人脸检测和对齐方法 (multi-task CNN, 简称 MTCNN) 比较适合本实验的数据预处理过程，下面就对 MTCNN 进行简单介绍。

### 3.2.3 MTCNN

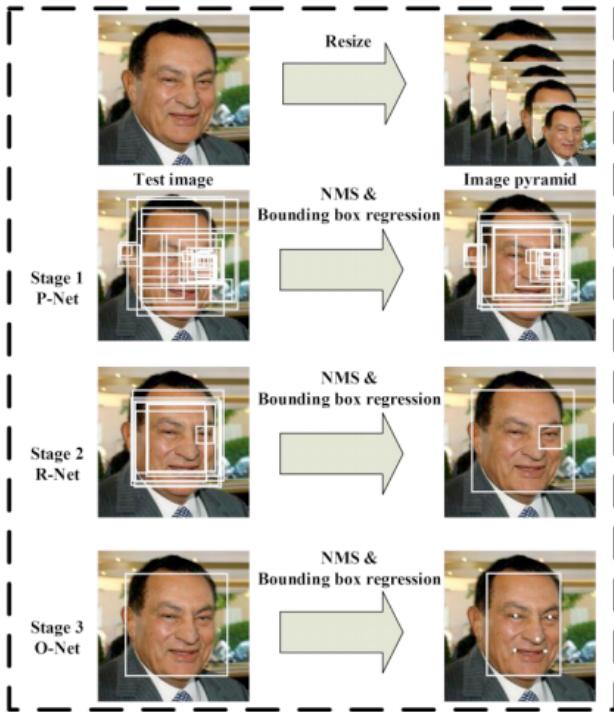
#### 3.2.3.1 整体流程框架

MTCNN 是一个基于 CNN 的级联型框架，用于联和人脸检测和对齐，整体框架如图 2 所示：

<sup>1</sup> [http://www.robots.ox.ac.uk/~vgg/data/vgg\\_face2/](http://www.robots.ox.ac.uk/~vgg/data/vgg_face2/)

<sup>2</sup> <http://vis-www.cs.umass.edu/lfw/>

图 2. MTCNN 整体框架



给定一张图片，首先将该图片重新调整到不同尺度大小，得到一个图像金字塔，该图像金字塔就是后面三阶段级联结构的输入：

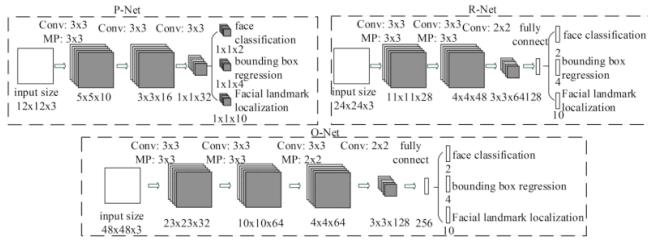
阶段 1：利用一个全卷积网络，称为 Proposal Network (P-Net)，来获得候选窗口和它们的边界框回归 (bounding box regression)，然后利用边界框回归来调整候选框。之后，利用非极大值抑制 (non-maximum suppression , NMS) 来合并那些高度重合的候选框。

阶段 2：第二阶段是一个带有全连接层的卷积神经网络，称为 Refine Network (R-Net)。该阶段的作用是接收第 1 阶段产生的所有候选框作为输入，利用边界框回归和非极大值抑制进一步排除掉大量错误的候选框。

阶段 3：Output Network (O-Net) 和 R-Net 相似，但这个阶段会输出 5 个人脸关键点位置。

本文三个阶段的网络结构如图 3 所示：

图 3. MTCNN 网络结构



### 3.2.3.2 训练

该文章利用三个任务来训练 CNN 检测器：人脸/非人脸分类，边界框回归 (bounding box regression) 和人脸关键点定位 (facial landmark localization) 。

1) 人脸分类器。人脸/非人脸分类是一个二分类问题，对于每个样本  $x_i$ ，使用交叉熵损失函数 (Cross-Entropy Loss)，其中， $y_i$  表示网络预测该样本是人脸的概率， $p_i$  是真实值标记：

$$L_i^{det} = -(y_i^{det} \log(p_i) + (1 - y_i^{det})(1 - \log(p_i)))$$

2) 边界框回归 (Bounding box regression)。对于每个候选窗口，我们预测该候选窗口与其最近的 ground truth 的偏移。这是一个回归问题，对每个样本  $x_i$ ，使用欧几里得距离 (Euclidean loss)：

$$L_i^{box} = \|\hat{y}_i^{box} - y_i^{box}\|_2^2$$

3) 人脸关键点定位 (Facial landmark localization)。与 Bounding box regression 相似，损失函数如下所示：

$$L_i^{landmark} = \|\hat{y}_i^{landmark} - y_i^{landmark}\|_2^2$$

此外，该篇文章提出了 multi-source training。由于在每个卷积网络中使用了不同的任务，因此，在学习阶段有不同种类的训练图像，比如人脸、非人脸，部分对齐的人脸。在这种情况下，一些损失函数用不到。整体的学习目标如下，其中， $N$  表示训练样本的数量， $\alpha_j$  表示任务的重要性：

$$\min \sum_{i=1}^N \sum_{j \in \{det, box, landmark\}} \alpha_j \beta_i^j L_i^j$$

