深度学习在图像识别中的研究进展与展望

王晓刚

香港中文大学电子工程系

深度学习是近十年来人工智能领域取得的最重要的突破之一。它在语音识别、自然语言处理、计算机视觉、图像与视频分析、多媒体等诸多领域都取得了巨大成功。本文将重点介绍深度学习在物体识别、物体检测、视频分析的最新研究进展，并探讨其发展趋势。

1. 深度学习发展历史的回顾

现有的深度学习模型属于神经网络。神经网络的历史可追述到上世界四十年代，曾经在八九十年代流行。神经网络试图通过模拟大脑认知的机理，解决各种机器学习的问题。1986 年 Rumelhart，Hinton 和 Williams 在《自然》发表了著名的反向传播算法用于训练神经网络[1]，直到今天仍被广泛应用。

但是后来由于种种原因，大多数学者在相当长的一段的时间内放弃了神经网络。神经网络有大量的参数，经常发生过拟合问题，即往往在训练集上准确率很高，而在测试集上效果差。这部分归因于当时的训练数据集规模都较小。而且计算资源有限，即便是训练一个较小的网络也需要很长的时间。总体而言，神经网络与其它模型相比并未在识别的准确率上体现出明显的优势，而且难于训练。

因此更多的学者开始采用诸如支持向量机、Boosting、最近邻等分类器。这些分类器可以用具有一个或两个隐含层的神经网络模拟，因此被称作浅层机器学习模型。它们不再模拟大脑的认知机理；相反，针对不同的任务设计不同的系统，并采用不同的手工设计的特征。例如语音识别采用高斯混合模型和隐马尔可夫模型，物体识别采用 SIFT 特征，人脸识别采用 LBP 特征，行人检测采用 HOG 特征。

2006 年，Geoffrey Hinton 提出了深度学习。之后深度学习在诸多领域取得了巨大成功，受到广泛关注。神经网络能够重新焕发青春的原因有几个方面。首先是大数据的出现在很大程度上缓解了训练过拟合的问题。例如 ImageNet[2]训练集拥有上百万有标注的图像。计算机硬件的飞速发展提供了强大的计算能力，使得训练大规模神经网络成为可能。一片 GPU 可以集成上千个核。此外神经网络的模型设计和训练方法都取得了长足的进步。例如，为了改进神经网络的训练，学者提出了非监督和逐层的预训练。它使得在利用反向传播对网络进行全局优化之前，网络参数能达到一个好的起始点，从而训练完成时能达到一个较好的局部极小点。

深度学习在计算机视觉领域最具影响力的突破发生在 2012 年，Hinton 的研究小组采用深度学习赢得了 ImageNet [2] 图像分类的比赛[3]。ImageNet 是当今计算机视觉领域最具影响力的比赛之一。它的训练和测试样本都来自于互联网图片。训练样本超过百万，任务是将测试样本分成 1000 类。自 2009 年，包括工业界在内的很多计算机视觉小组都参加了每年一度的比赛，各个小组的方法逐渐趋同。在 2012 年的比赛中，排名 2 到 4 位的小组都采用的是传统的计算机视觉方法，手工设计的特征，他们准确率的差别不超过 1%。Hinton 的研究小组是首次参加比赛，深度学习比第二名超出了 10%以上。这个结果在计算机视觉领域产生了极大的震动，掀起了深度学习的热潮。

计算机视觉领域另一个重要的挑战是人脸识别。Labeled Faces in the Wild (LFW) [4] 是当今最著名的人脸识别测试集，创建于 2007 年。在此之前，人脸识别测试集大多采集于实验

