Alibaba Group 阿里巴巴集团 | 阿里技术

阿里巴巴机器智能

计算机视觉技术精选专集

顶级学术会议ACM MM-2017收录论文

卷积神经网络 |

深度神经网络 |

视频异常检测 |

生成对抗网络(GAN) |

编者序：

在 ACM MM 2017 会议上，阿里巴巴有 3 篇论文被收录，论文中的技术研究起点均来自“城市大脑”项目，这些研究成果也保证了“城市大脑”的真实落地。

2016 年阿里巴巴推出了人工智能“城市大脑”项目，通过为城市安装人工智能枢纽，以摄像头为核心进行图像数据采集与计算，对整个城市进行全局实时分析，自动调配公共资源，修正城市运行中的 Bug。2017 年，城市大脑成为首批国家新一代人工智能开放创新平台。

“城市大脑”项目研究内容涵盖交通事故识别、人流轨迹判断以及交通数据样本的汇总等。包含多项人工智能技术，如视觉认知、优化决策、视觉搜索、预测、大规模实时视频处理系统等。

通过“城市大脑”项目，我们发现诸多待解决的问题，比如人流、车辆轨迹如何准确识别，如何提取三位空间中的物体特征等，经过不断实践，我们找到了一些比较好的解决方法，并将这些方法投入到实际场景中去正向增强应用的落地性。

在论文《Spatio-Temporal AutoEncoder for Video Anomaly Detection》里，我们为城市大脑提供监控交通异常的方法。受动作识别等领域的最新研究成果启发，设计了一种时空自编码进行视频异常检测，同时提出一种权重递减的预测误差计算方法。经真实的交通场景评测，该算法在重要指标上已经超过了此前的最好方法。

在论文《Deep Siamese Networkwith Multi-level Similarity Perception for Person Re-identification》，我们为人流轨迹的识别判断提供技术支持。结合深度学习的 Siamese 网络和分类网络模型优势，同时将相似度扩展到其他层次。与经典的大规模同人鉴别公开数据集对比，目前检索精确度的最优结果已达业内最高水平。

在实际应用中，我们发现城市大脑交通视频数据样本不足的问题，因此提出了一种图像生成算法，在论文《Stylized AdversarialAutoencoder for Image Generation》做了详细介绍。受条件对抗生成网络和风格迁移学习的启发，采用内容提取网络和风格提取网络分别从内容图片和风格图片中提取特征，将两者融合后，通过图片生成网络获得融合相应内容和风格的图片。

图像的数据量正覆盖着我们的星球，对计算机视觉能力的挖掘，人类也刚迈出一小步，并在个别可见的领域取得些许进展。我们将 ACM MM-2017 会议上收录的论文编辑成册，希望在便于读者观看的同时，能够和更多学术界、工业界同仁一起探讨，共同推进计算机视觉技术的发展和应用。

阿里巴巴机器智能计算机视觉技术精选编写组

2018 年 4 月

目录

Spatio-Temporal Auto Encoder for Video Anomaly Detection

时空自编码器的视频异常检测模型

主要作者（中英文）：
赵一儒 Yiru Zhao/ 邓兵 Bing Deng /申晨 Chen Shen/刘垚 Yao Liu/
卢宏涛 Hongtao Lu/华先胜 Xian-Sheng Hua
论文原文地址：https://dl.acm.org/citation.cfm?id=3123451

Spatio-Temporal AutoEncoder for Video Anomaly Detection

Yiru Zhao*
Shanghai Jiao Tong University
Alibaba Group
yiru.zhao@sjtu.edu.cn

Bing Deng
Alibaba Group
dengbing.db@alibaba-inc.com

Chen Shen†
Zhejiang University
Alibaba Group
zjushenchen@gmail.com

Yao Liu
Alibaba Group
xuanyao0111@gmail.com

Hongtao Lu‡
Shanghai Jiao Tong University
htlu@sjtu.edu.cn

Xian-Sheng Hua§
Alibaba Group
huaxiansheng@gmail.com

摘要

真实世界视频场景中的异常事件检测是一个高难度的问题，因为“异常”本身很复杂而且场景中还存在杂乱的背景、物体和运动。大多数已有的方法都是在局部空间区域中使用人工设计的特征来识别异常。在本论文中，我们提出了一种称为时空自编码器（Spatio-Temporal AutoEncoder，简称 ST AutoEncoder 或 STAE）的全新模型，使用深度神经网络来自动学习视频表征以及通过执行三维卷积来从空间维度和时间维度提取特征。在经典的自编码器中所使用的重建损失之外，我们为未来帧的生成引入了一种权重递减型预测损失，这能够增强视频中的运动特征学习。因为大多数异常检测数据集都局限于外观异常或不自然的运动异常，所以我们收集了一个新的高难度数据集，该数据集是由真实世界的交通监控视频构成的。我们在公开数据集和我们的交通数据集上进行了多项实验，结果表明我们提出的方法的表现显著优于之前最佳的方法。

1 引言

自动检测视频流中的异常事件是智能视频监控系统面临的一大基本难题，并且已经在过去几年中受到了学术界和工业界的高度关注。不同于动作识别和事件检测等监督式视频分析问题，视频异常检测主要面临着两大难题：一是正例样本和负例样本之间的数据不平衡（即作为正例样本的异常事件的数量远远少于常规事件）；二是正例样本内部存在很大的差异性（异常事件可能包含很多不同的情况，但一般而言可用的训练数据却很有限）。由于正例样本的稀疏性，经典的监督式事件检测和识别算法无法应用于这个任务。这个问题的通常解决方式是使用无监督方法训练一个表征正常视频序列中的模型，然后将异常值（模型的外点）看作是异常事件。

鉴于训练数据通常只包含普通视频，所以学习常规活动的特征表征是一个无监督学习问题。之前的一部分异常检测研究侧重于建模局部 2D 图像图块或 3D 视频立方体的时空事件模式，这个过程中会用到从低层面外观和运动中提取的人

工设计的特征，比如方向梯度直方图（HOG）、光流直方图（HOF）、3D 时空梯度等。但是，由于人工设计的特征的表征能力有限，这一类之前的方法并不适合用来分析复杂的视频监控场景。

深度学习方法已经展现出了在特征学习方面的优势，而且研究已经证明其可以非常有效地解决鉴别式视觉任务。基于自编码器网络的无监督深度学习方法也已被提出用作解决视频异常检测问题的又一类方法。但是，这些方法只依赖于全连接的自编码器或 2D 卷积自编码器，而没有利用来自时间维度的特征，因此无法获取异常事件的时间线索，而这对于识别视频事件异常而言是至关重要的。

受 3D 卷积网络在视频分析中的优越表现的启发，我们提出了用于视频异常检测的时空（ST）自编码器：通过在编码器中应用 3D 卷积和在解码器中应用 3D 反卷积，能够增强模型从时间维度中提取运动模式的能力。除了经典的自编码器所使用的重建损失，我们还引入了一种权重递减型预测损失来预测未来帧，这可以引导模型获取运动目标的轨迹以及增强编码器以更好地提取时间特征。经过在正常视频数据上的训练之后，该自编码器应该能够以较低误差重建出常规视频片段，而在重建非常规视频片段时则会出现高误差。然后模型再根据这个误差计算视频序列中每一帧的规律性分数（regularity score），然后再将其用于确定异常事件，如图 1 所示。

