Deep Learning Face Representation from Predicting 10,000 Classes(1)(1)(23-56-27)

摘要：

论文主要目的是通过深度学习去学习到一个高水平的特征表达集(**DeepID**)用于人脸验证。

DeepID 特征集是从深度卷积网络(**ConvNets**)的最后一个隐藏层神经元提取到的。这种特征是从人脸的不同区域提取的，可以形成一个互补的完备的人脸特征表达。

结合简单的人脸对齐，在 LFW 上实现了97.45%的识别率。

第一节：

当前有着最优表现性能的人脸验证算法是采用完备的低水平特征，紧接着是浅模型来表征人脸；

近来，深模型(如ConvNets) 【24】 也能有效提取高水平可见特征 【11,20,14】，并被用于人脸验证 【18,5,31,32,36】——其中：【18】无监督地学习一个泛化的深模型；【5】学习深度非线性矩阵集；【31】通过二值人脸验证目标来监督性地学习深模型。

本文是采用深模型来学习高水平的人脸特征集，简言之，就是把一个训练样本分进10000个身份中，高维空间的操作虽然更有难度，但学习到的特征表征有更好的泛化性。

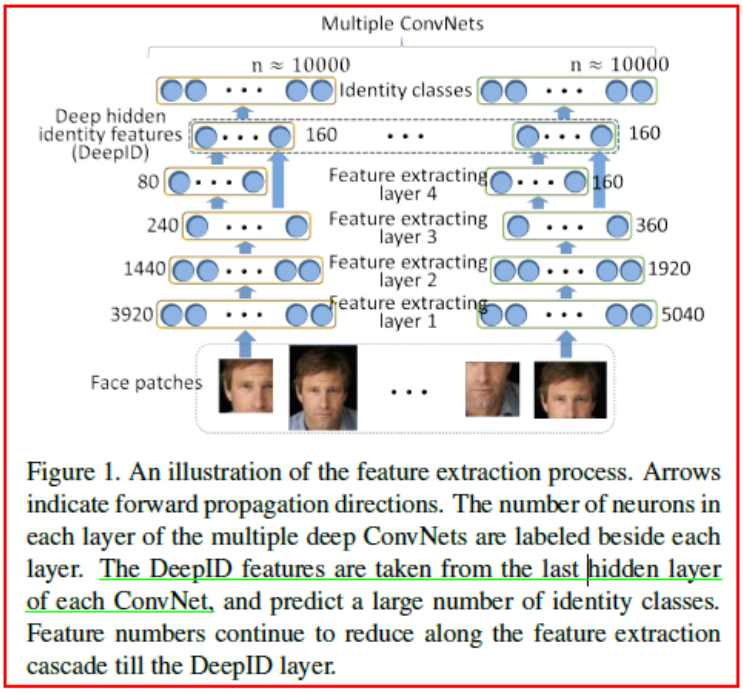

Figure 1. An illustration of the feature extraction process. Arrows indicate forward propagation directions. The number of neurons in each layer of the multiple deep ConvNets are labeled beside each layer. The DeepID features are taken from the last hidden layer of each ConvNet, and predict a large number of identity classes. Feature numbers continue to reduce along the feature extraction cascade till the DeepID layer.

每一个卷积网络将一个人脸块作为输入，并在底层提取局部的低水平特征；
在级联的特征提取过程中，特征维度不断减少，而顶层的整体的高水平特征不断形成；
如此，在级联的最后一级获得高度紧凑的160维 DeepID, 这个特征包含丰富的识别信息，并且能直接预测更大的类；
紧接着将不同人脸区域提取到的 DeepID 特征连接起来，组成完备的特征表达。

同时分类所有的身份而不是使用二分法训练分类器【21,2,3】是基于以下两点考虑：
1、将一个样本归类至多个类别比二分类更困难，这种极具挑战性的任务可以充分利用神经网络的超学习容量来为人脸识别提取高效的特征
2、这向卷及网络中暗中加入了强规则，这种规则能较好地分类多个身份类别。

