指导教师：杨涛　　提交时间：2016/03/18

CVPR2015 Paper Translation

No: 01

姓名：杨瑞安

学号：2013302507

学号：10011302

FaceNet：人脸识别和聚类统一的嵌入

Florian Schroff
fschroff@google.com
Google Inc.

Dmitry Kalenichenko
dkalenichenko@google.com
Google Inc.

James Philbin
jphilbin@google.com
Google Inc.

摘要

尽管近些年来我们在人脸识别领域取得了显著的进展[10，14，15，17]，但是有效地实现大规模的人脸校验和识别对当前的方法提出了严峻的挑战。本文中我们提出了一个叫做 FaceNet 的系统。该系统直接构造了一个从脸图像到一个紧凑的欧几里得空间的映射，使距离直接对应人脸相似度。一旦产生该空间，比如人脸识别、校验和聚类的任务就可以容易地通过以 FaceNet 嵌入为特征向量的标准技术来执行。

我们的方法是采用一个经过训练的深卷积网络去直接优化嵌入本身，不同于以前的深度学习方法采用中间瓶颈层实现优化。我们使用由一种新的在线三重挖掘方法衍生的大致对齐的匹配/不匹配脸补丁三元组进行训练。这种方法的好处是有更好的表征效率：我们对每张脸仅使用 128 字节就达到了最先进的人脸识别效果。

在广泛使用的任免数据库中，我们的系统达到了 99.63%的新纪录精度。而在 YouTube 人脸数据库中它达到 95.12%的精度。在两个数据库中，我们的系统相比于目前最好的已知结果[15]都降低了 30%的出错率。

1. 简介

在本文中我们提出了一个人脸校验（是否为同一个人）、识别（这个人是谁）和聚类（在这些人脸中寻找有共同点的人）相统一的系统。我们的方法是基于使用一个深卷积网络对每张图片构造一个欧式嵌入。该系统经过训练使嵌入空间里欧氏距离的平方直接对应人脸相似度：同一个人的不同的脸图片间只有很小的距离，而不同的人的脸图片之间存在很大的距离。

一旦这个嵌入已经产生，那么上述任务就会变得直截了当：人脸校检仅仅涉及两个嵌入之间距离的阈值；识别变成了一个 K-NN 分类问题；而聚类则可以通过现有的技术比如 K-means 或者凝结聚类来实现。

以前基于深层网络的人脸识别方法采用一个通过一组人脸身份的训练的分类层[15,17]，然后取一个中间瓶颈层作为一个表示用来推广不属于该组训练中身份的识别。这种方法的缺点是其间接性和低效率：一方面希望该瓶颈表示可以很好地在新的人脸中推广，但使用瓶颈层后每张人脸表示的大小通常会非常大（1000 维以上）。最近的一些研究[15]通过 PCA 减少了这个维度，但这是一个可以很容易从网络中的一层知道的线性变换。

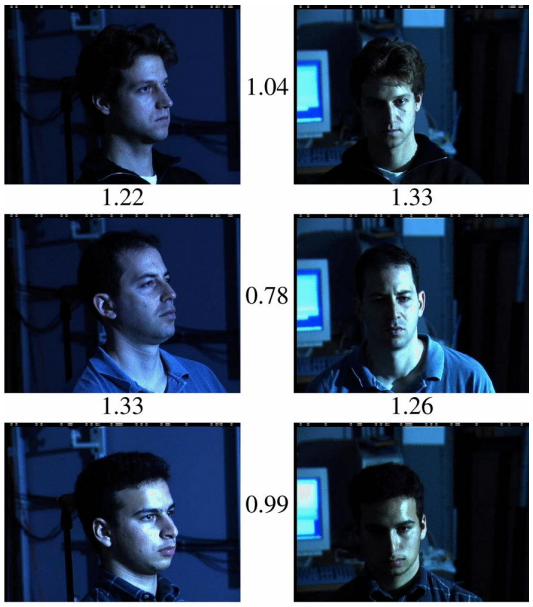

图 1.　光照和姿态不变。姿态和光照长久以来一直是人脸识别的一个问题。这张图显示了几组同一张脸和不同的脸在不哦那个的姿势和光照组合下的 FaceNet 输出距离。距离 0.0 意味着是同一张脸，距离 4.0 则对应着相反的频谱，即两张不同的面孔。你可以看到 1.1 的阈值可以把每一对图都正确分类。

与这些方法相比，FaceNet 用一个基于 LMNN[19]的三重损失函数直接把它的输出训练成一个紧凑的 128 维的嵌入。我们的三元组包括两个匹配的人脸略缩图和一个不匹配的人脸略缩图，而损失的目的是通过距离边界把有把握的配对和无把握的配对分离开。略缩图是人脸区域的紧缩结果，除了规模和转换的表现外并没有 2D 或者 3D 对齐。