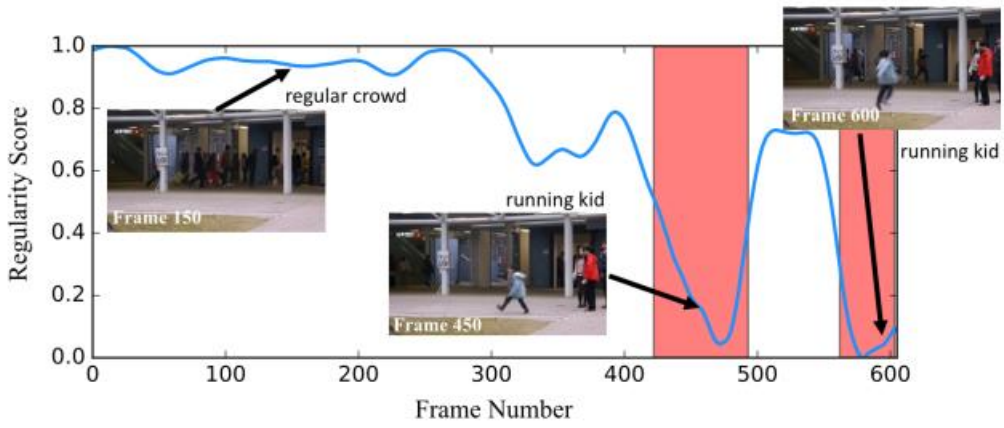

图 1：来自 CUHK Avenue 数据集的一段视频序列的规律性分数。红色区域表示基本真值异常帧。规律性分数会在异常事件发生时下降。

大多数真实世界情形中的异常事件都非常复杂，而大多数当前的异常检测数据集都只包含外观异常或人为制造的运动异常。为了评估我们提出的方法的实用性，我们收集了一个新的高难度数据集，其由真实世界交通监控视频构成。实验表明我们的模型可以应用于这一复杂应用。

本论文的主要贡献总结如下：

- 我们提出了一种全新的时空自编码器深度网络，可以通过执行 3D 卷积同时根据空间维度和时间维度来建模常规视频数据。据我们所知，这是首个基于 3D 卷积的视频异常检测模型。
- 我们在模型训练中引入了一个权重递减型预测损失，这能提升检测异常事件的表现。
- 我们收集了一个新的由真实世界交通监控视频构成的异常检测数据集，并且表明我们的方法的表现在公共基准和我们的 Traffic 数据集上都优于之前最佳的方法。

2 相关工作（略）

3 我们的方法

为了具体描述，我们首先简要介绍一下 3D 卷积，然后再详细讨论我们提出的模型。

3.1 3D 卷积

典型的 2D 卷积网络是在 2D 特征图上应用卷积来提取空间维度的特征。2D 卷积网络在图像识别方面表现优越，但它们却无法获取用于视频分析问题的连续帧中所编码的时间信息。Ji et al. [5] 提出执行 3D 卷积来同时计算来自时间维度和空间维度的特征，具体做法是将一个 3D 核卷积到通过连接时间维度中的多个连续帧而形成的立方体上。

3.2 3D 卷积自编码器

输入数据。在大多数用于图像识别的典型 CNN 中，输入数据都是具有 3 个通道（比如 R、G、B 颜色通道）的单张图像。而在异常检测网络中，输入数据是一段包含多帧的视频片段。Hasan et al. [3] 通过使用滑动窗口（滑动窗口的长度为 T）的时间立方体来构建输入。但其中的时间特征很少得到保留。为了解决这个问题，我们以超立方体的形式构建输入——通过在第 4 维（通常被称为时间维）上堆叠 T 帧，然后再在其上执行 3D 卷积。

数据增强。通过在从视频序列中采样的片段上应用多种变换（随机裁剪、亮度变化和高斯模糊），我们可以生成更多输入超立方体。在我们的方法中，我们使用恒定步幅来采样帧，这样目标的运动速度保持不变。

网络架构。图 3 给出了我们提出的时空自编码器网络示意图。

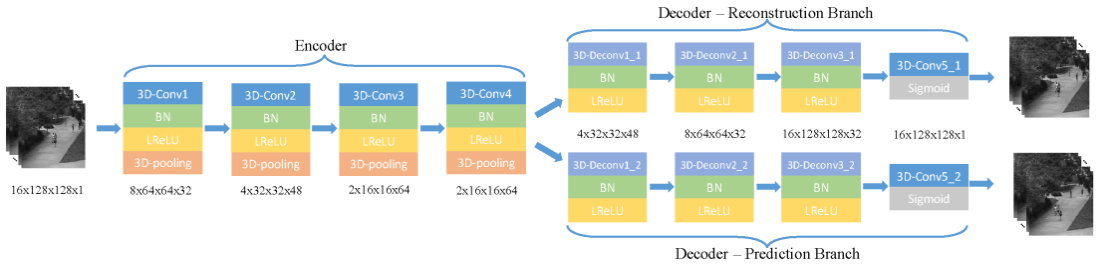

图 3：网络的架构。在编码器之后有两个分支的解码器，分别用于重建过去的帧和预测未来的帧。

3.3 权重递减型预测损失

之前已有研究证明预测网络有助于学习视频表征，受这些研究的启发，我们在解码器部分设计了一个预测分支来预测输入视频片段之后的未来 T 帧。具体来说，重建分支和预测分支具有相同的隐藏特征层，但执行的是不同的任务，分别是：重建过去的序列和预测未来的序列。其中预测任务可以引导模型获取运动目标的轨迹以及让编码器更好地提取时间特征。

在大多数视频异常检测场景中，视点是固定的，各种目标进进出出。新目标的出现难以预测，从而会影响预测网络在训练阶段的收敛性。我们应用了预测损失来增强模型的能力，以提取已有目标的运动特征和预测它们在未来近期的运动，而不会预测相对遥远的未来的新目标的出现。新目标出现的概率会随时间推移逐

渐增大，因此我们在预测得到的视频片段的每一帧上施加了一个递减的权重。

3.4 规律性分数

由常规事件组成的视频序列有更高的规律性分数，因为它们接近于特征空间中的正常训练数据。相反，异常序列的规律性分数更低，因此可以被用于定位异常事件。

4 实验

4.1 数据集

我们在三个数据集上评估了我们提出的时空自编码器，其中包含 UCSD Pedestrian 和 CUHK Avenue 这两个已有的数据集，另外还有新收集的 Traffic 数据集。

Dataset	#Scene	#Train	#Test Nor	#Test Abn
UCSD Pedestrian	2	9350	3569	5641
CUHK Avenue	1	15328	11612	3712
Traffic	5	248543	19784	59562

表 1：异常检测数据集比较。Nor 表示正常帧，Abn 表示异常帧。

4.2 异常的可视化

当我们训练完模型之后，规律性分数可以根据重建误差计算得出。由正常事件组成的视频序列有更低的误差，而异常序列有更高的误差。重建误差是根据每一帧中的每个像素计算得出的，这让我们可以将误差分解到每一帧以及定位图片中的异常区域。

图 4 给出了 5 组来自不同数据集的示例。

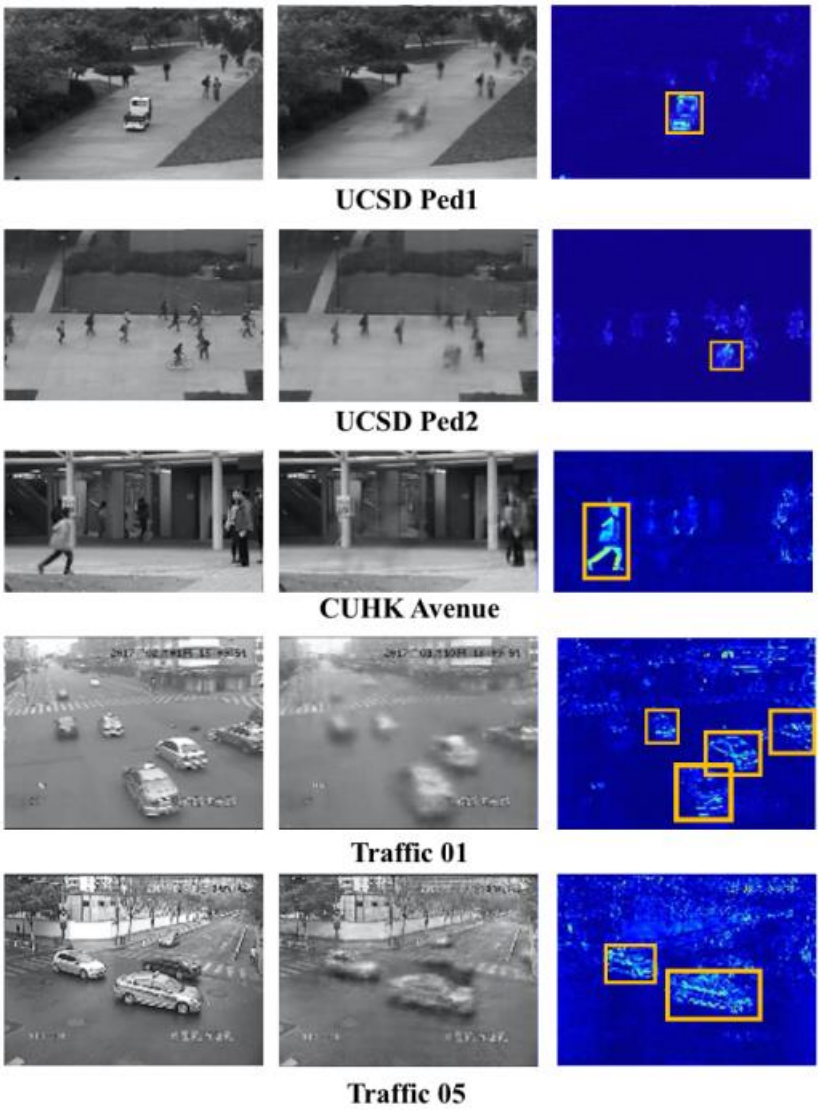

图 4：异常的可视化。左列：来自不规则视频片段的帧。中列：我们的模型的重建输出。右列：重建误差图。橙色矩形突出强调了误差图中的异常区域。在前三个场景中都只有单个目标存在异常，后两个场景则与多个目标有关。