室可控的条件下。LFW 从互联网收集了五千多个名人的人脸照片，用于评估人脸识别算法在非可控条件下的性能。这些照片往往具有复杂的光线、表情、姿态、年龄和遮挡等方面的变化。LFW 的测试集包含了 6000 对人脸图像。其中 3000 对是正样本，每对的两张图像属于同一个人；剩下 3000 对是负样本，每对的两张图像属于不同的人。随机猜的准确率是 50%。有研究表明[5]，如果只把不包括头发在内的人脸的中心区域给人看，人眼在 LFW 测试集上的识别率是 97.53%。如果把整张图像，包括背景和头发给人看，人眼的识别率是 99.15%。经典的人脸识别算法 Eigenface [6] 在这个测试集上只有 60%的识别率。在非深度学习的算法中，最好的识别率是 96.33% [7]。目前深度学习可以达到 99.47%的识别率[8]。

在学术界收到广泛关注的同时，深度学习也在工业界产生了巨大的影响。在 Hinton 的科研小组赢得 ImageNet 比赛之后 6 个月，谷歌和百度发都布了新的基于图像内容的搜索引擎。他们沿用了 Hinton 在 ImageNet 竞赛中用的深度学习模型，应用在各自的数据上，发现图像搜索的准确率得到了大幅度的提高。百度在 2012 年就成立了深度学习研究院，于 2014 年五月又在美国硅谷成立了新的深度学习实验室，聘请斯坦福著名教授吴恩达担任首席科学家。Facebook 于 2013 年 12 月在纽约成立了新的人工智能实验室，聘请深度学习领域的著名学者，卷积网路的发明人 Yann LeCun 作为首席科学家。2014 年 1 月，谷歌四亿美金收购了一家深度学习的创业公司，DeepMind。鉴于深度学习在学术和工业界的巨大影响力，2013 年 MIT Technology Review 将其列为世界十大技术突破之首。

2. 深度学习有何与众不同？

许多人会问，深度学习和其它机器学习方法相比有哪些关键的不同点，它成功的秘密在哪里？我们下面将对这此从几个方面作简要的阐述。

2.1 特征学习

深度学习与传统模式识别方法的最大不同在于它是从大数据中自动学习特征，而非采用手工设计的特征。好的特征可以极大提高模式识别系统的性能。在过去几十年模式识别的各种应用中，手工设计的特征处于同统治地位。它主要依靠设计者的先验知识，很难利用大数据的优势。由于依赖手工调参数，特征的设计中只允许出现少量的参数。深度学习可以从大数据中自动学习特征的表示，其中可以包含成千上万的参数。

手工设计出有效的特征是一个相当漫长的过程。回顾计算机视觉发展的历史，往往需要五到十年才能出现一个受到广泛认可的好的特征。而深度学习可以针对新的应用从训练数据中很快学习得到新的有效的特征表示。

一个模式识别系统包括特征和分类器两个主要的组成部分，二者关系密切，而在传统的方法中它们的优化是分开的。在神经网络的框架下，特征表示和分类器是联合优化的，可以最大程度发挥二者联合协作的性能。

以 2012 年 Hinton 参加 ImageNet 比赛所采用的卷积网络模型[9]为例，这是他们首次参加 ImageNet 图像分类比赛，因此没有太多的先验知识。模型的特征表示包含了 6 千万个参数，从上百万样本中学习得到。令人惊讶的是，从 ImageNet 上学习得到的特征表示具有非常强的泛化能力，可以成功地应用到其它的数据集和任务，例如物体检测、跟踪和检索等等。在计算机视觉领域另外一个著名的竞赛是 PSACAL VOC。但是它的训练集规模较小，不适合训练深度学习模型。有学者将 ImageNet 上学习得到的特征表示用于 PSACAL VOC 上的物体检测，将检测率提高了 20%[10]。

既然特征学习如此重要，什么是好的特征呢？一幅图像中，各种复杂的因素往往以非线性的方式结合在一起。例如人脸图像中就包含了身份、姿态、年龄、表情和光线等各种信息。深度学习的关键就是通过多层非线性映射将这些因素成功的分开，例如在深度模型

的最后一个隐含层，不同的神经元代表了不同的因素。如果将这个隐含层当作特征表示，人脸识别、姿态估计、表情识别、年龄估计就会变得非常简单，因为各个因素之间变成了简单的线性关系，不再彼此干扰。