在训练时，该文章在人脸分类中采用在线的 hard sample mining 来自适应训练。在每个 mini-batch 中，从所有样本的前向传播中将计算得到的 loss 排序，然后只取其中 loss 最高的前 70% 作为 hard samples。然后在反向传播 (BP) 中只计算这些 hard samples，忽略那些简单的样本。

### 3.2.4 数据增广

在实验中，我们遇到了两类问题：

1. 每人仅一张图片可以进行 fintune

2. 由于光照遮挡等原因，MTCNN 无法检测给定图片中的人脸。如下面的两张图，由于光照和拍摄角度问题，MTCNN 无法在图片上检测并框出人脸的位置。



我们对上图不能识别的两张图片进行处理，得到可处理图片，如下图。处理后的图片可被 MTCNN 检测出人脸位置。



因此，在本实验中，我们对数据集进行了增广，使用方法如下：

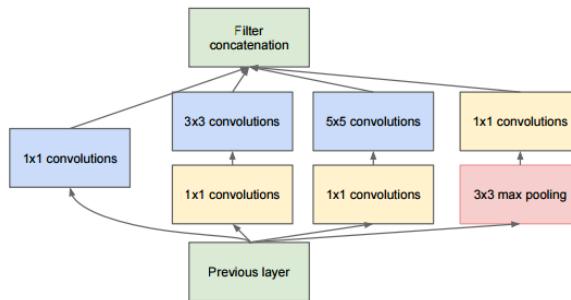
- (1) 对像素位置的变化：水平翻转，旋转变换等
- (2) 对颜色的数据增强：图像亮度，对比度，锐化等

### 3.3 基于 Inception-Resnet-v1 的特征提取

#### 3.3.1 Inception-Resnet-v1 结构

在实验中，我们使用 Inception-Resnet-v1 模型来对人脸图像进行特征提取。之前针对图像分类任务，Szegedy 提出了包含 Inception 结构的 CNN 网络结构 GoogleLeNet。它的网络层数比传统的 CNN 更深，表达能力更强，不仅分类效果有较大提升，同时保持了较低的计算开销。其网络结构示意图如图 4 所示：

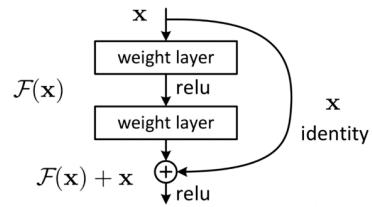
图 4. Inception 结构



该模型是由众多的  $1 \times 1$ ,  $3 \times 3$  和  $5 \times 5$  大小的卷积过滤器并行组合而成，其创新性的地方在于首先会通过一个  $1 \times 1$  的卷积过滤器使特征数量减少，这一般被称为瓶颈层，借此可以减少运算的数量，进而保持较低的计算开销和推理时间，同时不会损失性能。

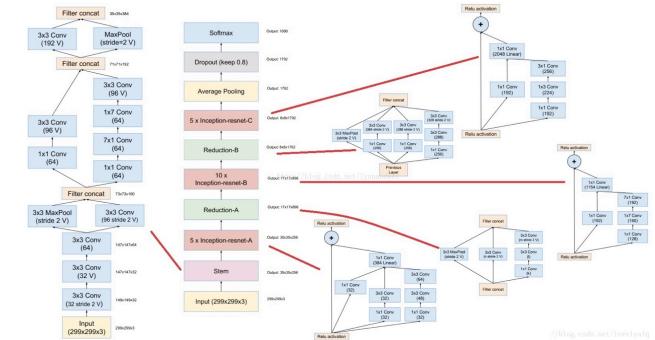
进一步，近年来还提出了 residual connections 的网络连接结构，它最根本的动机是解决模型深度加深带来的“退化”问题，如图 5 所示，作者提出了 residual block 的结构。residual block 通过 shortcut 将这个 block 的输入和输出进行逐元素相加，这个简单的加法并不会给网络增加额外的参数和计算量，同时却可以大大增加模型的训练速度、提高训练效果。