选择使用哪一种三元组对于实现良好的性能是非常重要的。受课程学习启发[1]，我们提出一种新的在线无把握的样本挖掘方案来确保随着网络训练不断提高的三元组难度。为了提高聚类精度，我们也探索鼓励单人嵌入的球状聚集的高效挖掘技术。

我们的方法可以达到的不可思议的改变如图 1 所示。图中那些来自 PIE[13]的图像对在以前在人脸校验方面是非常难的。

本文其余部分概述如下：在第 2 节我们回顾这方面的文献；第 3.1 节界定三重损失，而 3.2 节描述了我们新的三元组选拔和培训程序；在 3.3 节我们描述了我们使用的模型架构。最后，在第 4 节和第 5 节我们提出一些嵌入的定量结果并且定性地探讨一些聚类结果。

2. 相关工作

类似于最近的其他采用深网络[15，17]的研究，我们的方法是一个纯粹的数据驱动方法，该方法直接从该脸的像素获悉其特征。不同于使用设计的特征的方法，我们使用大量的标记人脸数据集来达到对应姿势、光照和其他可变的条件的合适的不变性。

在本文中，我们探讨了两种最近在计算机视觉界取得巨大成功的不同的深层网络架构。两者都是深卷积网络[8,11]。第一个结构是基于 Zeiler&Fergus[22]模型，它由多个卷积交错层，非线性活动，局部响应归一和最大池层组成。此外，我们还添加了几个受[9]研究影响的 1×1×d 卷积层。第二种架构是基于 Szegedy 等的初始模型，最近它是 ImageNet 2014

年[16]的优胜方法。这些网络使用混合层运行几个不同的并行卷积和池层并且连接他们的反应。我们已经发现，这些模型可以减少高达 20 倍的参数量，并具有降低可比性表现所需的每秒浮作业数的潜力。

人脸校验和识别有一个巨大的语料库。考虑到这超出了本文的范围,所以我们只会简要讨论近期最相关的工作。

研究[15，17，23]都使用一个复杂的多阶段系统,它们结合了深卷积网络的输出与降维 PCA 和分类 SVM。

Zhenyao 等人[23]采用了深网络去把面孔“扭曲”成规范化主视图,然后学习 CNN,把每张脸分类到一个已知的身份。对于脸部校验,他们使用了在网络输出上与一个 SVM 集合的 PCA。

Taigman 等[17]提出一种多级方法,将人脸和通用 3 D 形状模型对齐。多级网络被训练成在超过四千个身份中进行人脸识别任务。作者还尝试了一个称作 Siamese 的网络，在该网络中他们直接优化两张人脸特征之间的曼哈顿距离。他们最好的 LFW 成效(97.35%)是来自一个三个网络使用不同的排列和颜色通道的集合。这些网络的预测的距离（基于 χ^2 内核非线性 SVM 预测）通过一个非线性 SVM 进行组合。

Sun 等人[14、15]提出一个压缩,因此对计算网络会相对廉价。他们采用一个由 25 个这样的网络组成的集合,每个都在不同的补丁下操作。对他们的最终性能 LFW(99.47%[15])作者结合了 50 种响应(常规的和翻转的)。PCA 和能在嵌入空间中有效地对应一个线性变换的联合贝叶斯模型[2]都被采用。他们的方法不要求明确的 2 D / 3 D 对齐。其网络通过分类和校验损失的组合进行训练。校验损失类似于我们采用三重损失[12,19]，因为它最小化了相同的身份的脸之间的欧式距离并固定了不同身份的脸的距离之间的界限。主要的区别在于,只有成对图像进行比较,而三重损失鼓励相对距离约束。

在这里所使用的一种类似的损失是由 Wang[8]等人进行了探究,用于通过语义和视觉相似度给图像排序。

3.方法

FaceNet 采用了深卷积网络。我们讨论了两种不同的内核架构：Zeiler&Fergus 风格的网络和最近的 Inception 类型的网络。这些网络的细节在 3.3 节中将给予描述。已知这些模型细节，并把它当作一个黑盒子（见图 2），我们的方法最重要的部分就在于端对端地对整个系统的学习。为此我们采用直接反映我们想要在人脸校验、识别和聚类上达到的目标的三重损失。也就是说我们寻求一个嵌入函数 $f(x)$，将一张图像 x 转换成一个特征空间 R^d，这样的话忽略成像条件，所有属于同一个人的脸之间的平方距离会很小，而属于不同的人的脸之间