优势：这样学习到的高水平特征拥有好的泛化性，且在训练人脸集的小子集中也不会存在过拟合，同时这种特征可以和任意的先进人脸分类器(如联合贝叶斯【8】)进行集成，来进行人脸验证。
关键：强制使得 DeepID(集合的大小) 远远小于要预测的类别。

第三节：

3.1 深度卷积网络

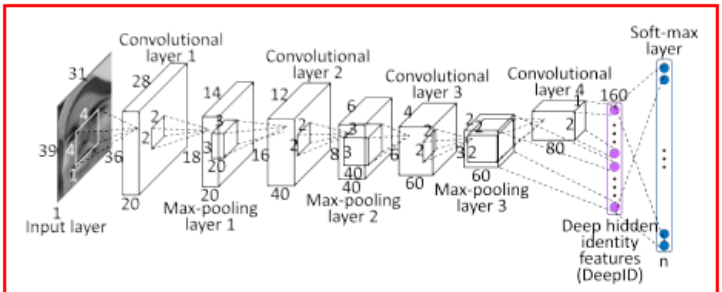

Figure 2. ConvNet structure. The length, width, and height of each cuboid denotes the map number and the dimension of each map for all input, convolutional, and max-pooling layers. The inside small cuboids and squares denote the 3D convolution kernel sizes and the 2D pooling region sizes of convolutional and max-pooling layers, respectively. Neuron numbers of the last two fully-connected layers are marked beside each layer.

该深度卷积网络通过四个卷积层(紧随polling层)来分级提取特征，紧接其后的是全连接的 DeepID 层和表明身份类别的 softmax 输出层。

输入：1、矩形图块—39*31*k ； 2、正方形图块—31*31*k ；（k=3是彩色图块，k=1是灰度图块）

左图的深度卷积网络结构是将 39*31*1 作为输入，预测 n=10000个类别，DeepID层的维度设置为160，输出层的维度 n 由它要预测的类别总数决定。

卷积运算表达式：$$y^{j(r)} = \max\left(0,\ b^{j(r)} + \sum_i k^{ij(r)} * x^{i(r)}\right), \quad (1)$$ 卷积运算一个重要特点就是：可以使原信号特征增强，并且降低噪音。

注：x^i 是上一层的输入神经元，y^j 是本层的输出神经元，k^{ij} 是相应权值，b^j 是本层神经元的偏置，r 表示权值共享的局部区域块；
第三个卷积层中，每个2*2区域共享部分权值，第四个卷积层中，权值不共享；
此处采用 ReLU 非线性(y = max(0, x))作为激活函数，比 sigmoid 函数具有更好的适应性。

下采样层表达式：$$y^i_{j,k} = \max_{0 \le m,n < s} \{x^i_{j \cdot s+m,\, k \cdot s+n}\}, \quad (2)$$ 利用图像局部相关性原理，对图像进行子抽样，可以减少数据处理量同时保留有用信息。

注：y^j 神经元是从上一层非重叠区域的神经元采样得到的。下采样怎么训练参数？

DeepID隐藏层：$$y_j = \max\left(0,\ \sum_i x^1_i \cdot w^1_{i,j} + \sum_i x^2_i \cdot w^2_{i,j} + b_j\right), \quad (3)$$ 全连接于第三层和第四层(下采样之后)，由于第四卷积层的特征比第三层更具有全局性，因此学习到的特征是多尺度的。

输出：$$y_i = \frac{\exp(y'_i)}{\sum_{j=1}^{n} \exp(y'_j)}, \quad (4)$$ 卷积网络的输出是在n个类上的概率分布。表征该样本被预测到不同类中的可能性大小。