4.3 异常事件检测

基于重建误差可以计算得到规律性分数，而规律性分数又可被进一步用于检测异常事件。如图 5 所示，视频片段的规律性分数会在异常发生时下降。

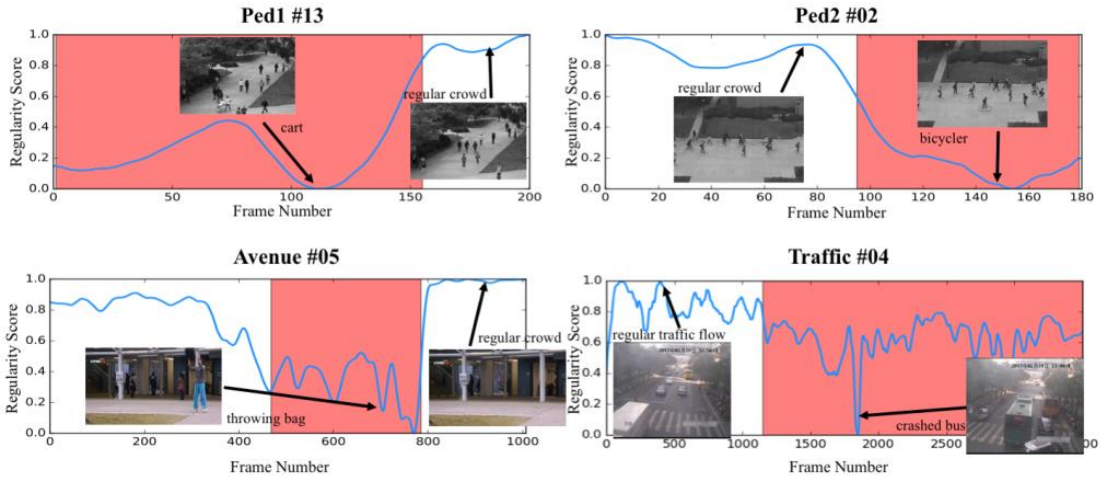

图 5：来自三个数据集的四段测试视频片段的规律性分数曲线。红色区域表示基本真值异常帧。结果表明规律性分数会在异常发生时下降。每个场景都给出了几帧采样，用以展示常规/非常规事件。

表 2 给出了我们的方法与几种当前最佳方法在 UCSD Pedestrian 和 CUHK Avenue 数据集上的表现比较。

	Ped1		Ped2		Avenue	
Algorithm	AUC	EER	AUC	EER	AUC	EER
MPPCA[7]	59.0	40.0	69.3	30.0	-	-
SF[16]	67.5	31.0	55.6	42.0	-	-
SF+MPPCA[14]	66.8	32.0	61.3	36.0	-	-
MDT[14]	81.8	25.0	82.9	25.0	-	-
SCL[12]	91.8	**15.0**	-	-	-	-
AMDN[32]	92.1	16.0	90.8	17.0	-	-
ConvAE[3]	81.0	27.9	90.0	21.7	70.2	25.1
STAE-grayscale	**92.3**	15.3	**91.2**	**16.7**	77.1	33.8
STAE-optflow	87.1	18.3	88.6	20.9	**80.9**	**24.4**

表 2：在 UCSD Pedestrian 和 CUHK Avenue 数据集上的比较

结果还表明我们的时空自编码器模型可用于不同类型的输入数据。

我们还在新收集的 Traffic 数据集上进行了同样的评估。我们将 ConvAE[3] 设为当前最佳方法，因为它有一定的揭示时间特征的能力。表 3 给出了 5 种场景的结果，另外也报告了平均结果。所有被测模型的输入都是灰度帧。

	#01		#02		#03		#04		#05		Average	
Algorithm	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
ConvAE-1frame	54.9	43.3	59.3	44.1	68.7	34.7	66.1	40.1	84.1	22.7	66.6	37.0
ConvAE[3]	57.2	48.6	75.7	28.7	71.2	31.2	71.0	38.3	82.8	24.2	71.5	34.2
STAE-3d-w/o-pred	71.4	36.4	78.7	29.8	77.4	28.3	75.6	34.9	85.8	19.6	77.8	29.8
STAE-3d-constant-pred	73.2	34.8	80.0	27.2	**77.6**	27.7	77.8	29.2	86.1	**19.0**	78.9	27.6
STAE-3d-decreasing-pred	**74.3**	**33.3**	**81.4**	**25.5**	76.6	**27.1**	**79.3**	**26.1**	**86.5**	19.1	**79.6**	**26.2**

表 3：在 Traffic 数据集上的比较

4.4 预测未来帧

如前所述，我们在时空自编码器网络中设计了一个预测分支，以通过跟踪视频序列中运动目标的轨迹来增强视频表征学习的能力。

图 6 给出了两个示例。我们的 STAE 模型可以重建输入的规则视频片段，也能预测未来帧。运动中的车辆（用绿框标出）的轨迹在未来帧中被很好地预测了出来。我们还给出了有新车辆（用红框标出）进入该场景的示例，这表明我们的模型无法预测新出现的目标。

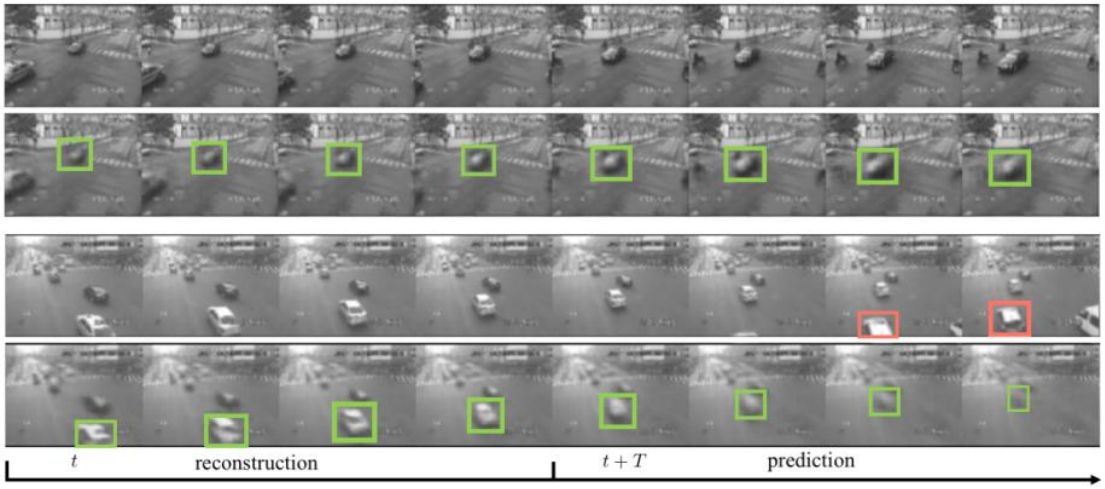

图 6：在 Traffic 数据集上的两组帧预测示例。每一组的上一行都是基本真值视频序列，下一行则是我们的网络重建和预测的输出。左侧部分是从 T 个输入帧中采样的，右侧部分是从未来片段中采样的。运动汽车用绿框标出，新进入场景的汽车用红框标出。