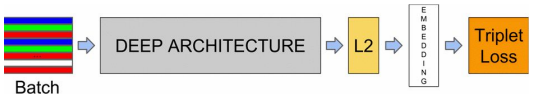

图 2. 模型结构。我们的网络包括一批输入层和一个深胞状类神经网路（CNN），接着是可以生成人脸嵌入的标准化欧氏距离。然后是训练时的三重损失。

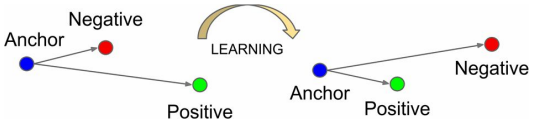

图 3. 三重损失把标本和属于同一身份的匹配图像之间的距离最小化，并最大化和不同身份的图像之间的距离。

的平方距离是很大的。

虽然我们没有和其他的损失直接作比较，比如在研究[14]的公式(2)中使用正反对，但我们认为三重损失会更加适合人脸校验。原因是[14]里面的损失鼓励所有属于同一个身份的脸按照 â˜AŸ映射到嵌入空间的一个独立的点的â˜ A´Z。而三重损失试着设立来自一个人的所有面孔和其他面孔的边界。这使得属于同一个身份的脸都在同一个类别中，同时还保证了和其他身份的距离和区分度。

以下部分介绍这种三重损失和它是怎样有效地大规模学习。

3.1 三重损失

该嵌入是由属于 R^d 的函数 $f(x)$ 表示的。它把图像 x 嵌入一个 d 维欧几里德空间。此外，我们限制该嵌入存在于 d 维球面中，即 $\|f(x)\|_2=1$。这种损失是在研究[19]的最邻近分类的上下文中提出的。再这里我们想确保某人的图像 x_i^a（标本）更接近同一个人的其他所有图像 x_i^p（匹配）而不是其他任何人的图像 x_i^n（不匹配）如图 3 所示。

所以我们希望

$$\|x_i^a-x_i^p\|_2^2+\alpha<\|x_i^a-x_i^n\|_2^2,\forall(x_i^a,x_i^p,x_i^n)\in\Gamma \quad (1)$$

这里 α 是设定的正误对之间的界限。Γ 是在训练组中全部可能三元组的集合而且其基数为 N。

最小化损失是 $L=$

$$\sum_i^N[\|f(x_i^a)-f(x_i^p)\|_2^2-\|f(x_i^a)-f(x_i^n)\|_2^2+\alpha]_+ \quad (2)$$

生成的所有可能三元组将使许多三元组容易地满足条件（即满足公式(1)中的约束）。这些三元组无助于训练而且会导致更慢的收敛，因为他们仍通过该网络传递。关键是选择生硬的三元组，他们是活跃的所以有益于改进模型。下面的章节将谈到在三元组选择中我们采用的不同方法。

3.2 三元组选择

为了确保快速收敛，选择那些不符合公式(1)的约束的三元组是至关重要的。这意味着给定 x_i^a，我们想选择一个 x_i^p（绝对匹配）比如 $\arg\max_{x_i^p}\|f(x_i^a)-f(x_i^p)\|_2^2$ 以及一个 x_i^n（绝对不匹配）比如 $\arg\min_{x_i^n}\|f(x_i^a)-f(x_i^n)\|_2^2$。

对整个训练集计算 argmax 和

argmin 是不可行的。另外，它还可能导致低效训练，因为贴错标签和不好的脸图像会导致绝对正确和绝对错误结果。要避免这个问题有两个明显的选择：

- 每 n 步离线产生三元组，使用最新的网络检查点，并计算数据子集的 argmax 和 argmin。
- 在线产生三元组。这可以通过从一个小批次中选择绝对正确/错误样本来实现。

在这里，我们专注于在线生成和在几千样本的顺序下使用大量小批次并仅在一个小批量内计算 argmax 和 argmin。

为了得到一个有意义的标本正距离的表示，我们需要确保任何一个身份的样本最小值在每个小批次里表示出了。在我们的实验中，我们像每个小批次每个身份选 40 张脸图像来采集数据。另外，随机取样的错误面孔会加到每个小批次里面。

不同于选择最符合的图像对，我们采用小批次里面全部的标本匹配对，当然包括那些不符合的图像对。我们没有把一个小批次里的绝对正确的标本和所有匹配标本并排比较，但是我们在实践中发现所有的标本匹配方法在开始训练的时候都更加稳定并更快地收敛。