2.2 深层结构的优势

深度学习模型意味着神经网络的结构深，由很多层组成。而支持向量机和 Boosting 等其它常用的机器学习模型都是浅层结构。有理论证明，三层神经网络模型（包括输入层、输出层和一个隐含层）可以近似任何分类函数。既然如此，为什么需要深层模型呢？

理论研究表明，针对特定的任务，如果模型的深度不够，其所需要的计算单元会呈指数增加。这意味着虽然浅层模型可以表达相同的分类函数，其需要的参数和训练样本要多得多。浅层模型提供的是局部表达。它将高维图像空间分成若干局部区域，每个局部区域存储至少一个从训练数据中获得的模板。浅层模型将一个测试样本和这些模板逐一匹配，根据匹配的结果预测其类别。例如在支持向量机模型中，这些模板就是支持向量；在最近邻分类器中，这些模板是所有的训练样本。随着分类问题复杂度的增加，图像空间需要被划分成越来越多的局部区域，因而需要越来越多的参数和训练样本。

深度模型能够减少参数的关键在于重复利用中间层的计算单元。例如，它可以学习针对人脸图像的分层特征表达。最底层可以从原始像素学习滤波器，刻画局部的边缘和纹理特征；通过对各种边缘滤波器进行组合，中层滤波器可以描述不同类型的人脸器官；最高层描述的是整个人脸的全局特征。

深度学习提供的是分布式的特征表示。在最高的隐含层，每个神经元代表了一个属性分类器，例如男女、人种和头发颜色等等。每个神经元将图像空间一分为二，N 个神经元的组合就可以表达 2^N 个局部区域，而用浅层模型表达这些区域的划分至少需要个 2^N 模板。由此我们可以看到深度模型的表达能力更强，更有效率。

2.5 提取全局特征和上下文信息的能力

深度模型具有强大的学习能力，高效的特征表达能力，从像素级原始数据到抽象的语义概念逐层提取信息。这使得它在提取图像的全局特征和上下文信息方面具有突出的优势。这为解决一些传统的计算机视觉问题，如图像分割和关键点检测，带来了新的思路。

以人脸的图像分割为例。为了预测每个像素属于哪个脸部器官（眼睛、鼻子、嘴、头发），通常的作法是在该像素周围取一个小的区域，提取纹理特征（例如局部二值模式），再基于该特征利用支持向量机等浅层模型分类。因为局部区域包含信息量有限，往往产生分类错误，因此要对分割后的图像加入平滑和形状先验等约束。

事实上即使存在局部遮挡的情况下，人眼也可以根据脸部其它区域的信息估计被遮挡处的标注。这意味着全局和上下文的信息对于局部的判断是非常重要的，而这些信息在基于局部特征的方法中从最开始阶段就丢失了。理想的情况下，模型应该将整幅图像作为输入，直接预测整幅分割图。图像分割可以被当作一个高维数据转换的问题来解决。这样不但利用到了上下文信息，模型在高维数据转换过程中也隐式地加入了形状先验。但是由于整幅图像内容过于复杂，浅层模型很难有效地捕捉全局特征。深度学习的出现使这一思路成为可能，在人脸分割[11]、人体分割[12]、人脸图像配准[13]和人体姿态估计等各个方面都取得了成功[14]。

2.4 联合深度学习

一些计算机视觉学者将深度学习模型视为黑盒子，这种看法是不全面的。事实上我们可以发现传统计算机视觉系统和深度学习模型存在着密切的联系，而且可以利用这种联系提出新的深度模型和新的训练方法。这方面一个成功的例子是用于行人检测的联合深度学习[15]。一个计算机视觉系统包含了若干关

键的组成模块。例如一个行人检测器就包括了特征提取、部件检测器、部件几何形变建模、部件遮挡推理、分类器等等。在联合深度学习中[15]，深度模型的各个层和视觉系统的各个模块可以建立起对应关系。如果视觉系统中一些有效的关键模块在现有深度学习的模型中没有与之对应的层，它们可以启发我们提出新的深度模型。例如大量物体检测的研究工作证明对物体部件的几何形变建模可以有效地提高检测率，但是在常用的深度模型中没有与之相对应的层。于是联合深度学习[15]及其后续的工作[16]都提出了新的形变层和形变池化层实现这一功能。