图 5. Residual Block



本实验用到的 Inception-Resnet-v1 模型就是 Inception Architecture 和 residual connection 的结合。具体地，该模型使用了最新版本的 Inception architecture，并且将其中的 filter concatenation 结构替换为了 residual connections 结构。根据作者介绍，使用 residual connections 结构可以显著加速 Inception 网络的训练过程，同时其性能要优于网络结构复杂度相近的纯 Inception 网络。Inception-ResNet-v1 模型的网络结构如图 6 所示：

图 6. Inception-Resnet-v1



该模型是将 Inception architecture 和 residual connections 结构进行了合并。

现在效果最好的人脸识别模型是 facenet，在 facenet 中作者提出了多种模型，我们将与 facenet 中提出的各网络结构对比来说明我们选择 Inception-Resnet-v1 的原因。

在 facenet 的论文中，主要使用了两种 CNN 模型。一种是 NN1，它以 Zeiler&Fergus 提出的网络结构为原型，并在其基础上添加了一些  $1 \times 1 \times d$  的卷积层。另一种即为上文介绍的基于 Inception architecture 的模型，包括 NN2, NN3, NN4, NNS1, NNS2，分别对应不同的网络规模以适应不同的图像输入大小或者不同的应用场景。这两类模型共同存在的缺点是，在 VGG 数据集上较难训练，不容易收敛，即使在没有正则项的条件下依然维持相对较大的训练误差。所以我们发现 facenet 中的这些模型并不适合用来训练 VGG 等人脸图像数据集。而 Inception-Resnet-v1 这个网络结构可以在训练集上较好地收敛，而且可以在 LFW 测试集上有很好的性能表现。综上，我们决定在本实验中使用 Inception-Resnet-v1 网络结构。

#### 3.3.2 Loss Function

在确定网络结构后，我们需要选择一个合理的 loss function 来辅助训练。在本实验中，我们选择两种 loss 函数分别训练网络，分别是：softmax loss 和 softmax loss + center loss。

我们两种损失函数中都包含 softmax loss，这意味着我们将把 Inception-Resnet-v1 作为分类器来训练。与 facenet 不同，我们并没有采用 triplet loss 来训练网络，因为 triplet loss 会使得

训练集规模骤增，并且利用 triplet loss 训练收敛缓慢，稳定性降低，同时论文中提到利用 hard positive 和 hard negative 来辅助训练，这带来采样困难的问题。所以比较之下，利用 softmax loss 进行训练更加的简单和高效。Softmax loss 的定义如下：

$$\mathcal{L}_S = - \sum_{i=1}^m \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^n e^{W_j^T x_i + b_j}}$$

其中， $m$  为 mini-batch 的大小， $x_i$  代表第  $i$  个样本通过 Inception-ResNet-v1 提取的 deep feature， $W^T$  为 softmax 的权重矩阵， $b$  为 softmax 的偏置。这种 loss 的定义又叫交叉熵。可以看出，我们使用 Inception-ResNet-v1 对图像进行 deep feature 的提取，再在 deep feature 的基础上联合训练一个 softmax 分类器。

我们第一个模型就是利用 softmax 训练，但是同时我们注意到了仅 softmax loss 可能存在的问题：softmax loss 的目的仅仅是使得模型提取出的特征是可分的，虽然简单，但训练出的特征只具备可分性，而不具备好的区分度，判别能力较差，尤其在面对训练集以外的新的身份的人脸时，其泛化能力较差。

而 facenet 中使用的 triplet loss 可以使得模型具有更好的区分度，triplet loss 的目的是使得模型提取的特征在欧氏空间上有良好的性质：特征向量在欧氏空间中的距离能够直接反应人脸的相似度。更直接的说：triplet loss 的训练目的是使得同一身份的人脸特征之间的欧氏距离更近，更紧凑，使得不同身份的人脸特征之间的欧氏距离更远，更分散，好比每个身份的人脸形成一个社区，社区内紧密，社区之间松散。我们认为这样的人脸特征更具有区分度，判别能力更强，尤其在做人脸识别时，可以直接使用 kNN 方法就能完成任务而不用额外使用分类器进行分类。

与 triplet loss 相比，softmax loss 并不能使得特征具有上述性质，所以为了使的训练特征可能具有更好的可分性，而又避免使用 triplet loss，我们决定引入 center loss。