我们也探究过离线三元组生成和在线生成的结合，这样可以使用更小的批量，但是并没有经过实验验证。

在实践的时候选择最不符合的图像会很早地在训练中导致不好的局部最小值，具体来说他会导致一个收缩模型（即 $f(x)=0$ ）。选择一个合适的 x_i^n 可以减轻该现象，例如

$$\|f(x_i^a)-f(x_i^p)\|_2^2<\|f(x_i^a)-f(x_i^n)\|_2^2 \quad (3)$$

我们把这些不完全符合的样本称作半适应，因为他们还和那些符合的样本差很多，但是因为平方距离接近适应标本距离所以仍然认为适应。这些不完全符合的样本处于边界 α 之内。

如前面提到的，正确的三元组选择是快速收敛的关键。一方面我们想用迷你小批量，因为他们在随机梯度下降算法（SGD）[20]中可以有助于收敛。另一方面，实施细节使得几十上百的样本批次更加高效。对于批次大小的主要约束是我们在迷你批次中选择绝对相关三元组的方式。在大多数实验中我们使用大小约为 1,800 个样本的批次。

3.3 深卷积网络

在我们所有的实验中我们都是使用带标准反向传播和 AdaGrad 的随机梯度下降算法（SGD）来训练胞状类神经网路的。在大多数实验里面我们降低学习率以保证模型完成，所以先以 0.05 的学习率开始。模型是随机初始化的，类似于研究[16]，并且在一个 CPU 集群中训练了 1,000 到 2,000 小时。损失的下降在经过 500 小时的训练后大幅度地减缓，但是额外的训练仍然可以显著地提高性能。边界 α 设定为 0.2。

layer	size-in	size-out	kernel	param	FLPS
conv1	220×220×3	110×110×64	7×7×3, 2	9K	115M
pool1	110×110×64	55×55×64	3×3×64, 2	0	
rnorm1	55×55×64	55×55×64		0	
conv2a	55×55×64	55×55×64	1×1×64, 1	4K	13M
conv2	55×55×64	55×55×192	3×3×64, 1	111K	335M
rnorm2	55×55×192	55×55×192		0	
pool2	55×55×192	28×28×192	3×3×192, 2	0	
conv3a	28×28×192	28×28×192	1×1×192, 1	37K	29M
conv3	28×28×192	28×28×384	3×3×192, 1	664K	521M
pool3	28×28×384	14×14×384	3×3×384, 2	0	
conv4a	14×14×384	14×14×384	1×1×384, 1	148K	29M
conv4	14×14×384	14×14×256	3×3×384, 1	885K	173M
conv5a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv5	14×14×256	14×14×256	3×3×256, 1	590K	116M
conv6a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv6	14×14×256	14×14×256	3×3×256, 1	590K	116M
pool4	14×14×256	7×7×256	3×3×256, 2	0	
concat	7×7×256	7×7×256		0	
fc1	7×7×256	1×32×128	maxout p=2	103M	103M
fc2	1×32×128	1×32×128	maxout p=2	34M	34M
fc7128	1×32×128	1×1×128		524K	0.5M
L2	1×1×128	1×1×128		0	
total				140M	1.6B

表 1. NN1. 本表展示了我们受研究[9]启发的基于 Zeiler&Fergus [22]带 1x1 卷积的模型的结构。输入和输出大小由行 x 列 x#滤波数描述。其内核有特定的行 x 列，跨步和尺寸为 P=2 的 maxout[6]池。

我们使用了两种类型的体系结构而且在实验部分更详细地探讨了它们的得失。它们的实际区别在于参数和 FLOPS 的不同。应用的不同可能导致最佳模式的不同。例如一个在数据中心运行的模型可能需要大量的参数和很大的 FLOPS，而一个在手机上运行的模型只需要很少的参数，以便于它可以放到内存中。我们所有的模型都使用修正线性单元作为非线性激活功能。

表 1 所示的第一个类别，如研究[9]中建议的，在 Zeiler&Fergus[22]架构的标准卷积层之间添加了 1x1xd 卷积层，并且引起了模型达到 22 层深。它总共有 1.4 亿的参数而且要求每张图象有大约 160 亿的 FLOPS。

我们使用的第二个类别基于 GoogleNet 风格的 Inception 模型[16]。这些模型有 20 倍的参数（大约 660 万到 750 万）和最高 5 倍的 FLOPS（在 5 亿到 16 亿之间）。有些模型的尺寸（深度和滤波器数量）显著减少，所以他们可以在移动电话上运行。其一，NNS1，有 2,600 万的参数而且对每张图像仅需要 2.2 亿的 FLOPS。另一个，NNS2，有 430 万的参数以及 2,000 万的 FLOPS。表 2 详细介绍了我们最大的网络 NN2。NN3 在构架上是相同的，但其输入的大小减少为 160x160。NN4 的输入尺寸只有 96x96，从而大大地降低了 CPU 需求（2.85 亿 FLOPS 对比 NN2 的 16 亿）。除了减小输入尺寸，它并没有在更高的层里使用 5x5 卷积，因为那样的话接受场会太小。通常我们发现 5x5 卷积会被轻微的精度下降所消除。图 4 中我们比较了我们所有的模型。