注：$y'_j = \sum_{i=1}^{160} x_i \cdot w_{i,j} + b_j$

3.2 特征提取

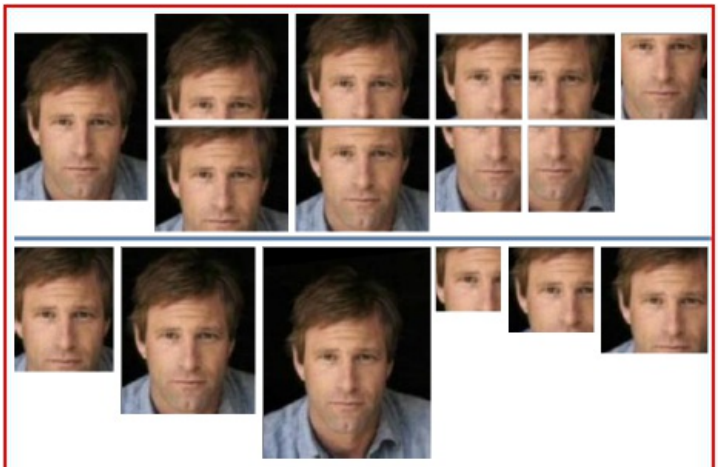

Figure 3. Top: ten face regions of medium scales. The five regions in the top left are global regions taken from the weakly aligned faces, the other five in the top right are local regions centered around the five facial landmarks (two eye centers, nose tip, and two mouse corners). Bottom: three scales of two particular patches.

预处理部分：

面部关键点检测是参考文献【30】中五点检测；

采用相似性变换对人脸进行全局校正；

特征是从 60 个人脸块中提取到的（60=10*3*2，10个区域，3个尺度，2个颜色空间）。

我们训练了60个卷积神经网络，每个网络都提取2个160维的DeepID向量（从某一人脸图块及它水平翻转后的图块）得到的 **DeepID 特征集**的维度是 19200（160*2*60）。

3.3 人脸验证

我们采取的是基于 DeepID 的**联合贝叶斯方法**【8】进行人脸验证，该模型在文献【9,6】的人脸验证中均有良好表现。

采用两个独立的高斯变量的和来表征提取到的面部特征集 x（去除均值）：

$$x = \mu + \epsilon, \quad (5)$$

注：其中$\mu \sim N(0, S_\mu)$代表人脸身份信息（类间），$\epsilon \sim N(0, S_\epsilon)$代表同一人物的变化（类内）；

在给定类间和类内标识后，联合贝叶斯模型算出两个人脸之间的联合概率：$P(x_1, x_2 \mid H_I)$ 和 $P(x_1, x_2 \mid H_E)$.

同时，我们也专门训练了用于人脸验证的**神经网络**（利用提取到的特征进行验证这一环节），并且将它同联合贝叶斯方法相比。以此观测是否其他模型也能从提取到的特征（DeepID）中学习,也观测在验证性能上，特征和一个好的人脸验证模型的贡献各有多少。

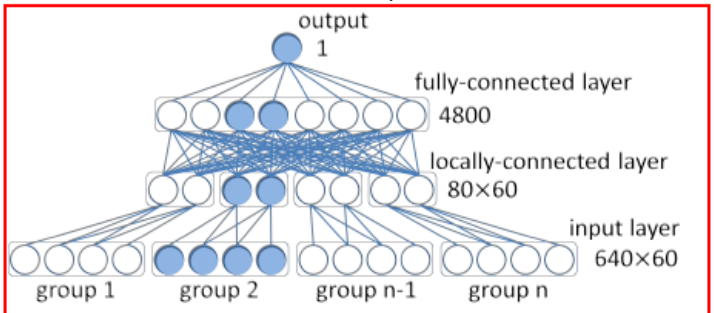

Figure 4. The structure of the neural network used for face verification. The layer type and dimension are labeled beside each layer. The solid neurons form a subnetwork.