5 结论

未来的研究方向包括研究其它网络架构，融合多模态输入数据（比如 RGB 帧和光流），在实例层面而非像素层面评估规律性分数，以及将我们的框架应用于更复杂的场景。

Deep Siamese Network with Multi-level Similarity Perception for Person Re-identification

基于多层相似度感知的深度神经网络及其在同人鉴别中的应用

论文作者（中英文）：

申晨 Chen Shen/金仲明 Zhongming Jin/赵一儒 Yiru Zhao/付志航 ZhaoZhihang Fu/蒋荣欣 Rongxin Jiang/陈耀武 Yaowu Chen/华先胜 Xian-Sheng Hua

论文原文地址：

https://dl.acm.org/citation.cfm?id=3123452&dl=ACM&coll=DL#URLTOKEN#

Deep Siamese Network with Multi-level Similarity Perception for Person Re-identification

Chen Shen *
Zhejiang University, Alibaba Group
zjushenchen@gmail.com

Zhongming Jin
Alibaba Group
zhongming.jinzm@alibaba-inc.com

Yiru Zhao *
Shanghai Jiao Tong University, Alibaba Group

Zhihang Fu *
Zhejiang University, Alibaba Group
fostor.hunt@gmail.com

Rongxin Jiang †
the State Key Laboratory of Industrial Control Technology, Zhejiang University

Yaowu Chen
Zhejiang Provincial Key Laboratory for Network Multimedia Technologies, Zhejiang University

Xian-Sheng Hua †
Alibaba Group
huaxiansheng@gmail.com

摘要

行人重识别（person re-ID）的目的是识别多个摄像头视角中的相关行人，这项任务在计算机视觉社区中已经得到了越来越多的关注。我们在本论文中提出了一种基于卷积神经网络（CNN）和多级相似度感知的全新深度孪生架构。根据不同特征图的不同特性，我们有效地在训练阶段将不同的相似度约束应用到了低层级和高层级特征图上。因此，我们的网络可以有效地学习不同层级的有判别性的（discriminative）特征表征，这能显著提升 re-ID 的表现。此外，我们的框架还有另外两个优势。第一，可以轻松地将分类约束整合到该框架中，从而形成一个带有相似度约束的统一的多任务网络。第二，因为相似度的信息已经通过反向传播被编码在了该网络的学习参数中，所以在测试时并不必需成对的输入。这就意味着我们可以提取每张图库图像的特征并以一种离线的方式来构建索引，这对大规模真实世界应用而言至关重要。我们在多个有挑战性的基准上进行了实验，结果表明我们的方法相比于当前最佳方法表现出色。

1 引言

行人重识别（person re-ID）的目的是匹配一个行人在多个无交集的摄像头视角中的图像，这项任务凭借其研究和应用价值正获得越来越大的关注。但是，行人重识别仍然是一项非常具有挑战性的任务，因为不同身份实体之间的外观可能差异不大（见图 1(a)），而同一身份实体在不同光照、视角和部分遮挡（见图 1(b)、1(c)、1(d)）情况下又可能差异很大。

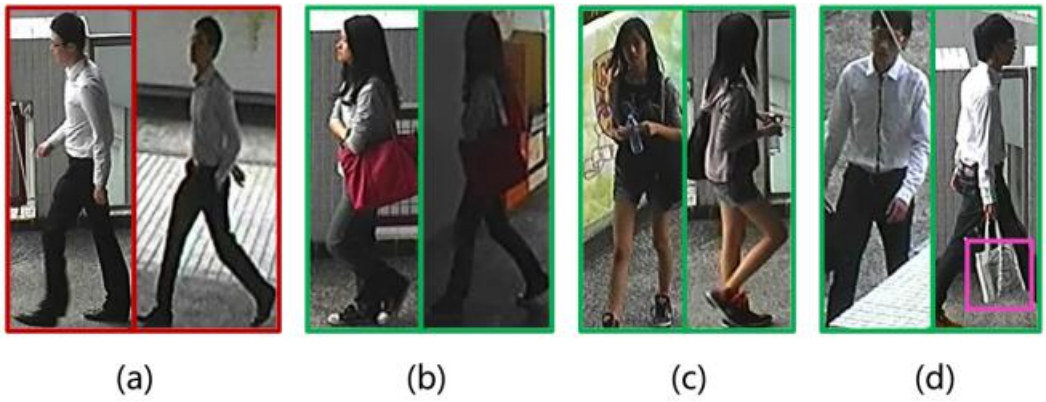

图 1：行人重识别的各种复杂性示意图，来自 CUHK03 数据集的。绿框表示同一个身份，而红框则表示不同的身份。(d) 中的粉色框标示了一个突出的局部图案（手提袋），由于部分遮挡这很容易丢失。

从技术上讲，行人重识别有两大基本组成：特征表征和距离度量。最近，基于 CNN 的深度学习方法已经在行人重识别上表现出了出色的优越性，因为它能够联合学习复杂的特征表征和可区分的距离度量。

在本论文中，我们提出了一种全新的基于 CNN 的行人重识别方案，称为多级相似度感知卷积神经网络（MSP-CNN）。在训练阶段，我们会使用一种孪生模型（Siamese model），其使用图像对作为输入，并且所有图像都要经过同样的共享参数的深度 CNN 网络的处理。该基准网络是精心设计的，其中使用了非常小的卷积过滤器和 Inception 模块。接下来，我们深入思考了如何有效地将相似度约束应用到不同的特征图上。

图 2 给出了我们提出的网络在训练阶段的整体架构。图 3 给出了该网络在测试阶段的整体架构。

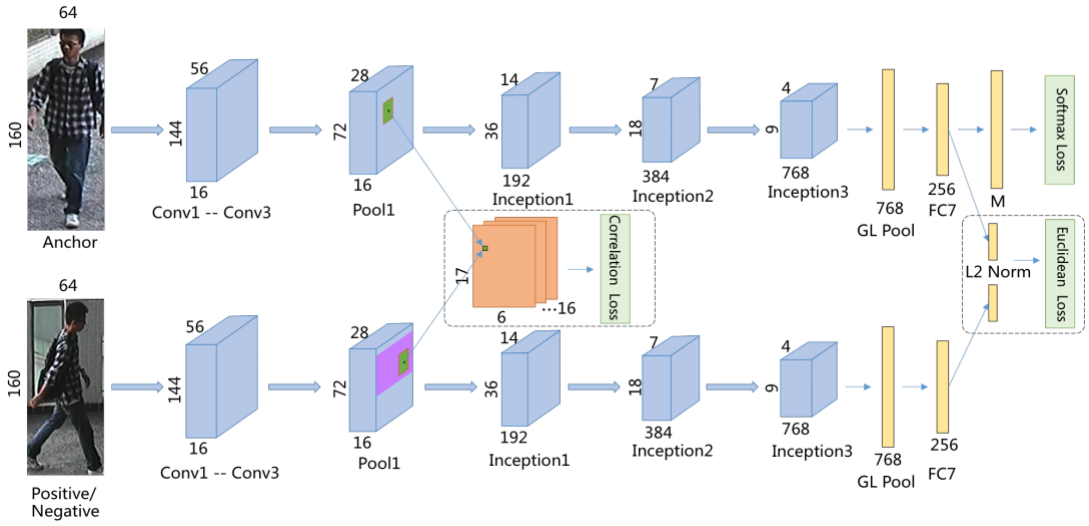

图 2：训练阶段的多任务框架示意图。具体说明一下，我们在低层级的 Pool1 层和高层级的 FC7 层分别优化相似度约束。正例（或负例）图像的 Pool1 层特征图上的紫色区域表示在获取局部形义模式时互相关所使用的宽搜索区域。另外也同时使用了 softmax 损失来优化分类约束，M 表示行人身份实体的数量。