从训练方式上看，计算机视觉系统的各个模块是逐一训练或手工设计的；在深度模型的预训练阶段，各个层也是逐一训练的。如果我们能够建立起计算机视觉系统和深度模型之间的对应关系，在视觉研究中积累的经验可以对深度模型的预训练提供指导。这样预训练后得到的模型至少可以达到与传统计算机视觉系统可比的结果。在此基础上，深度学习还会利用反向传播对所有的层进行联合优化，使它们之间的相互协作达到最优，从而使整个网络的性能得到重大提升。

3. 深度学习在物体识别中的应用

3.1 ImageNet 图像分类

深度学习在物体识别中最重要的进展体现在 ImageNet ILSVRC 挑战中的图像分类任务。传统计算机视觉方法在这个测试集上最低的 top5 错误率是 26.172%。2012 年 Hinton 的研究小组利用卷积网络在这个测试集上把错误率大幅降到 15.315%。这个网络的结构被称作 Alex Net [3]。与传统的卷积网络相比，它有三点比较重要的不同。首先是采用了 dropout 的训练策略，在训练过程中将一些输入层和中间层的神经元随机置零。这模拟了由于噪音和对输入数据的各种干扰，从而使一些神经元对一些视觉模式产生漏检的情况。Dropout 使训练过程收敛更慢，但得到的网络模型更加鲁棒。其次，它采用整流线型单元作为非线性的激发函数。这不仅大大降低了计算的复杂度，而且使神经元的输出具有稀疏的性质。稀疏的特征表示对各种干扰更加鲁棒。第三，它通过对训练样本镜像映射，和加入随机平移扰动产生了更多的训练样本，以减少过拟合。

ImageNet ILSVRC2013 比赛中，排名前 20 的小组使用的都是深度学习，其影响力可见一斑。获胜者是来则纽约大学 Rob Fergus 的研究小组，所采用的深度模型还是卷积网络，对网络结构作了进一步优化。Top5 错误率降到 11.197%，其模型称作 Clarifai[17]。

2014 年深度学习又取得了重要进展，在 ILSVRC2014 比赛中，获胜者 GooLeNet[18]将 top5 错误率降到 6.656%。它突出的特点是大大增加了卷积网络的深度，超过 20 层，这在之前是不可想象的。很深的网络结构给预测误差的反向传播带了困难。因为预测误差是从最顶层传到底层的，传到底层的误差很小，难以驱动底层参数的更新。GooLeNet 采取的策略是将监督信号直接加到多个中间层，这意味着中间和低层的特征表示也需要能够准确对训练数据分类。如何有效地训练很深的网络模型仍是未来研究的一个重要课题。

虽然深度学习在 ImageNet 上取得了巨大成功，但是一个现实的问题是，很多应用的训练集是较小的，如何在这种情况下应用深度学习呢？有三种方法可供读者参考。（1）可以将 ImageNet 上训练得到的模型做为起点，利用目标训练集和反向传播对其进行继续训练，将模型适应到特定的应用[10]。ImageNet 起到预训练的作用。（2）如果目标训练集不够大 ，也可以将低层的网络参数固定，沿用 ImageNet 上的训练集结果，只对上层进行更新。这是因为底层的网络参数是最难更新的，而从 ImageNet 学习得到的底层滤波器往往描述了各种不同的局部边缘和纹理信息，而这些滤波器对一般的图像有较好的普适性。（3）直接采用 ImageNet 上训练得到的模型，把最高的隐含层的输出作为特征表达，代替常用的手工设计的特征[19][20]。