4. 数据集和评估

我们在四组数据集上评估我们的方法，并除去在 Wild 和 YouTube 面孔库里面的标记面孔，我们用人脸校验任务评估我们的方法，即给定一对人脸图像的欧氏距离的平方阈值 $D(x_i, x_i)$，用于确定是否为相同的身份。所有的统一身份的人脸对 (i, j) 都标以 $\mathrm{P}_{同}$，而全部不同身份的人脸对都标以 $\mathrm{P}_{异}$。

我们定义全部判断正确的组为

type	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj (p)	params	FLOPS
conv1 (7×7×3, 2)	112×112×64	1							9K	119M
max pool + norm	56×56×64	0						m 3×3, 2		
inception (2)	56×56×192	2		64	192				115K	360M
norm + max pool	28×28×192	0						m 3×3, 2		
inception (3a)	28×28×256	2	64	96	128	16	32	m, 32p	164K	128M
inception (3b)	28×28×320	2	64	96	128	32	64	L_2, 64p	228K	179M
inception (3c)	14×14×640	2	0	128	256,2	32	64,2	m 3×3,2	398K	108M
inception (4a)	14×14×640	2	256	96	192	32	64	L_2, 128p	545K	107M
inception (4b)	14×14×640	2	224	112	224	32	64	L_2, 128p	595K	117M
inception (4c)	14×14×640	2	192	128	256	32	64	L_2, 128p	654K	128M
inception (4d)	14×14×640	2	160	144	288	32	64	L_2, 128p	722K	142M
inception (4e)	7×7×1024	2	0	160	256,2	64	128,2	m 3×3,2	717K	56M
inception (5a)	7×7×1024	2	384	192	384	48	128	L_2, 128p	1.6M	78M
inception (5b)	7×7×1024	2	384	192	384	48	128	m, 128p	1.6M	78M
avg pool	1×1×1024	0								
fully conn	1×1×128	1							131K	0.1M
L2 normalization	1×1×128	0								
total									7.5M	1.6B

表 2. NN2. NN2 Inception 模型可视化的细节。此模型和研究[16]中描述的几乎一模一样。主要的两个区别是此模型特别地使用了欧式距离池而不是最大池。池大小始终是 3x3（除了最终的平均池），而且在每个 Inception 模块里都并行于卷积模块。如果在该池用 p 表示之后有降维，那么 1x1，3x3，5x5 的池会相连接得到最终的输出。

$$TA(d)=\{(i,j)\in \mathrm{P}_{同}, 并且 D(x_i,x_j)\le d\} \quad (4)$$

这些人脸对(i,j)都是在阈值d下判定为统一身份的。相似地，

$$FA(d)=\{(i,j)\in \mathrm{P}_{异}, \text{ 而且} D(x_i,x_j)\le d\} \quad (5)$$

是那些被错误地判定为统一身份的对（容错）。

一个给定的人脸距离d的验证率$VAL(d)$和容错率$FAR(d)$则定义为

$$VAL(d)=\frac{|TA(d)|}{|\mathrm{P}_{同}|}$$

$$FAR(d)=\frac{|FA(d)|}{|\mathrm{P}_{异}|} \quad (6)$$

4.1 抵抗测试集

我们保留了一组由大约一百万张图像组成的抵抗测试集。它有和我们的训练集一样的分布，但其中身份都不一样。在评估时我们把它划分成五个互不相关的组，每组 20 万张图像。验证率FAR和容错率VAL以 100,000x 10,000 的图像单元计算。通过这五组可以体现出标准误差。

4.2 个人照片

这个测试集和我们的训练集有着类似的分布，但里面的图像已经经过手动分类到具体的标签下了。它一共有约 12,000 张图像，里面包含了三个人的照片集。我们对全部$12{,}000^2$对图像计算FAR和VAL。