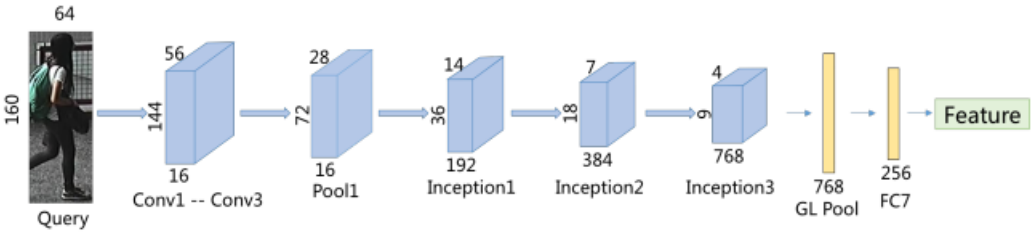

图 3：测试时间的网络架构

因此，我们的工作有三大关键优势和主要贡献。

- 我们提出了一种用于行人重识别的全新孪生模型，并且创新地在不同的特征图上应用了相应的距离度量。这种多级相似度感知机制能巧妙地匹配不同层级特征图的特性并显著提升表现。
- 我们使用了一种多任务架构来同时优化分类约束和相似度约束。多任务学习可以在解决多个相关任务的同时实现知识共享，从而将两者的优势组合到一起。
- 在测试时间，我们可以避免成对输入的时间低效的流程并且可以提取图像特征来事先构建索引，这对于大规模真实世界应用场景而言至关重要。

2 相关工作（略）

3 我们提出的方法

3.1 方法概述

受 [42] 的启发，我们首先仔细设计了一个用于行人重识别的基本深度 CNN 网络，并期望它能仅使用单个 softmax 损失就得到优于大多数已有深度学习框架的强大基准结果。为了适应大多数行人图像的尺寸（通常很小而且不是正方形的），所有的输入图像都重新调整为 160×64 大小，并且为了数据增强而随机裁剪为 144×56 大小。

然后，我们从一种互补的角度考虑了相似度约束，并构建了一种分类任务的多任务架构。这种设计的目的是兼取二者之长，即充分利用行人重识别标注以及正例负例对之间相似度相当的信息。为了利用不同层级的特征图的相关性信息来更好地描述相似度约束（之前的大多数研究都忽略了这一点），我们可视化了我们的基本 CNN 分类网络所学习到的某些典型层的特征图。

如图 4 所示，低层级特征图的响应通常很密集并且反映了局部形义区域。比如，来自 Conv1 层 #0 通道的特征会强烈响应黑色区域（头发和裤子），而来自 Conv2 层 #9 通道的特征则重点强调明亮的白色区域（短袖衫）。这种现象也可以根据 Pool1 层的特征图进行验证。随着层越来越深，它们的特征图也会逐渐变得稀疏，而且往往会编码更加抽象的全局特征。比如，Inception(1a) 层的某些通道仍然反映的是局部形义区域（红色背包，#11），但大部分通道反映的都非常稀疏（#91）。其内部机制是：低层级卷积层所得到的可区分的局部特征会传播给高层级层（尤其是全连接层），这些特征会变得抽象并形成全局表征。

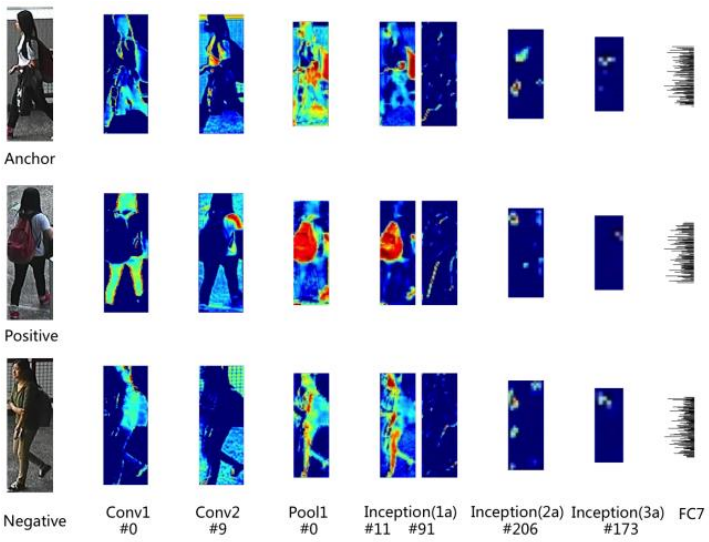

图 4：我们的基本 CNN 分类网络学习到的特征图的可视化。每一行都表示一个人的图像从低层级到高层级的某些典型特征图。第一行：锚图像；第二行：正例图像；最后一行：负例图像。我们也标注了每个特征图对应的通道号，用 # 号标记。

就低层级特征图而言，典型的局部区域（比如图 4 中的红色背包）在同一个人的图像中都有，这对区分正例对和负例对而言至关重要。因此，我们假设，对于第一张图像的特征图上的特定图块（patch），如果我们可以从另一张正例图像的对应特征图上找到其最相似的图块，那么这样的图块对就非常有可能表示了可区分的局部区域。根据以上假设，我们可以自然而然地设计出寻找和强调这个可区分的区域的目标，这样我们的网络就能向更高层传播更相关的特征。受 [35] 的启发，我们采用了归一化互相关（normalized cross-correlation）作为一种非精确的匹配技术来匹配广大区域中的像素区域。[35] 中已经证明，使用归一化互相关和在更大区域上进行搜索能在视角变化很大、光照变化或部分遮挡的情况中保持稳健（robust）。归一化互相关分数取值范围为 [-1,1]，其中 -1 表示特征向量完全不相似，1 表示特征向量非常相似。因此，对于第一个特征图的每个局部图块，我们都要通过选择图块来找到第二个特征图中与其最相似的图块——该图块有最大的互相关响应。之后，我们设计了损失函数，其目的是根据正例对（即可区分的区域）来增强互相关分数，同时根据负例对（即某种扰动）减弱互相关分数。因此，我们的设计可以适应性地更加重点关注正例对之间共有的局部语义可区分区域，并且能够通过前向传播沿更高层的方向放大这种局部相似度。应当指出，正例图像和负例图像之间也有一些共有的语义模式（比如黑发），但这些模式不能看作是可区分的信息。因此，对于负例对而言，我们会忽略这样的情况。

至于高层级特征图，尤其是全连接层的特征图，我们直接在 L2 归一化之后使用欧几里德距离来表示它们的相似度，并设计了用来降低正例对之间的距离并增加负例对之间的距离的损失函数。

3.2 多级相似度感知

低层级相似度。我们在 Pool1 层的特征图上应用了低层级相似度约束，如图 2 所示。

高层级相似度。高层级相似度约束（即优化欧式距离）应用在最后的全连接层的特征上（即图 2 中的 FC7 层）。

3.3 多任务网络架构

联合训练。前面已经提到，我们提出了一种全新的孪生网络，可以在训练阶段将不同的相似度约束应用到对应的特征图上。此外，我们将相似度约束和分类约束结合到一起构建了一个统一的多任务网络。

如图 2 所示，低层级和高层级相似度约束分别应用在 Pool1 层和 FC7 层上。这个选择也是由验证集决定的。

我们提出的训练 MSP-CNN 的流程分为两个阶段。我们首先使用 softmax 损失和欧式距离损失在对应的数据集上从头开始训练一个精心设计的 CNN 多任务网络。然后，我们加入低层级的互相关损失并继续训练该 CNN 几个 epoch。因为低层级层的梯度通常很小，所以直接为低层级层提供梯度的互相关损失应该在相对稳定的阶段得到更好的准确优化。此外，互相关损失收敛速度很快，所以我们只需要优化它少数几个 epoch 来防止过拟合即可。

测试。前面已经提到，每张图库图像的特征都可以按照图 3 给出的过程事先提取出来。当发生查询时，图库中的图像根据它们与探针图像（probe image）的相似度进行排序，其中图像特征之间的相似度是根据欧式距离计算的。我们甚至可以利用某些索引技术（比如倒排索引或哈希）来基于这些图像特征构建索引，从而进一步提升检索效率（尤其是对于大规模数据集）。

4 实验

4.1 数据集和协议

我们在大数据集 CUHK03、Market-1501 和小数据集 CUHK01 上进行了实验。在我们的实验中，我们使用了最常用的累积匹配特征（CMC）top-k 准确度来评估所有方法。我们还为 Market-1501 数据集使用了平均精度均值（mAP）。所有的评估结果都是单次查询的结果。