Center loss 的思想很直观，对于每个类别  $j$ ，都会维护一个中心向量  $c_j$ （与提取出的特征维度相同），代表类别  $j$  的中心点，然后极小化属于类别  $j$  的特征与  $c_j$  的距离。Center loss 的定义如下：

$$\mathcal{L}_C = \frac{1}{2} \sum_{i=1}^m \|x_i - c_{y_i}\|_2^2$$

其中， $m$  为 mini-batch 的大小， $x_i$  代表第  $i$  个样本的 deep feature， $c_{y_i}$  代表类别  $y_i$  的中心点。

理论上，类别  $j$  的中心点  $c_j$  等于训练集中所有属于类别  $j$  的人脸特征的均值，显然这样开销计算过大，所以在实际训练过程中，我们以 mini-batch 中各个类别对应的特征的均值近似代表各个类别的中心点。

所以我们第二种损失函数表示为 softmax loss 和 center loss 两部分之和：

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C$$

其中， $\lambda$  代表 center loss 的权重因子。

相比于 triplet loss，使用 center loss 训练时的 batch size 更小，计算方式更简单，更易收敛，且同样能达到类内紧凑，类间分散的效果。我们的第二个模型就是利用 center loss 训练的。

虽然我们以分类器的方式来训练 Inception-ResNet-v1，但是我们的最终目的是能够从人脸图像中提取出有良好性质的特征，这部分特征会在后续的人脸识别实验中用到。

### 3.4 基于 HW\_1\_Face 数据集 fine-tune

我们希望利用 HW\_1\_Face 数据集进一步提高我们模型的人脸识别命中率，所以我们准备利用 HW\_1\_Face 中的 \*\_1.jpg 对模型进行 finetune，这样可以使得模型的特征能够更好的表示 HW\_1\_Face 中的人脸。但是 HW\_1\_Face 数据集中，每个人只有两张图片，这意味着我们仅有 2000 张图片可以进行 finetune，我们做了一系列的实验发现，除了可以偶然得到有提高的模型外，模型的命中率、LFW 的准确率和验证率会随着训练不断变低。所以我们通过对 \*\_1.jpg 进行旋转、变化亮度、对比度等图像处理进行数据增广，然后在利用增广后的数据或者继续结合一部分 VGG 数据在模型上进行 finetune 继续比较效果。

### 3.5 在 HW\_1\_Face 和 LFW 数据集上做验证

#### 3.5.1 在 HW\_1\_Face 上测试 1vn 人脸识别测试

在 Inception-ResNet-v1 模型训练过程中，我们会不断在 HW\_1\_Face 数据集上做 1vn 人脸识别测试。HW\_1\_Face 数据集包含 2000 个人，每个人各自有两张人脸图像，每张图像的 title 形式为：id\_1.jpg 和 id\_2.jpg，其中 id 是人的编号。我们用 \*\_2.jpg 用作测试，\*\_1.jpg 做索引的数据库。

对于每张测试的图片 \*\_2.jpg，我们计算它和所有下标为 1 的图片的距离，并通过统计对应的 \*\_1.jpg 是否是距离最近的图片或者在最近的五张图片中来计算 top 1 命中率和 top 5 命中率。

在测试时，我们发现有部分图片无法使用 MTCNN 检测出人脸，这影响了人脸识别的准确率。因为当图片无法检测出人脸时，我们只能将整张图片 resize 之后输入网络。如果是 \*\_2.jpg，直接 resize 的图片与对应的 \*\_1.jpg 可能距离非常远，使得识别失败。如果是 \*\_1.jpg，直接 resize 的图片会影响所有测试图片计算距离，对人脸识别的效果影响非常大。

我们利用前面介绍的数据增广来解决问题。对无法识别出人脸的图片对应的增广数据集用 MTCNN 检测人脸，绝大部分增广数据集都存在一张或者多张图片能够检测出人脸。针对 \*\_2.jpg，我们计算其对应的所有能检测出人脸的增广图片与索引数据集所有图片的距离并统计总的 top 1 和 top 5 图片做为 \*\_2.jpg 的输出。针对 \*\_1.jpg，我们在其对应的所有能检测出人脸的增广图片中随机选取一张图片作为替代图片。

在用 MTCNN 对输入人脸图片进行检测和对齐时，MTCNN 不仅会返回 bounding box，还会返回人脸的五官位置。所以在利用 bounding box 对图片 crop 后，我们还利用人两眼的坐标计算角度并对图片进行仿射变换，再对仿射变换后的图片 resize。这也可能提升人脸识别的效果。