该网络由一个输入层（DeepID集）、一个局部连接层、一个完全连接层和一个表示人脸相似度的单个神经元组成。

输入特征被分为60个组，每组包含640维特征——用特定的卷积网络在特定的图块对上提取的特征（320*2=640），同一个组内的特征是高度相关的。

各层功能分析：

在局部连接层的神经元只连接到某一个组来学习这个局部的关联并且同时降低了特征的维度。

第二个完全连接层，完全连接到第一个隐藏层，以此学习整体的关联。

单输出的神经元完全连接到第二个隐藏层，隐藏的神经元都是ReLU函数，输出的那个神经元是sigmoid函数。

问题：

所有隐藏神经元都是通过 Dropout 学习【16】的方式得到，输入神经元不能丢掉，因为它学习到的特征很简洁并且是分散表达人脸身份的（用少量的神经元表达大量的人脸身份类），这些神经元需要相互合作才能更好的表达身份信息。然而由于渐层扩散的原因，不丢弃（神经元）地学习高维的特征是很困难的。

策略：

1、先训练了60个子网络，每个子网络只输入一组特征（对应60个区块中的某一个提取到的特征）；

2、然后使用第一层子网的权重来初始化原始网络的权重，再使用第一层权重的截尾（剪除）来调整原始网络的第二层和第三层。

第四节：

实验部分

样本集合：前期样本集准备：由于LFW库只有85人拥有超过15张图片，4096个人只有一张图片，这远远不足我们的训练输入。因此我们在CelebFaces【31】库上训练我们的模型，在LFW上测试。

CelebFaces有5436个社会名流，共计87628张图片，平均每人16张，它和 LFW库是互不相容的。

后期样本集扩展：将CelebFaces扩展至CelebFaces+，共10177个人（202599张图片）

特征学习：随机挑选 CelebFaces 中80%的人（4349）来学习DeepID，剩下的20%用来学习人脸验证模型（联合贝叶斯或者神经网络）。

在特征学习阶段，卷积网络有监督的同时对4349个人进行分类——依上文所示，每张图衍生出60张图块。 我们随机的选择每个训练的人（身份类）的10%的图片用于生成验证数据。

每个训练级结束后，我们观测使验证集错误率最优的模型，并将这个模型用于支持验证结果最差的那个模型。

人脸验证：在学习联合贝叶斯模型之前，我们使用PCA降维将特征维度降为150维。验证的性能在很大的维度范围内都可以保持稳定。

测试阶段，通过比较联合贝叶斯似然比和一个阈值进行人脸对的分类，而这个阈值是在训练数据中按照最优化原则得到的。

算法有效性验证

多尺度的卷积网络：

DeepID层由第三和第四卷积层（在pooling层之后）同时连接而来，因此称为多尺度的卷积网络，这样做能用更少的网络学到更具表达能力的特征。

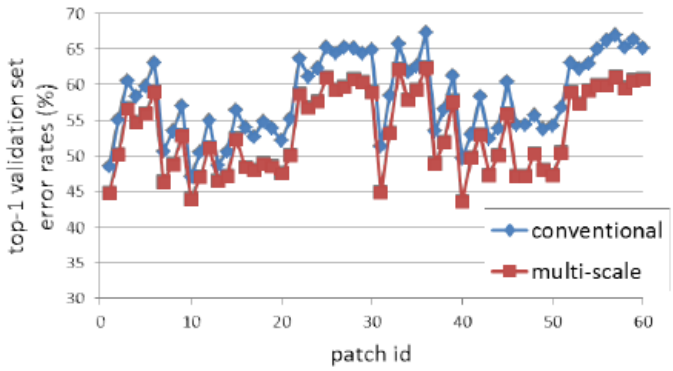

Figure 5. Top-1 validation set error rates of the 60 ConvNets trained on the 60 different patches. The blue and red markers show error rates of the conventional ConvNets (without the skipping layer) and the multi-scale ConvNets, respectively.

这里的**跳跃层**（skipping layer）指的是将第三卷积层与DeepID层相连，这样做减少了第四卷积层上的信息损失，连接函数

$$y_j = \max\left(0, \sum_i x_i^1 \cdot w_{i,j}^1 + \sum_i x_i^2 \cdot w_{i,j}^2 + b_j\right), \quad (3)$$

学习有效特征：

同时对所有的身份进行分类对于学习高度判别力（能认识更多的类）的压缩（特征维数低）隐藏特征是非常关键的。因此试验中将类数分别从136增加到了4349，以下图一表明：类别数成倍增加时，验证的准确率线性提升，同时验证的错误率也有所下降。——说明学习到了高度判别力的特征