4.2 实现细节

受 [42] 的启发，我们精心设计了一个基本 CNN 网络，它主要由 3 个 CONV 模块、6 个 Inception 模块和 1 个 FC 模块组成。CONV 模块中有一个卷积（conv）层，后面跟着批归一化（BN）层和 ReLU 层。conv 层使用了一个非常小的过滤器（3×3）（受 VGGNet 的启发），BN 层的作用是加快收敛速度。Inception 模块是指 GoogLeNet，而且我们将每个 5×5 卷积都替换成了 2 个 3×3 卷积，就像 Inception-v3 建议的那样。FC 模块由一个全连接层以及后续的 BN 层、ReLU 层和 dropout 层构成。表 1 给出了详细结构。

name	channel num	output size	stride
Input	3	144 × 56	-
Conv11-conv3	16	144 × 56	1
Pool1	16	72 × 28	2
Inception(1a)	128	72 × 28	1
Inception(1b)	192	36 × 14	2
Inception(2a)	256	36 × 14	1
Inception(2b)	384	18 × 7	2
Inception(3a)	512	18 × 7	1
Inception(3b)	768	9 × 4	2
Global_Pool	768	-	-
FC7	256	-	-

表 1：基本网络结构

我们的算法是基于深度学习框架 Caffe 实现的，运行在配置了一块英伟达 M40 GPU 卡的工作站上。

4.3 训练策略

我们设计了一种训练阶段的采样策略，让负例对的数量和正例对的数量之比为 2:1。如图 5 所示。

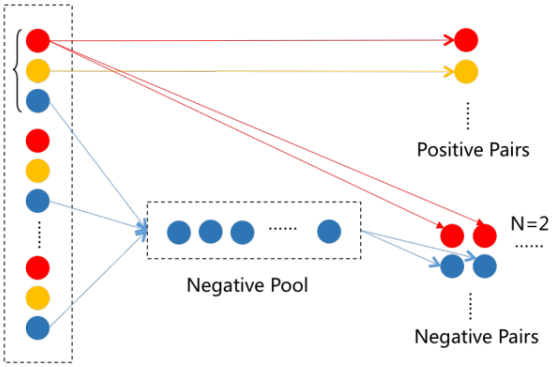

图 5：采样过程示意图。红色圆圈表示锚图像，橙色圆圈表示正例图像，蓝色圆圈表示负例图像。

数据增强。我们遵循了 AlexNet [15] 提出的经典技术。

5 结果和讨论

5.1 与当前最佳方法的比较

(a) Manually labeled setting.

Method	rank-1	rank-5	rank-10	rank-20
LDML [10]	13.5	-	52.1	70.8
KISSME [14]	14.2	-	52.6	70.0
FPNN [18]	20.7	51.3	68.7	83.1
LOMO+XQDA [19]	52.2	82.2	92.1	96.3
Ahmed *et al.* [1]	54.7	86.5	93.9	98.1
SS-SVM [47]	57.0	84.8	92.5	96.4
MLAPG [20]	58.0	-	94.7	98.0
DNS [46]	58.9	85.6	92.5	96.3
Ensembles [26]	62.1	89.1	94.3	97.8
Fused Model [35]	72.4	91.0	95.5	98.4
WARCA [13]	78.4	94.6	97.5	99.1
MSP-CNN (Ours)	**85.7**	**97.6**	**99.2**	**99.8**

(b) Detected setting.

Method	rank-1	rank-5	rank-10	rank-20
LDML [10]	10.9	32.3	47.0	65.0
KISSME [14]	11.7	33.3	48.1	64.9
FPNN [18]	19.9	49.3	64.8	81.1
LOMO+XQDA [19]	46.3	78.9	88.6	94.3
Ahmed *et al.* [1]	45.0	76.0	83.5	93.2
SS-SVM [47]	51.2	81.5	89.9	95.0
MLAPG [20]	51.2	83.6	92.1	96.9
DNS [46]	53.7	83.1	93.0	94.8
Gated S-CNN [38]	61.8	80.9	88.3	-
Fused Model [35]	72.0	90.5	96.0	98.3
MSP-CNN (Ours)	**83.6**	**96.9**	**98.4**	**99.0**

表 2：在 CUHK03 数据集上 CMC rank 分别为 1、5、10、20 时当前最佳方法的表现比较。（a）人工标注人类框的设置。（b）检测得到人类框的设置。

(a) 100 identities for testing.

Method	rank-1	rank-5	rank-10	rank-20
LDML [10]	26.5	-	72.0	84.7
FPNN [18]	27.9	58.2	73.5	86.3
KISSME [14]	29.4	-	72.4	86.1
Ahmed *et al.* [1]	65.0	88.7	93.1	97.2
Deep Metric [31]	69.4	90.8	96.0	-
SIR-CIR [39]	71.8	91.6	96.0	98.0
Fused Model [35]	**81.2**	95.1	97.4	98.6
MSP-CNN (Ours)	79.3	**95.2**	**97.6**	**98.9**

(b) 486 identities for testing.

Method	rank-1	rank-5	rank-10	rank-20
Semantic [32]	31.5	52.5	65.8	77.6
MirrorRep [6]	40.4	64.6	75.3	84.1
Ahmed *et al.* [1]	47.5	71.6	80.3	87.5
DNS [46]	65.0	85.0	89.9	94.4
Fused Model [35]	65.0	83.9	89.8	94.5
WARCA [13]	65.6	85.3	90.5	95.0
SS-SVM [47]	**66.0**	**89.1**	**92.8**	96.5
MSP-CNN (Ours)	63.7	87.9	92.4	**96.9**

表 3：在 CUHK01 数据集上 CMC rank 分别为 1、5、10、20 时当前最佳方法的表现比较。（a）测试中有 100 个身份。（b）测试中有 486 个身份。

Method	rank-1	rank-5	rank-10	mAP
LOMO+XQDA [19]	26.1	-	-	7.8
BoW [50]	35.8	52.4	60.3	14.8
WARCA [13]	45.2	68.2	76.0	-
DNS [46]	55.4	-	-	29.9
Gated S-CNN [38]	65.9	-	-	39.6
MSP-CNN (Ours)	**81.9**	**92.8**	**95.2**	**63.6**

表 4：在 Market-1501 数据集上当前最佳方法的表现比较。

5.2 算法组成部分的有效性

以 CUHK03 有标注数据集为例，我们还详细研究了我们提出的算法中各个模块的效果，包括单独的基本分类深度网络、与高层级或低层级相似度约束相结合的分类约束、以及上述三者的综合。表 5 给出了结果。

Combination Method	rank-1
cls alone	78.9
cls + sim_high	84.2
cls + sim_low	82.1
cls + sim_high + sim_low	**85.7**

表 5：在 CUHK03 有标注数据集上，算法的不同组成部分以及它们的组合所得到的表现比较。cls 是指分类约束（即 softmax 损失），sim_high 是指高层级相似度约束（即欧式距离损失），sim_low 是指低层级相似度约束（即归一化互相关损失）。

6 结论和未来工作

对于未来，我们打算寻找一个用于中层级层的合适优化目标并探索利用更多层特征图的效果。

Stylized Adversarial Autoencoder for Image Generation

基于风格化对抗自编码器的图像生成算法

论文作者（中英文）：
赵一儒 Yiru Zhao/邓兵 Bing Deng/黄建强 Jianqiang Huang/卢宏涛 Hongtao Lu/华先胜 Xian-Sheng Hua

论文原文地址：https://dl.acm.org/citation.cfm?id=3123450

Stylized Adversarial AutoEncoder for Image Generation

Yiru Zhao*
Shanghai Jiao Tong University
Alibaba Group
yiru.zhao@sjtu.edu.cn

Bing Deng
Alibaba Group
dengbing.db@alibaba-inc.com

Jianqiang Huang
Alibaba Group
jianqiang.hjq@alibaba-inc.com

Hongtao Lu†
Shanghai Jiao Tong University
htlu@sjtu.edu.cn

Xian-Sheng Hua‡
Alibaba Group
huaxiansheng@gmail.com

摘要

在本论文中，我们提出了一种用于自动图像生成的基于自编码器的生成对抗网络（GAN），我们称之为“风格化对抗式自编码器”。不同于已有的生成式自编码器（通常会在隐向量上施加一个先验分布），我们提出的方法是将隐变量分成两个分量：风格特征和内容特征，这两个分量都是根据真实图像编码的。这种隐向量的划分让我们可以通过选择不同的示例图像来任意调整所生成图像的内容和风格。此外，这个 GAN 网络中还采用了一个多类分类器来作为鉴别器，这能使生成的图像更具真实感。我们在手写数字、场景字符和人脸数据集上进行了实验，结果表明风格化对抗式自编码器能实现优异的图像生成结果，并能显著改善对应的监督识别任务。