### 3.5.2 基于阈值在 LFW 上做验证

我们同时选择在 LFW 数据集上进行验证特征的普遍性。实验中用到的测量指标定义如下：

$$\text{accuracy} = \frac{|tp| + |tn|}{|tp| + |tn| + |fp| + |fn|}$$

$$\text{VAL} = \frac{|tp|}{|tp| + |fn|}$$

$$\text{FAR} = \frac{|fp|}{|tn| + |fp|}$$

其中， $|tp|$ 代表 true positive 样本的个数， $|tn|$ 代表 true negative 样本的个数， $|fp|$ 代表 false positive 样本的个数， $|fn|$ 代表 false negative 样本的个数。所以，accuracy 代表所有预测正确的样本数占所有样本的比例，VAL 全称 validation rate，代表正样本的召回率，FAR 全称 false accept rate，代表 false positive 样本占所有负样本的比例。

不同于 HW\_1\_Face，我们在 LFW 数据集上完成人脸验证任务，即给定一对人脸图像，判断它们是否为同一个人，具体步骤如下：

1. 在 LFW 上生成待验证的人脸图像对以及对应的标签
2. 将所有人脸图像输入 Inception-ResNet-v1 网络中，生成各自的特征向量，然后计算每对人脸图像对应的特征向量的距离
3. 使用 10 折交叉验证对模型进行评估，包括模型在测试集上的 accuracy VAL 曲线

具体地，第三步中进行 10 折交叉验证时，我们首先将数据集随机分为 10 份，在每折的验证中，将 9 份作为训练集，剩余的 1 份作为测试集，即每份数据集分别做一次测试集，共进行十次训练/测试的验证。同时，我们以 0.001 的间隔从 0 到 4 之间选择 4000 个值作为候选阈值。

对于每折验证，我们在训练集上进行训练以从候选阈值集中寻找一个最佳阈值，然后使用此阈值在测试集上进行测试，统计在该阈值下测试集的 accuracy，然后进一步计算 10 折验证的 accuracy 的平均值和标准差。

同时，对于每折验证，我们在测试集上计算候选阈值集合中每个阈值的 VAL 和 FAR，然后计算 10 折验证中每个阈值对应的 VAL 和 FAR 的均值。

综上，在 LFW 数据集上验证时，我们使用 Inception-ResNet-v1 对图像进行特征提取，然后基于两张图像的特征向量的距离和阈值来完成人脸验证，借助 10 折交叉验证来评估模型的各项性能。

## 4. 实验和评估

### 4.1 实验环境

我们程序的运行环境如下：

操作系统：Ubuntu 16.04

深度学习平台：tensorflow 1.8.0

GPU：GeForce GTX 1080 Ti/PCIe/SSE2

CPU：Intel® Core™ i7-7700K CPU @ 4.20GHz × 8

在分别用 softmax loss 和 softmax loss + center loss 训练时，我们均设置 batch 为 90，epoch\_size 为 1000，但是两者一些参数具有较大差异。表 1 列出两者不同的参数：

表 1. 不同损失函数下训练参数的差异

	Softmax	Softmax + Center
Embedding size	512	128
Image standardization	Fixed	Per image
Base learning rate	0.05	0.1
Optimizer	Adam	RMSProp
Dropout keep prob	0.4	0.8

我们将利用 softmax loss + center loss 训练时的 center\_loss\_factor 设置为 0.01，并没将其设置得非常大，毕竟我们将网络作为分类器，衡量分类性能主要还是以 softmax loss 为主，center loss 只是辅助，如果 center\_loss\_factor 过大，center loss 会淹没 softmax loss 的作用，使得网络训练很差，同时 center\_loss\_factor 较大会使得中心向量迭代学习的步数增加，导致网络容易发散，不收敛。

在利用 softmax loss 训练时，学习率本采用如表 2 所示策略进行衰减。经过 3 天加 18 个小时的训练后，人脸识别的精度达到稳定。

表 2. Softmax Loss 训练的学习率

Epoch number	学习率
0	0.05
100	0.005
140	0.0001
220	0.00005
276	-1

而利用 softmax loss + center loss 进行训练时，学习率采用如表 3 所示策略进行衰减。在经过 2 天加 12 个小时的训练后，人脸识别的精度达到稳定。

表 3. Softmax Loss + Center Loss 训练的学习率

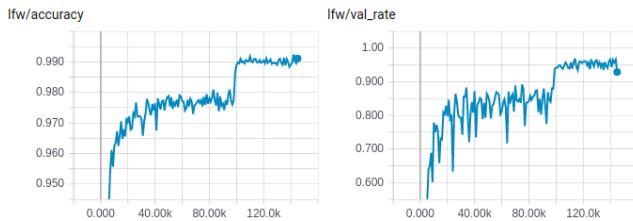
Epoch number	学习率
0	0.1
80	0.01
150	0.001
300	0.0001
1000	-1