3.2 人脸识别

深度学习在物体识别上了另一个重要突破是人脸识别。人脸识别的最大挑战是如何区分由于光线、姿态和表情等因素引起的类内变化和由于身份不同产生的类间变化。这两种变化分布是非线性的且极为复杂，传统的线性模型无法将它们有效区分开。深度学习的目的是通过多层的非线性变换得到新的特征表示。该特征须要尽可能多地去掉类内变化，而保留 类间变化。

人脸识别有两种任务，人脸确认和人脸辨识。人脸确认的任务是判断两张人脸照片是否属于同一个人，属二分类问题，随机猜的正确率是 50%。人脸辨识的任务是将一张人脸图像分为 N 个类别之一，类别是由人脸的身份定义的。这是个多分类问题，更具挑战性，其难度随着类别数的增多而增大，随机猜的正确率是 1/N。两个任务都可以用来通过深度模型学习人脸的特征表达。

2013 年，[21]采用人脸确认任务作为监督信号，利用卷积网络学习 人脸特征，在 LFW 上取得了 92.52%的识别率。这一结果虽然与后续的深度学习方法相比较低，但也超过了大多数非深度学习的算法。由于人脸确认是一个二分类的问题，用它学习人脸特征效率比较低。这个问题可以从几个方面理解。深度学习面临的一个主要问题是过拟合。作为一个二分类问题，人脸确认任务相对简单，容易在训练集上发生过拟合。与之不同，人脸辨识是一个更具 挑战性的多分类问题，不容易发生过拟合，更适合通过深度模型学习人脸特征。另一方面，在人脸确认中，每一对训练样本被人工标注成两类之一，所含信息量较少。而在人脸辨识中，每个训练样本都被人工标注成 N 类之一，信息量要大的多。

2014 年 CVPR，DeepID[22]和 DeepFace[23] 都采用人脸辨识作为监督信号，在 LFW 上取得了 97.45%和 97.35%的识别率。他们利用卷积网络预测 N 维标注向量，将最高的隐含层作为人脸特征。这一层在训练过程中要区分大量的人脸类别（例如在 DeepID 中要区分 1000 类人脸），因此包含了丰富的类间变化的信息，而且有很强的泛化能力。虽然训练中采用的是人脸辨识任务，得到特征可以应用到人脸确认任务，以及识别训练集中没有新人。例如 LFW 上用于测试的任务是人脸确认任务，不同于训练中采用的人脸辨识任务；DeepID[22]和 DeepFace[23]的训练集与 LFW 测试集的人物身份是不重合的。

通过人脸辨识任务学习得到的人脸特征包含了较多的类内变化。DeepID2[24]联合使用人脸确认和人脸辨识作为监督信号，得到的人脸特征在保持类间变化的同时最小化类内变化，从而将 LFW 上的人脸识别率提高到 99.15%。利用 Titan GPU，DeepID2 提取一幅人脸图像的特征只需要 35 毫秒，而且可以离线进行。经过 PCA 压缩最终得到 80 维的特征向量，可以用于快速人脸在线比对。在后续的工作中，DeepID2+[25]对 DeepID2 通过加大网络结构，增加训练数据，以及在每一层都加入监督信息进行了进一步改进，在 LFW 达到了 99.47%的识别率。

一些人认为深度学习的成功在于用具有大量参数的复杂模型去拟合数据集。这个看法也是不全面的。事实上，进一步的研究[25]表明 DeepID2+的特征有很多重要有趣的性质。例如，它最上层的神经元响应是中度稀疏的，对人脸身份和各种人脸属性具有很强的选择性，对局部遮挡有很强的鲁棒性。以往的研究中，为了得到这些属性，我们往往需要对模型加入各种显示的约束。而 DeepID2+通过大规模学习自动拥有了这些引人注目的属性，其背后的理论分析值得未来进一步研究。

4. 深度学习在物体检测中的应用

深度学习也对图像中的物体检测带来了巨大提升。物体检测是比物体识别更难的任务。一幅图像中可能包含属于不同类别的多个物体，物体检测需要确定每个物体的位置和类别。深度学习在物体检测中的进展也体现在 ImageNet ILSVRC 挑战中。2013 年比赛的组织者增加了物体检测的任务，需要在四万张互联网图片中检测 200 类物体。当年的比赛中