1 引言

生成式自然图像建模是计算机视觉和机器学习领域的一个基本研究问题。早期的研究更关注生成网络建模的统计原理，但由于缺乏有效的特征表征方法，相应结果都局限于某些特定的模式。深度神经网络已经展现出了在学习表征方面的显著优势，并且已经被证明可有效应用于鉴别式视觉任务（比如图像分类和目标检测），与贝叶斯推理或对抗训练一起催生出了一系列深度生成模型。

研究表明，正则化神经网络的在实际工作中的表现通常优于无约束的网络。常用的正则化形式包括 L1 范数 LASSO、L2 范数岭回归（ridge regression）以及 dropout 等一些现代技术。尤其是对于自编码器神经网络，研究者近期已经提出了相当多的正则化方法。但是，所有这些正则化方法都会在隐变量（也被称为隐藏节点）上施加一个先验分布，经常使用的是高斯分布。对于相对简单的生成任务（比如建模灰度数字图像）而言，这种方法效果很好，但却不适合用于生成彩色字母数字图像或人脸等复杂图像，因为这些图像的隐变量的真实分布是不可见的，也无法用简单的模型进行建模。

如图 1 所示，我们在本论文中提出了一种名为风格化对抗式自编码器（SAAE）

的全新生成模型，该模型是使用一种对抗式方式来训练风格化自编码器。不同于已有的自编码器，我们会将隐向量分成两部分，一部分与图像内容有关，另一部分与图像风格有关。内容特征和风格特征都是根据示例图像编码的，并且不会在隐变量的分布上使用任何先验假设。带有给定内容和风格的目标图像可以根据组合起来的隐变量解码得到，这意味着我们可以通过选择不同的示例内容和/或风格图像来调整输出图像。此外，受 [9, 21, 26] 中方法的启发，我们在 模型训练阶段采用了一种对抗式的方法。我们的 GAN 网络没有使用典型的二元分类器作为鉴别器，而是使用了一个多类分类器，它在鉴别真实图像和虚假图像时能更好地建模生成图像的变化情况。此外，由于 GAN 模型训练是博弈形式的最小-最大目标，所以非常难以收敛，因此我们根据经验开发出了一种有效的三步式训练算法，可以改善我们提出的 GAN 网络的收敛表现。

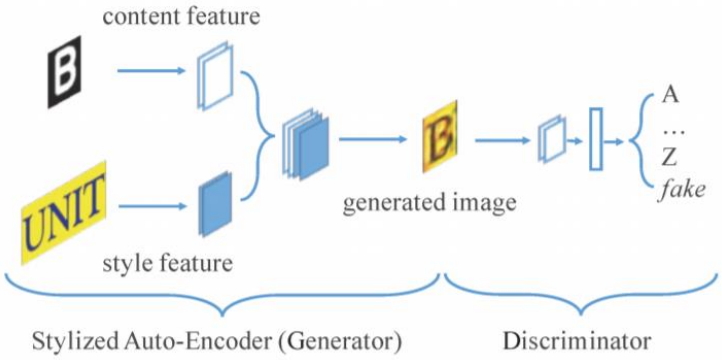

图 1：我们的模型的图示，其分别从内容图像和风格图像中提取特征，然后再将这些特征融合起来，解码得到目标图像。多类鉴别器会迫使生成的图像更具真实感。

本工作的主要贡献可以总结为：

- 我们提出了一种全新的深度自编码器网络，它可以分别编码来自两个示例图像的内容特征和风格特征并根据这两个特征解码得到新图像。
- 使用了多类分类器作为鉴别器，这能更好地建模生成的图像的变化情况，并能有效地迫使生成网络生成更具真实感的结果。
- 我们开发了一种三步式训练策略，以确保我们提出的风格化对抗式自编码器的收敛。

2 相关工作（略）

3 风格化对抗式自编码器

为方便起见，我们将使用文本字符图像生成（比如场景文本生成等）作为背景应用来介绍我们的算法，但我们还会在实验部分展示更多应用（比如人脸生成）。我们的目标是通过定义和训练一个神经网络，根据两张示例图像（内容图像 c 和风格图像 s）来生成图像。就字符图像生成而言，内容图像是指没有任何风格或纹理或背景的合成字符图像，比如 A 到 Z，0 到 9；风格图像是一张示例图像，比如是一张真实的单词图像。

正如前面提到的，我们将揭示了真实数据的先验分布的隐变量分成两个部分：风格特征和内容特征。内容特征是从内容图像中导出的（通过一个卷积网络），而风格特征是从风格图像中导出的。

3.1 生成器

生成网络由两个编码器（Enc_c 和 Enc_s）和一个解码器（Dec）构成。其中 Enc_c 将内容图像编码成内容隐含表征或特征 z_c，Enc_s 将风格图像编码成风格隐含表征或特征 z_s。Dec 解码组合后的隐含表征并得到输出图像。为了方便起见，我们使用生成器 G 表示 Enc_c、Enc_s 和 Dec 的组合。

3.2 鉴别器

已有 GAN 中的鉴别器的输出是表示该输出 x 是真实图像的概率 $y = Dis(x) \in [0,1]$。而鉴别器 D 的训练目标是最小化二元交叉熵：$L_{dis} = -\log(Dis(x)) - \log(1 - Dis(G(z)))$。

G 的目标是生成 D 无法将其与真实图像区分开的图像，即最大化 L_{dis}。前面已经提到，已有的 GAN 网络在 D 中使用二元分类器来确定图像是真实图像还是生成图像。但是，将所有真实图像放入一个大型的正例类别中将无法利用这些训练图像的内在形义结构。因此，我们提出使用多类分类器作为鉴别器，该分类器将确定输入是生成图像还是属于某个特定的真实图像类别（比如特定字符）。

3.3 网络架构

卷积神经网络（CNN）已经在特征表征和图像生成方面展现出了巨大的优势，我们提出的 SAAE 网络就基于 CNN 架构，如图 2 所示。

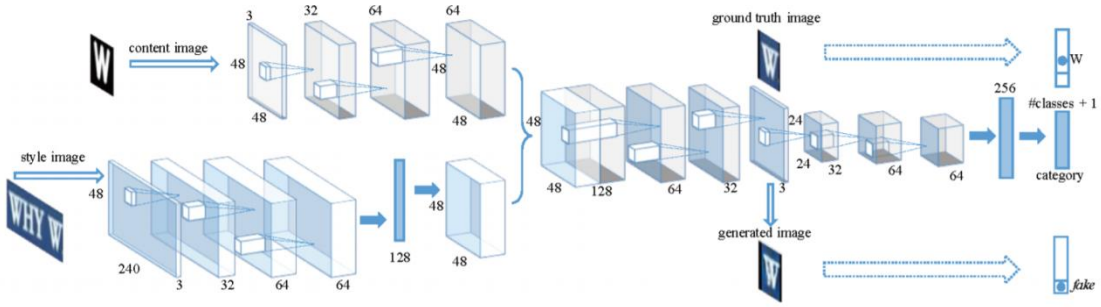

图 2：网络架构

实际上，我们提出的生成网络包含两个特征提取网络流程，之后再跟上一个生成网络。内容特征提取器和风格特征提取器都有三个无下采样的卷积层，这样能尽可能多地保留示例图像的细节信息。输入的风格图像和内容图像可能有不同的尺寸。比如，当生成场景文本字符图像时，内容图像是含有一个字符的图像，风格图像是含有一个词或多个字符的图像。在三个卷积层之后，风格特征图（feature map）的形状会由一个全连接层重新调整为风格特征向量。为了与根据内容图像解码得到的内容特征图拼接到一起，风格特征向量需要重新调整回一个特征图，且该特征图与内容特征图具有一样的尺寸。内容特征提取网络没有任何全连接层，因为内容图像的二维空间信息需要保留。我们在通道维度中合并内容特征图和风格特征图，这意味着组合后的特征图有一半通道来自内容特征，另一半则来自风格特征。之后，生成网络使用三个卷积层将组合后的特征图解码成一张目标字符图像。