### 4.2 训练结果

#### 4.2.1 Softmax loss

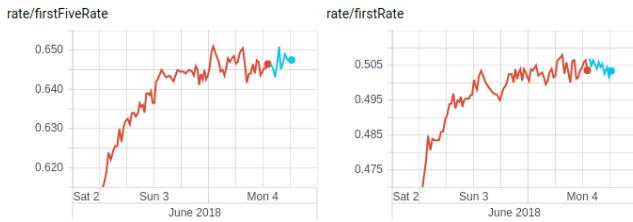
在模型训练的初始阶段，我们仅监测模型在 LFW 数据集上的效果，如图 7，我们可以看到，最初学习率为 0.05 时，LFW 的准确率在 10 个 epoch 后就快速收敛到 90%以上，而且抖动较小，但是验证率始终只能达到 85%左右。直到学习率调整为 0.005，准确率陡升到 98%以上，验证率陡升到 95%以上，此时 top 1 命中率达到 47%，top 5 命中率达到 60%。

图 7. LFW 初期变化



接下来，我们进一步将学习率调小为 0.0001，准确率进一步提高到 99%以上，验证率提升到 97%以上，如图 8，top 1 rate 和 top 2 rate 达到 50%和 65%。之后，我们试图继续通过调小学习率到 0.00005 来提高命中率，但是实验结果表明进一步降低学习率不能对效果有提升。所以最后我们利用 softmax 训练的网络达到的最好效果为：top 1 50.8%，top 5 65.1%。

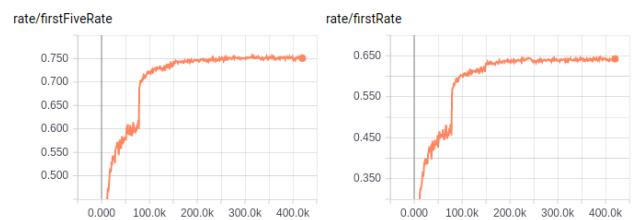
图 8. 学习率 0.0001 和 0.00005 时人脸识别命中率变化



#### 4.2.2 Softmax Loss + Center Loss

如图 9，初始学习率为 0.1 时，LFW 的准确率在 30 个 epoch 后就快速收敛到 97%以上，抖动较小，验证率达到 85%左右，但是抖动剧烈。与之趋势相同的，模型在 HW\_1\_Face 上的 top 1 和 top 5 命中率达到了 46%和 57%。

图 9. 人脸识别命中率随训练的变化



当学习率降低到 0.01 时，LFW 的准确率和验证率稳步提升到 99%和 96%，抖动变小，在 HW\_1\_Face 上的 top 1 和 top 5 命中率迅速提升到了 60%和 72%。

最终在我们将学习率调整到 0.001 后，在 HW\_1\_Face 上的 top 1 和 top 5 命中率最终达到了 64.7%，75.6%。我们会在这个基础上继续 finetune 以进一步提升命中率。

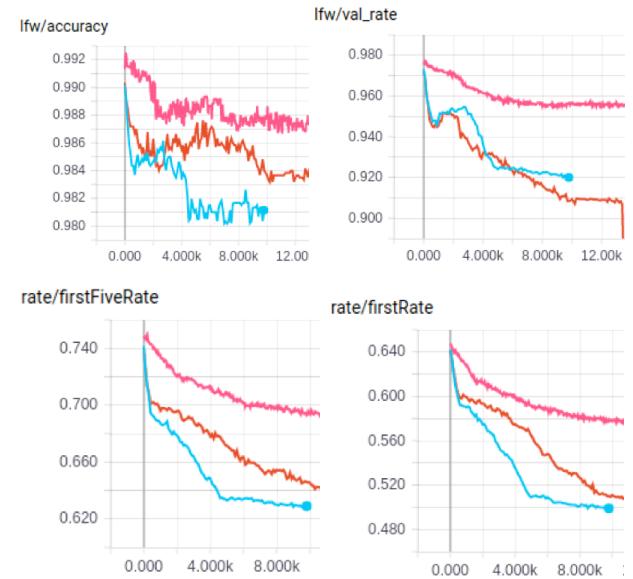
### 4.3 fine-tune 的效果

我们分别做了 3 组 finetune 实验，分别是用 2000 张图片、2000 张图片增广后的数据集以及增广数据集混合 VGG 数据集来 finetune。

虽然我们希望通过 VGG 数据和增广数据集带来命中率效果上的提升，但是我们发现在三种数据集上 finetune 时，如图 10 所示，命中率始终在不断下降，直到收敛到某个值。其中橙线是 VGG 混合增广数据集，蓝线代表增广数据集，红线代表