4.3 学术数据集

LFW 实际上是人脸校验的学术测试集。我们遵循无限制的、数据外部标记的标准协议，并报告平均分类精度以及平均值的标准误差。

YouTube 人脸数据库是一个新的数

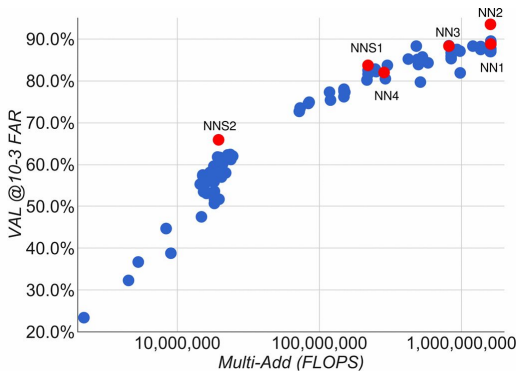

图 4. FLOPS vs. 精度权衡。图示为对一个大范围的不同模型尺寸和结构的 FLOPS 和精度间的权衡。高亮的是我们在实验中关注的四种模型。

architecture	VAL
NN1 (Zeiler&Fergus 220×220)	87.9% ± 1.9
NN2 (Inception 224×224)	89.4% ± 1.6
NN3 (Inception 160×160)	88.3% ± 1.7
NN4 (Inception 96×96)	82.0% ± 2.3
NNS1 (mini Inception 165×165)	82.4% ± 2.4
NNS2 (tiny Inception 140×116)	51.9% ± 2.9

表 3. 网络结构。 表中比较了我们的几种模型结构在抵抗测试集上的表现。图示了在 $10\,e^{-3}$ 的容错率下的平均校验率 VAL ，并且说明了五个测试分组的平均标准差。

据集，在人脸识别领域得到了普及。其设定类似于 FLW，但并不是用来图像验证，而是视频的校验。

5. 实验

如果之前没有提到，我们使用由大约 800 万不同身份组成的 1 亿到 2 亿的训练人脸缩略图。人脸检测器扫描每一张图片并产生紧密相关的边框。这些人脸缩略图会调整成不同网络的输入大小。我们的实验中输入大小从 96x96 像素到 224x224 像素不等。

5.1. 计算精度权衡

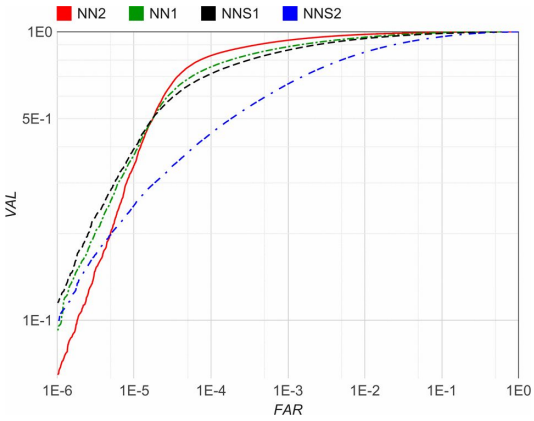

图 5. 网络结构。本图给出了 4.2 节中四种不同模型在我们的个人照片测试集上的完整 ROC 曲线。在 FAR 为 $10\,e^{-4}$ 处的急剧下降可以解释为实际情况标记中的噪音。模型按照性能排序为：NN2：输入像素为 224x224 的基于 Inception 的模型；NN1：基于 Zeiler&Fergus 的带 1x1 卷积的网络；NNS1：只有 2.2 亿 FLOPS 的小型 Inception 类模型；NNS2：只有 2,000 万 FLOPS 的微型 Inception 模型。

在深入更具体的实细节前，我们讨论了对应特殊模型需要的 FLOPS 数量的精度权衡。图 4 显示了 x 轴上的 FLOPS 和第 4.2 节我们的用户标记的测试数据集的 0.001 精度的误识率。我们可以看到模型需要的运算和它可以达到的精度有很大的联系。该图突出了我们在实验中更详细讨论的 5 个模型（NN1, NN2, NN3, NNS1, NNS2）。

考虑到模型参数的量，我们也研究了准确性权衡。然而，图片并不那么清晰。比如，基于 Inception 的模型 NN2 达到的性能可以和 NN1 相比，但是其参数只有它的 1/20。尽管它们的 FLOPS 的数值是差不多的。很明显，在某些情况下如果参数量大幅下降，性能也会下降。其他模型架构也许可以在不损失精度的情况下大量减少参

jpeg q	val-rate
10	67.3%
20	81.4%
30	83.9%
50	85.5%
70	86.1%
90	86.5%

#pixels	val-rate
1,600	37.8%
6,400	79.5%
14,400	84.5%
25,600	85.7%
65,536	86.4%

表 4. 图像质量。左边的表展示了对不同质量的 JPEG 图像我们的精度为$10e^{-3}$的校验率。右边的表则说明了图像像素大小是如何影响精度为$10e^{-3}$的校验率的。本次实验是在我们的测试抵抗数据集的第一组上采用 NN1 模型完成的。