赢得物体检测任务的方法使用的依然是手动设计的特征，平均物体检测率，即 mean Averaged Precision (mAP)，只有 22.581%。在 ILSVRC2014 中，深度学习将 mAP 大幅提高到 43.933%。较有影响力的工作包括 RCNN[10]，Overfeat[26]，GoogLeNet[18]，DeepID-Net[27]，network in network[28]，VGG[29]，和 spatial pyramid pooling in deep CNN[30]。被广泛采用的基于深度学习的物体检测流程是在 RCNN[10]中提出的。首先采用非深度学习的方法（例如 selective search[31]）提出候选区域，利用深度卷积网络从候选区域提取特征，然后利用支持向量机等线性分类器基于特征将区域分为物体和背景。DeepID-Net[27]将这一流程进行了进一步的完善使得检测率有了大幅提升，并且对每一个环节的贡献做了详细的实验分析。此外深度卷积网络结构的设计也至关重要。如果一个网络结构提高提高图像分类任务的准确性，通常也能使物体检测器的性能显著提升。

深度学习的成功还体现在行人检测上。在最大的行人检测测试集(Caltech[32])上，被广泛采用的 HOG 特征和可变形部件模型[33]平均误检率是 68%。目前基于深度学习最好的结果是 20.86%[34]。在最新的研究进展中，很多在物体检测中已经被证明行之有效的思路都有其在深度学习中的实现。例如，联合深度学习[15]提出了形变层，对物体部件间的几何形变进行建模；多阶段深度学习[35]可以模拟在物体检测中常用的级联分类器；可切换深度网络[36]可以表达物体各个部件的混合模型；[37]通过迁移学习将一个深度模型行人检测器自适应到一个目标场景。

5. 深度学习用于视频分析

深度学习在视频分类上的应用总体而言还处于起步阶段，未来还有很多工作要做。描述视频的静态图像特征，可以采用用从 ImageNet 上学习得到的深度模型；难点是如何描述动态特征。以往的视觉方法中，对动态特征的描述往往依赖于光流估计，对关键点的跟踪，和动态纹理。如何将这些信息体现在深度模型中是个难点。最直接的做法是将视频视为三维图像，直接应用卷积网络[38]，在每一层学习三维滤波器。但是这一思路显然没有考虑到时间维和空间维的差异性。另外一种简单但更加有效的思路是通过预处理计算光流场，作为卷积网络的一个输入通道[39]。也有研究工作利用深度编码器(deep autoencoder)以非线性的方式提取动态纹理[40]，而传统的方法大多采用线性动态系统建模。在一些最新的研究工作中[41]，长短记忆网络（LSTM）正在受到广泛关注，它可以捕捉长期依赖性，对视频中复杂的动态建模。

6. 未来发展的展望

深度学习在图像识别中的发展方兴未艾，未来有着巨大的空间。本节对几个可能的方向进行探讨。

在物体识别和物体检测中正趋向使用更大更深的网络结构。ILSVRC2012 中 Alex Net 只包含了 5 个卷积层和两个全连接层。而 ILSVRC2014 中 GooLeNet 和 VGG 使用的网络结构都超过了 20 层。更深的网络结构使得反向传播更加困难。与此同时训练数据的规模也在迅速增加。这迫切需要研究新的算法和开发新的并行计算系统更加有效的利用大数据训练更大更深的模型。

与图像识别相比，深度学习在视频分类中的应用还远未成熟。从 ImageNet 训练得到的图像特征可以直接有效地应用到各种与图像相关的识别任务（例如图像分类、图像检索、物体检测和图像分割等等），和其它不同的图像测试集，具有良好的泛化性能。但是深度学习至今还没有得到类似的可用于视频分析的特征。要达到这个目的，不但要建立大规模的训练数据集（[42]最新建立了包含一百万 YouTube 视频的数据库），还需要研究适用于视频分析的新的深度模型。此外训练用于视频分析的深度模型的计算量也会大大增加。