2000 张数据集。与之对应的，在 LFW 准确率变化较小的情况下，LFW 验证率显著下降。

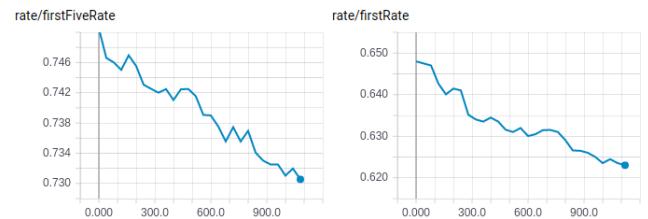
图 10. 随 finetune 下降剧烈的命中率和 LFW 验证率



出乎意料的，我们看到仅用 2000 张图片 finetune 虽然仍然带来了性能下降，但是最后达到收敛的命中率是三个数据集中最高的，我们分析可能是因为我们生成的增广数据集由于还是在同一张图片上生成的，这导致训练时出现了过拟合的情况。这和数据集本身有限相关，但是同时也和我们 finetune 的方法相关。

通过大量的实验，我们发现有极少数的情况，命中率能在第一个 epoch 后提升，如图 11，在第一个 epoch 后 top 1 和 top 5 命中率达到 64.8%和 75.05%，但是之后命中率仍然不断下降。并且 top 5 命中率没有提升效果。

图 11. 有少量效果提升的训练结果



### 4.4 数据增广以及仿射变换的效果

#### 4.4.1 仿射变换

在利用 MTCNN 对训练数据集 VGG 进行处理时，我们仅进行了人脸检测和利用 bounding box 进行粗粒度的 2D 人脸对齐切割。但是在 HW\_1\_Face 数据集上进行测试时，我们还利用 MTCNN 检测到的五官信息对切割后的图像进行仿射变换。这对人脸识别的命中率起到了提升效果。我们选取训练得到两个不同 loss 的模型上进行了对比试验，得到的结果如表 4：

表 4. 仿射变换的效果

	未仿射变换	进行仿射变换
Softmax	0.4695/0.617	0.474/0.624
Softmax + center loss1	0.613/0.714	0.6455/0.7495
Softmax + center loss2	0.647/0.7515	0.6785/0.7795

我们看到，利用五官信息进行仿射变换对模型命中率起到了提升效果，尤其是对于用 softmax loss + center loss 训练的模型，提升效果非常明显。

#### 4.4.2 数据增广

在方法 3.2 中，我们介绍了如何利用数据增广来处理 HW\_1\_Face 数据集中部分图片无法检测到人脸的问题，我们在多个模型上对其效果进行了对比实验。得到的结果如表 5：

表 5. 数据增广的效果

	未增广，已仿射	增广，已仿射
Softmax	0.508/0.651	0.513/0.6595
Softmax + center loss1	0.6135/0.7370	0.6195/0.744
Softmax + center loss2	0.6145/0.7380	0.622/0.739
Softmax + center loss3	0.6175/0.74	0.6205/0.74
Softmax + center loss4	0.6785/0.7795	0.6805/0.781

我们看到，利用五官信息进行数据增广对模型命中率起到了较好的提升效果

进一步的，我们统计数据增广对每张不能检测到人脸的图片的人脸识别效果的作用。表 6 展示了图片增广前和增广后是否和对应图片被判断为同一个人的情况，其中 G 代表判断为是。我们发现数据增广对不能检测到人脸的图片的 top 1 和 top 5 的改善是巨大的，尤其是 top 5，在增广前，仅有 2 张图片被正确归类，在增广后，14 张图片都能在前五张图片中找到正确的对应图片。

表 6. 数据增广对不能识别出人脸的情况的改善

	前 1	后 1	前 5	后 5
265_1.jpg		G		G
1107_1.jpg				
1701_2.jpg				
1344_2.jpg				
1869_2.jpg				G
257_1.jpg				
552_1.jpg		G		G
183_1.jpg	G	G	G	G
1993_2.jpg		G		G
1052_2.jpg				G
522_1.jpg				
1560_2.jpg				
806_1.jpg				G
1394_1.jpg		G	G	G

389_1.jpg				
895_2.jpg		G		G
1727_1.jpg		G		G
1431_1.jpg				
94_1.jpg				
1369_2.jpg				
178_1.jpg				
717_1.jpg				
1788_2.jpg				
1450_2.jpg				
1626_2.jpg				
986_2.jpg				
1195_1.jpg		G		G
202_1.jpg				
331_1.jpg				
1396_1.jpg				
1151_2.jpg				
665_2.jpg				G
1877_1.jpg				
996_2.jpg				
350_2.jpg				G
1724_2.jpg				G
1381_2.jpg				