图二表明相同的人脸有更多相同的激活神经元（相同位置的positive features）,因此学习到的特征能提取身份信息。

同时还将4349维的分类器输出（soft-max层）作为特征来做人脸验证，只有66%的准确率（贝叶斯）。而用神经网络作验证，结果失败，因为类别数太多，而每个类的样本数很少，因此分类器输出多样性。

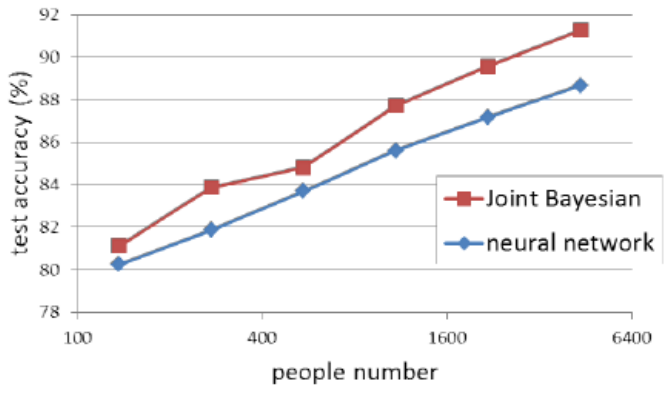

Figure 6. Face verification accuracy of Joint Bayesian (red line) and neural network (blue line) learned from the DeepID, where the ConvNets are trained with 136, 272, 544, 1087, 2175, and 4349 classes, respectively.

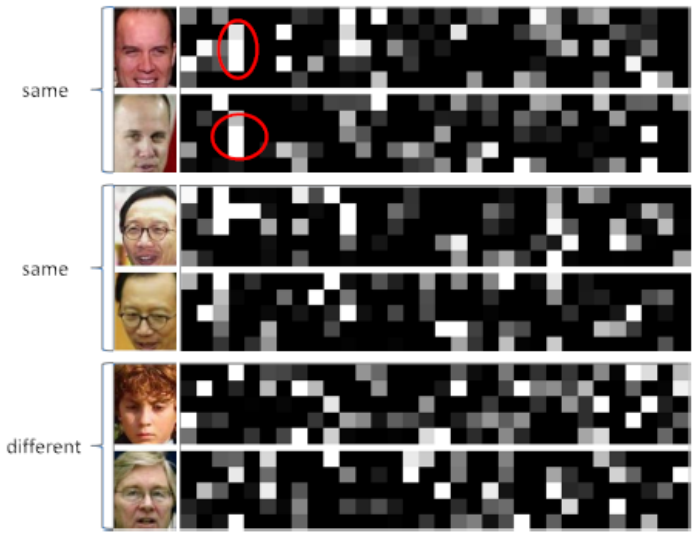

Figure 8. Examples of the learned 160-dimensional DeepID. The left column shows three test pairs in LFW. The first two pairs are of the same identity, the third one is of different identities. The corresponding features extracted from each patch are shown in the right. The features are in one dimension. We rearrange them as 5×32 for the convenience of illustration. The feature values are non-negative since they are taken from the ReLUs. Approximately 40% features have positive values. The brighter squares indicate higher values.

过完备表征：

这一小节做实验来判断多少的patch数提取到的特征，结合起来有最好的性能表现。

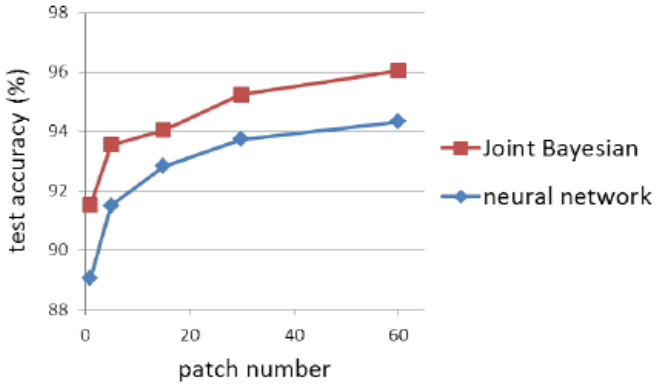

Figure 9. Test accuracy of Joint Bayesian (red line) and neural networks (blue line) using features extracted from 1, 5, 15, 30, and 60 patches. Performance consistently improves with more features. Joint Bayesian is approximately 1.8% better on average than neural networks.