鉴别网络是一个常见的 CNN 分类器，包含三个卷积层，其中第一个卷积层后有一个 2×2 最大池化层，最后一个卷积层后有两个全连接层。鉴别器的输出

层是一个 (k+1) 维的向量，表示输入图像属于每个类别的概率（真实图像有 k 类，虚假图像占 1 类）。

我们在每个卷积层上都应用了批归一化，这能加快训练阶段的收敛速度。除最后一层之外的每一层都使用了 Leaky ReLU，最后一层使用了 sigmoid 来将每个输出投射到 [0,1] 区间中（作为概率）。

3.4 训练策略

受 [30] 中所用的分步训练的启发，我们提出了一种三步式训练策略来优化我们的模型。这个三步式优化策略能帮助我们得到稳定的训练结果。

4 实验

我们使用 4 种不同的方法评估了我们的方法：在 MNIST 数据集上计算对数似然以衡量 SAAE 模型拟合数据分布的能力；在人脸生成任务上展示视觉属性迁移；在场景文本数据集上评估 SAAE 模型；为监督识别任务生成训练数据。

4.1 对数似然分析

受 [9,26] 中评估流程的启发，我们评估了作为生成模型的 SAAE 拟合数据分布的表现，具体方法是计算生成图像的估计分布与 MNSIT 测试集分布的对数似然。

表 1 比较了 SAAE 与六种当前最佳方法的对数似然结果。我们的方法在这一标准上表现最优，超过 AAE 大约 89。

Models	Log-likelihood
DBN [11]	138 ± 2
Stacked CAE [3]	121 ± 1.6
Deep GSN [2]	214 ± 1.1
GAN [9]	225 ± 2
GMMN + AE [24]	282 ± 2
AAE [26]	340 ± 2
SAAE-binary	402 ± 2.2
SAAE	**429 ± 2.5**

表 1：测试数据在 MNIST 数据集上的对数似然。值越高越好。最后两行结果来自我们的方法，分别使用了二元鉴别器和多类鉴别器。这里报告的数值是样本在测试集上的平均对数似然以及在多次实验结果计算得到的均值的标准误差。

遵照之前的方法，我们在图 3 中展示了一些来自训练后的 SAAE 生成器的样本。最后一列是与倒数第二列的生成图像最接近的训练图像（用像素级别的欧氏距离来度量），以证明 SAAE 模型没有单纯地记忆训练集。

图 3：我们的 SAAE 模型生成的样本示例

4.2 基于属性条件的人脸生成

我们在 Labeled Faces in the Wild（LFW）数据集上评估了我们的模型在人脸图像生成任务上的表现。

如图 4 所示，生成的样本在视觉上与属性迁移一致。比如，如果改变“眼镜”这样的属性，整体外观仍然能保存完好，但眼部区域会出现差异。

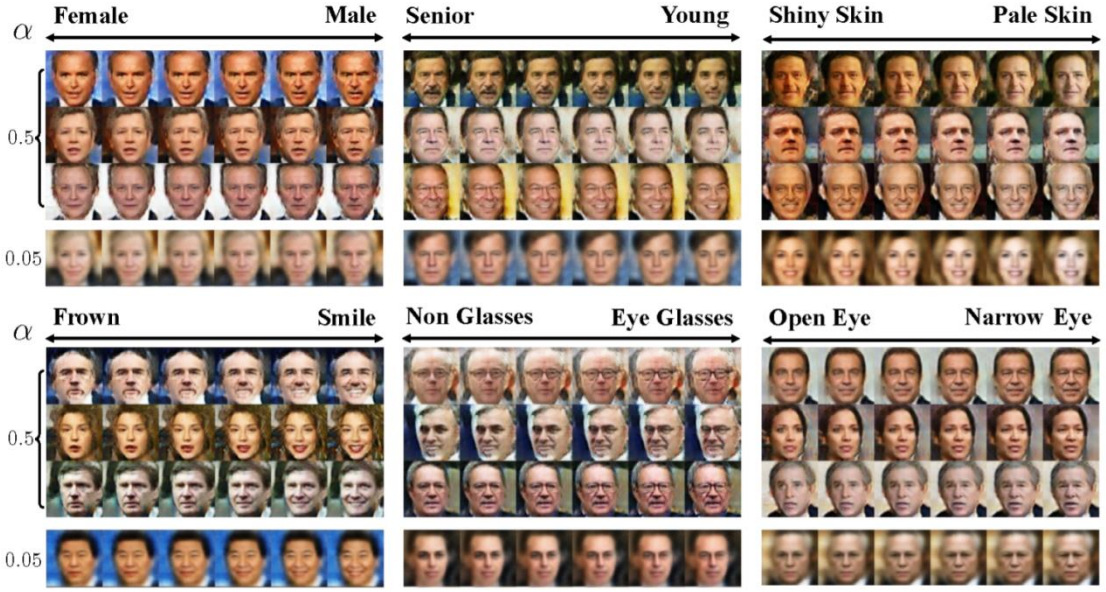

图 4：基于属性条件的图像生成，分成六个组（性别、年龄、肤色、表情、眼镜和眼睛大小）。

4.3 模型样本

我们在 IIIT 5k-word（IIIT5K）数据集和中国汽车牌照（PLATE）数据集上评估了我们的 SAAE 模型。

图 5 展示了我们的模型和 DCGAN 模型生成的图像中随机取出的样本，同时也给出训练数据以便比较。SAAE 生成的样本看起来更像字符而且有更清晰的

边缘和背景。

图 5：SAAE 和 DCGAN 的训练数据和模型样本。上行：IIIT5K 数据集，下行：PLATE 数据集

为了可视化我们的风格化对抗式自编码器的风格化属性，我们在图 6 中展示了几组生成样本，IIIT5K 和 PLATE 数据集上的都有。在每个数据集中，我们都选择了一张示例风格图像并遍历了所有的内容图像和标签。结果表明，SAAE 模型可以将示例风格图像的字符风格迁移给内容图像。

图 6：给定一张风格图像而生成的样本。上行：IIIT5K 数据集。下行：PLATE 数据集。对于每组生成样本，风格图像在左上角给出，用红色方框标出。对于 PLATE 数据集，我们因为隐私原因隐藏了汽车牌照的第一个汉语字符。

4.4 用于监督学习的数据生成

深度神经网络（DNN）已经在监督学习方面表现出了显著的优越性，但它却依赖于大规模有标注训练数据。在小规模训练数据上，深度模型很容易过拟合。我们还使用 SAAE 模型为识别中国汽车牌照任务生成了训练数据。

我们通过测量在 DR-PLATE 数据集上的识别准确度而对数据生成的质量进行了评估。图 7 表明加入到训练数据集中的生成数据越多，模型收敛得越慢，但分类准确度却越来越好。这个结果表明我们的 SAAE 模型能够通过生成数据提升监督学习的表现。

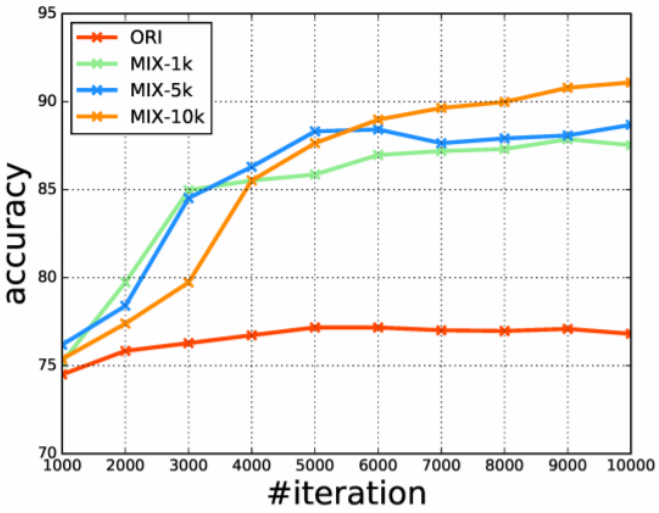

图 7：在不同训练集上对应迭代次数的识别准确度

5 结论

未来的研究重点是优化网络结构，以实现更高的生成质量。将这一框架扩展到其它应用领域（比如半监督特征学习）也会是一个有趣的研究方向。