以人物 1727 为例，如图 12，在增广前，1727\_2.jpg 的 top 1 和 top 5 中都没有正确的对应图片 1727\_1.jpg，我们可以看到 1727\_1.jpg 本身因为侧脸等原因无法识别出人脸，但是经过锐化，MTCNN 成功的检测出人脸并与 1727\_2.jpg 匹配。

图片 12. 1727\_1.jpg、1727\_2.jpg、锐化后人脸检测的效果



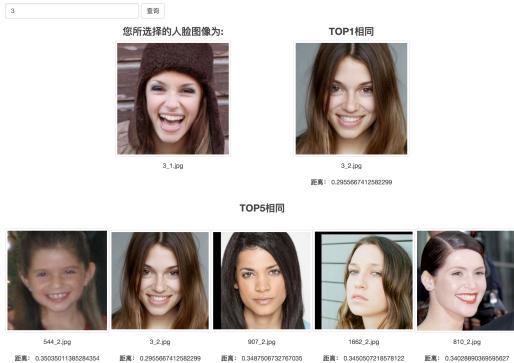
#### 4.5 可视化

我们针对本次人脸识别任务做了可视化，提供网页版的用户界面进行访问与查询，相关项目代码开源在 <https://github.com/ChaoyueWu/ChaoyueWu.github.io> 中，您可以通过访问 <http://chaoyuewu.github.io/> 来查看和使用我们的用户界面。下面对该用户界面的使用方式做简单说明。

访问 <http://chaoyuewu.github.io/>，在页面上的输入框中键入希望查询的人脸图片编号（编号范围为 0 到 1999），点击查询按钮即可进行查询。图 13 显示的是一个查询结果示例，其中第一行展示了所要查询的人脸图片，以及使用我们的模型进行查询得到的和给定人脸图片属于同一个人的人脸图和距离；第二行则展示了使用我们的模型进行判断得出的与待查

询人脸最相似的五张人脸图片（即 TOP5），以及他们各自和给定人脸的距离。

图 13. 可视化界面



## 5. 结论

在本次实验中，我们在网上公开的 VGG 数据集上，分别利用 softmax loss 和 center loss + softmax loss 训练 Inception-Resnet-v1 进行人脸识别任务，其中 center loss + softmax loss 训练得到的模型在 1vn 人脸识别任务上表现更好。我们在训练好的模型上利用 HW\_1\_Face 数据集进行 finetune，但是效果不佳。在具体的搜索算法中，我们利用仿射变换和数据增广进一步提高模型的命中率，最终模型在 1vn 人脸识别的任务中 top 1 命中率达到了 68.1%，top 5 命中率达到 78.1%。同时，我们针对本次人脸识别任务做了可视化，提供网页版的用户界面进行访问与查询。

## 6. 引用

- [1] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[2] Li H, Lin Z, Shen X, et al. A convolutional neural network cascade for face detection[C]// Computer Vision and Pattern Recognition. IEEE, 2015:5325-5334.

[3] Wu X, He R, Sun Z, et al. A Light CNN for Deep Face Representation with Noisy Labels[J]. IEEE Transactions on Information Forensics & Security, 2015:1-1.

[4] Wen Y, Zhang K, Li Z, et al. A Discriminative Feature Learning Approach for Deep Face Recognition[M]// Computer Vision – ECCV 2016. Springer International Publishing, 2016:499-515.

[5] Liu W, Wen Y, Yu Z, et al. SphereFace: Deep Hypersphere Embedding for Face Recognition[J]. 2017:6738-6746.

[6] Liu W, Wen Y, Yu Z, et al. Large-Margin Softmax Loss for Convolutional Neural Networks[C]//Proceedings of The 33rd International Conference on Machine Learning. 2016: 507-516.

[7] Deng J, Guo J, Zafeiriou S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition[J]. 2018.

## 7. 总结和反思

在本次作业之前，我们组的三个成员从未接触过深度学习，这次作业是我们第一次接触深度学习任务和代码。经验的缺失使得我们花了大量的时间在入门上。在模型的 fine tune 上，第一次提交前我们没有得到一个好的 fine tune 结果，我们认为是我们 fine tune 的方式不对，但是我们在第二次尝试时，修改了 fine tune 的方式，继续进行了大量的实验，但是仍然无法得到一个提升明显的结果。所以我们最终的结果还是主要依靠训练加上图片处理得到的。我们小组的成员在组里的项目上仍会负责 1vn 人脸识别，所以如何做好 1vn 人脸识别是我们会一直探寻的问题。