方法比较：

此节实验中，将CelebFaces数据集扩展至CelebFaces+，10177个人（202599张图片），随机选取8700人来学习DeepID特征，剩余的1477人训练贝叶斯分类器；同时将patch的数目增加至100，用的是5个尺度，这样一来得到的是32000维的DeepID特征，然后用PCA降维至150。

由于数据分布不同，能够很好迎合CelebFaces+的模型不一定在LFW上有相同的表现力，文献【6】中提出了一个实用的传输学习算法，能使贝叶斯模型很好的适应源域到目标域的转换。达到97.45%准确率。

Method	Accuracy (%)	No. of points	No. of images	Feature dimension
Joint Bayesian [8]	92.42 (o)	5	99,773	2000×4
ConvNet-RBM [31]	92.52 (o)	3	87,628	N/A
CMD+SLBP [17]	92.58 (u)	3	N/A	2302
Fisher vector faces [29]	93.03 (u)	9	N/A	128×2
Tom-vs-Pete classifiers [2]	93.30 (o+r)	95	20,639	5000
High-dim LBP [9]	95.17 (o)	27	99,773	2000
TL Joint Bayesian [6]	96.33 (o+u)	27	99,773	2000
DeepFace [32]	97.25 (o+u)	6 + 67	4,400,000 + 3,000,000	4096×4
DeepID on CelebFaces	**96.05** (o)	5	87,628	150
DeepID on CelebFaces+	**97.20** (o)	5	202,599	150
DeepID on CelebFaces+ & TL	**97.45** (o+u)	5	202,599	150

Table 1. Comparison of state-of-the-art face verification methods on LFW. Column 2 compares accuracy. Letters in the parentheses denote the training protocols used. r denotes the restricted training protocol, where the 6000 face pairs given by LFW are used for ten-fold cross-validation. u denotes the unrestricted protocol, where additional training pairs can be generated from LFW using the identity information. o denotes using outside training data, however, without using training data from LFW. o+r denotes using both outside data and LFW data in the restricted protocol for training. (o+u) denotes using both outside data and LFW data in the unrestricted protocol for training. Column 3 compares the number of facial points used for alignment. Column 4 compares the number of outside images used for training (if applicable). The last column compares the final feature dimensions for each face (if applicable). DeepFace used six 2D points and 67 3D points for alignment. TL in our method means transfer learning Joint Bayesian.

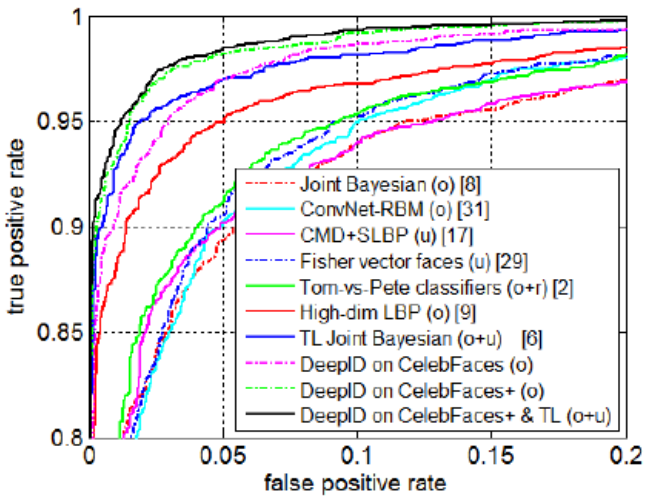

Figure 10. ROC comparison with the state-of-the-art face verification methods on LFW. TL in our method means transfer learning Joint Bayesian.