#dims	VAL
64	$86.8\% \pm 1.7$
128	$87.9\% \pm 1.9$
256	$87.7\% \pm 1.9$
512	$85.6\% \pm 2.0$

表 5. 嵌入维数。本表比较了我们的 NN1 模型在 4.1 节的抵抗集上各个嵌入维数的效果。除了精度为$10e^{-3}$的校验率我们还展示了对五个分组计算出的平均标准误差。

数，就像在这次试验中的 Inception 模型。

5.2. CNN 模型的效果

我们现在更详细地讨论我们选出来的四个模型的性能。一个是传统的带有 1×1 卷积的基于 Zeiler&Fergus 的架构（如表一所示）[22,9]。另一个是显著减小模型规模的基于 Inception 的模型。总的来说，两种结构的最顶尖模型的最后性能差不多。然而，一些基于 Inception 结构的模型，例如 NN3 在 FLOPS 和模型规模显著减小时依然表现出很好的性能。

关于我们个人照片测试集的详细评估如图 5 所示。虽然相比小型的 NNS2，我们最大的模型在精度方面有很大的提高，但 NNS2 可以在手机上每 30ms 扫描一张图像并且在人脸聚类上有足够的精确度。在 ROC 曲线上 $FAR < 10^{-4}$ 处的剧烈下降表明了测试数据实况的噪音标签。在极低的容错率下一个错误标记的图片就可以对曲线产生极大的影响。

5.3 图像质量的敏感度

表 4 显示了我们的模型在一个很大的图片大小范围内的鲁棒性。该网络在 JPEG 图片压缩上惊人地强大，而且对下至 20 质量的 JPEG 图片都能表现很好。人脸缩略图减小到 120x120 大小时性能下降也很小，甚至，对 80x80 像素的图片其表现也可接受。因为此网络是在 220x220 的输入图像下测试的，所以该结果值得注意。用更低分辨率的人脸图像训练可以更好地提高识别范围。

5.4. 嵌入维数

我们研究了各种嵌入维数，然后给所有实验都选择了 128，不同于表 5 中显示的。你可能会认为更大的嵌入维数在性能方面至少和更小的一样好，然而它们可能需要更多的训练来达到一样高的精度。也就是说，表 5 中体现的性能的差异没有统计学意义。

值得注意的是，在训练中我们使用了一个 128 维的浮点矢量，它可以量化为 128 字节而不损失精度。因此每

#training images	VAL
2,600,000	76.3%
26,000,000	85.1%
52,000,000	85.1%
260,000,000	86.2%

表 6.　训练数据大小。表中比较了对输入为 96x96 像素的小模型进行 700 小时的训练之后的性能。其模型结构类似于 NN2，但在 Inception 模块中不包含 5x5 卷积。

图 6.　LFW 失误。上图展示了所有在 LFW 中被错误分类的图像。

一张脸都由 128 维的字节矢量紧密表示，有利于大范围的聚类和识别。更小的嵌入也有可能仅损失少量精度并且可以用在移动设备上。

5.5. 训练数据量

表 6 显示的是大量训练数据的结果。由于时间的限制，我们在更小的模型上进行评估。如果在更大的模型上评估，效果可能更显著。很明显，在第 4.2 节个人照片测试集中使用几千万的示例结果可以明显提高精度。相对于只用几百万的图像，错误率减少了 60%；使用另一个数量级（几亿）的图像依然有少量提高，但是提高逐渐减少。

5.6 在 LFW 上的性能

在 LFW 上我们遵循无限制的，外标记数据的标准协议测试了我们的模型。9 个训练子集被用来挑选欧式距离阈值。在第 10 个训练子集中则表现出分类（相同或不同）。全部子集里挑选出的最佳阈值是 1.242，除了第八个子集（1.256）。

我们的模型用以下两种模式评估：

1. 修正 LFW 提供的缩略图的中心区域裁剪

2. 在 LFW 提供的缩略图上运行专有人脸探测器（类似于 Picasa[3]）。如果不能对齐人脸（发生在有两张图象时），就会使用 LFW 调正。

图 6 给出了所有失败案例的概况。它显示了最高的容错率和最低的拒错率。在使用（1）描述的修正中心裁剪时我们的分类精度达到了 98.87%±0.15，而且在使用（2）中的额外人脸对齐时我们打破了 99.63%±0.09 的标准平均误差记录。这比在[17]中的 DeepFace 所公布的误差减少了超过 7 个因子，而且比[15]中宣告的目前最先进的 DeepId2＋减少了 30%。这是模型 NN1 的性能，但是更小的 NN3 的性能与其在统计学上也无差异。