在与图像和视频相关的应用中，深度模型的输出预测（例如分割图或物体检测框）往往

具有空间和时间上的相关性。因此研究具有结构性输出的深度模型也是一个重点。

虽然神经网络的目的在于解决一般意义的机器学习问题，领域知识对于深度模型的设计也起着重要的作用。在于图像和视频相关的应用中，最成功的是深度卷积网络，它正是利用了与图像的特殊结构。其中最重要的两个操作，卷积和池化（pooling）都来自于与图像相关的领域知识。如何通过研究领域知识，在深度模型中引入新的有效的操作和层，对于提高图像识别的性能有着重要意义。例如池化层带来了局部的平移不变性，[27]中提出的形变池化层在此基础上更好的描述了物体各个部分的几何形变。在未来的研究中，可以将其进一步扩展，从而取得旋转不变性、尺度不变性、和对遮挡的鲁棒性。通过研究深度模型和传统计算机视觉系统之间的关系，不但可以帮助我们理解深度学习成功的原因，还可以启发新的模型和训练方法。联合深度学习[15]和多阶段深度学习[35]是两个例子，未来这方面还可以有更多的工作。

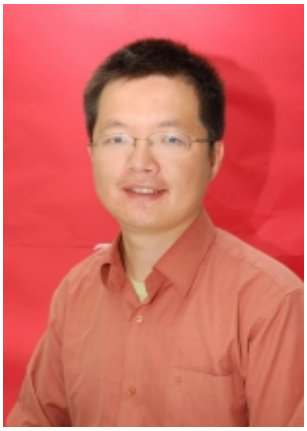
最然深度学习在实践中取得了巨大成功，通过大数据训练得到的深度模型体现出的特性（例如稀疏性、选择性、和对遮挡的鲁棒性[22]）引人注目，其背后的理论分析还有许多工作需要在未来完成。例如，何时收敛，如何取得较好的局部极小点，每一层变换取得了那些对识别有益的不变性，又损失了那些信息等等。最近 Mallat 利用小波对深层网络结构进行了量化分析[43]，是在这一个方向上的重要探索。

深度学习在图像识别上的巨大成功，必将对于多媒体相关的各种应用产生重大影响。我们期待着更多的学者在不久的将来研究如何利用深度学习得到的图像特征，推动各种应用的快速进步。

7. 结束语

2012 年以来，深度学极大的推动了图像识别的研究进展，突出体现在 ImageNet ILSVRC 和人脸识别，而且正在快速推广到与图像识别相关的各个问题。深度学习的本质是通过多层非线性变换，从大数据中自动学习特征，从而替代手工设计的特征。深层的结构使其具有极强的表达能力和学习能力，尤其擅长提取复杂的全局特征和上下文信息，而这是浅层模型难以做到的。一幅图像中，各种隐含的因素往往以复杂的非线性的方式关联在一起，而深度学习可以使这些因素分级开，在其最高隐含层不同神经元代表了不同的因素，从而使分类变得简单。深度模型并非黑盒子，它与传统的计算机视觉体统有着密切的联系，但是它使得这个系统的各个模块（即神经网络的各个层）可以通过联合学习，整体优化，从而性能得到大幅提升。与图像识别相关的各种应用也在推动深度学习在网络结构、层的设计和训练方法各个方面的的快速发展。我们可以预见在未来的数年内，深度学习将会在理论、算法、和应用各方面进入高速发展的时期，期待着愈来愈多精彩的工作对学术和工业界产生深远的影响。

作者



王晓刚 2011 年于中国科技大学少年班获得学士学位,2004 年于香港中文大学信息工程系获得硕士学位,2009 年于美国麻省理工学院获计算机博士学位。自 2009 年至今任香港中文大学电子工程系助理教授。他于 2012 年获得香港杰出青年学者奖,香港中文大学青年学者奖。他是国际期刊 Image and Visual Computing Journal 的副主编,IEEE Transactions on Circuits and Systems for Video Technology 的副主编, 2011 年

IEEE 国际计算机视觉大会(ICCV) 的领域主席, 2014 年欧洲计算机视觉大会(ECCV)的领域主席, 2014 年亚洲计算机视 觉大会(ACCV)的领域主席。他的研究兴趣包括计算机视觉、深度学习、群体视频监控、物体检测、和人脸识别等等。

参考文献

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Nature, 323(99):533–536, 1986.

[2] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2009.

[3] A. Krizhevsky, L. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Proc. Neural Information Processing Systems, 2012.

[4] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miler. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

[5] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In IEEE Int'l Conf. Computer Vision, 2009.

[6] M. Turk and A. Pentland. Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1):71–86, 1991.

[7] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: Highdimensional feature and its efficient compression for face verification. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2013.

[8] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. arXiv:1412.1265, 2014.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86:2278– 2324, 1998.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2014.

[11] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2012.

[12] P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional network. In Proc. IEEE Int'l Conf. Computer Vision, 2013.

[13] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2013.

[14] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2014.

[15] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In Proc. IEEE Int'l Conf. Computer Vision, 2013.

[16] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, C. Qian, Z. Zhu, R. Wang, C. Loy, X. Wang, and X. Tang. Deepidnet: multi-stage and deformable deep convolutional neural networks for object detection. arXiv:1409.3505, 2014.

[17] http://www.clarifai.com/

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. arXiv:1409.4842, 2014.

[19] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. arXiv:1403.6382, 2014.

[20] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep

convolutional activation features. arXiv:1403.1840, 2014.

[21] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for computing face similarities. In Proc. IEEE Int'l Conf. Computer Vision, 2013.

[22] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2014.

[23] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2014.

[24] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In Proc. Neural Information Processing Systems, 2014.

[25] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. arXiv:1412.1265, 2014.

[26] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Le- Cun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In Proc. Int'l Conf. Learning Representations, 2014.

[27] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, C. Qian, Z. Zhu, R. Wang, C. Loy, X. Wang, and X. Tang. Deepidnet: multi-stage and deformable deep convolutional neural networks for object detection. arXiv:1409.3505, 2014.

[28] M. Lin, Q.. Chen, and S. Yan. Network in network. arXiv:1312.4400v3, 2013.

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014, 2014.

[30] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. arXiv:1406.4729, 2014.

[31] J. R. R. Uijlings, K. E. A. Van de Sande, T. Gevers, and W. M. Smeulders. Selective search for object recognition. International Journal of Computer Vision, 104:154–171, 2013.

[32] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2009.

[33] P. Felzenszwalb, R. B. Grishick, D.McAllister, and D. Ramanan. Object detection with discriminatively trained part based models. IEEE Trans. PAMI, 32:1627–1645, 2010.

[34] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian Detection aided by Deep Learning Semantic Tasks," arXiv 2014.

[35] X. Zeng, W. Ouyang, and X. Wang. Multi-stage contextual deep learning for pedestrian detection. In Proc. IEEE Int'l Conf. Computer Vision, 2013.

[36] P. Luo, Y. Tian, X. Wang, and X. Tang. Switchable deep network for pedestrian detection. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2014.

[37] X. Zeng, W. Ouyang, and X. Wang. Deep learning of scene-specific classifier for pedestrian detection. In Proc. European Conf. Computer Vision, 2014.

[38] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 35(1):221–231, 2013.

[39] K. Simonyan and A. Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. arXiv:1406.2199, 2014.

[40] X. Yan, H. Chang, S. Shan, and X. Chen, Modeling Video Dynamics with Deep Dynencoder. In Proc. European Conf. Computer Vision, 2015.

[41] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. arXiv:1411.4389, 2014.

[42] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale

video classification with convolutional neural networks. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2014.

[43] J. Bruna and S. Mallat. Invariant scattering convolution networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 35(8):1872 – 1886, 2013.