图 7.　人脸聚类。图示为一个用户的样例聚类。该用户个人照片集中所有图像都聚类在一起了。

5.7. 在 YouTube 人脸数据库上的性能

我们使用了人脸探测器在每一个录像机上检测到的的前 100 帧的平均相似性，并得到了 95.12%±0.39 的分类精度，而使用前 1000 帧则得出 95.18%的精度。相比也评估了前 100 帧所得的 91.4%的研究[17]，我们几乎减少了一半的误差。DeepId2[15]的精度达到了 93.2%，而我们的方法把这个误差减小了 30%，比得上我们在 FLW 上的提高。

5.8 人脸聚类

我们紧凑的嵌入使它可以把用户地个人照片聚类成具有相同身份的人群。相比于单纯的校验任务，聚类人脸的强制分配的约束导致了惊人的结果。图 7 展示了一组使用聚类生成的用户个人照片集。这是一个清晰的展示案例，显示了难以置信的阻塞、亮光、姿势甚至年龄的不变性。

6. 总结

我们为人脸识别提供了一个直接构造一个嵌入到欧几里得空间的方法。这使它不同于那些使用 CNN 瓶颈层或需要而外的后期处理的方法，如多个模型和 PCA 的级联，以及 SVM 分类法。我们的端到端训练简化了设定并且说明了直接优化任务相关损失可以适度地提高性能。

我们模型的另一个优势是它只需要最少的校正（脸部周围区域轻度的剪裁）[17]，例如执行一个复杂的 3D 对齐。我们还试验了一个相似度转换对齐，并注意到它确实可以略微提高性能。但这些额外的复杂性是否值得目前尚不清楚。

未来的工作将专注于更好地理解错误情况，减小模型大小和降低 CPU 需求上面。我们也会寻找方法去减少目前极长的训练时间，例如改变我们的小批量在线/离线的合格/不合格数据挖掘的课程学习。

致谢

在此我们特别感谢 Johannes Steffens 在人脸识别方面的论述与深刻见解以及 Christian Szegedy 提供的如[16]中的新型网络结构和对网络设计选择的讨论。同时我们也很感谢 DistBelief [4]团队的贡献，尤其是 RajatMonga 在设定高效的训练体系上给出的帮助。

当然，我们工作的完成也离不开 Chuck Rosenberg, Hartwig Adam 和 Simon Han 的支持。

参考文献

[1] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In Proc. of ICML, New York, NY,USA, 2009. 2

[2] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In Proc. ECCV, 2012. 2

[3] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In Proc. ECCV, 2014.8

[4] J.Dean, G.Corrado, R.Monga, K.Chen, M.Devin, M.Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q. V. Le, and A. Y. Ng.Large scale distributed deep networks. In P. Bartlett, F. Pereira, C.Burges, L. Bottou, and K. Weinberger, editors, NIPS, pages 1232–1240. 2012.9

[5] J. Duchi, E. Hazan, and Y.Singer. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12: 2121-2159, July 2011.4

[6] I. J. Goodfellow, D. Warde- farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In In ICML, 2013.4

[7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07- 49, University of Massachusetts, Amherst, October 2007. 5

[8] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4):541–551, Dec. 1989. 2, 4

[9] M. Lin, Q. Chen, and S. Yan. Network in network. CoRR, abs/1312.4400, 2013. 2, 4, 6

[10] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with gaussianface. CoRR, abs/1404.3840, 2014. 1

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representa- tions by backpropagating errors. Nature, 1986. 2, 4

[12] M.Schultz and T.Joachims. Learning a distance metric from relative comparisons. In S.Thrun, L.Saul, and B.Schölkopf, editors, NIPS, pages 41–48. MIT Press, 2004. 2

[13] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In In Proc. FG, 2002. 2

[14] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. CoRR, abs/1406.4773, 2014. 1, 2, 3

[15] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. CoRR, abs/1412.1265, 2014. 1, 2, 5, 8

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. CoRR, abs/1409.4842, 2014. 2, 4, 5, 6, 9

[17] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In IEEE Conf. on CVPR, 2014. 1, 2, 5, 8

[18] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. CoRR, abs/1404.4661,2014. 2

[19] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In NIPS. MIT Press, 2006. 2, 3

[20] D. R. Wilson and T. R. Martinez. The general inefficiency of batch training for gradient descent learning. Neural Networks,16(10):1429–1451, 2003. 4

[21] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In IEEE Conf. on CVPR, 2011. 5

[22] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013. 2, 4, 6

[23] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonicalview faces in the wild with deep neural networks. CoRR, abs/1404.3543, 2014. 2