深度网络：他们为视觉做过什么？

**摘要：**这是一篇关于深度视觉网络的优缺点的观点论文。它们是人工智能最新进展的中心，在认知科学和神经科学中也越来越重要。他们取得了巨大的成功，但也有明显的局限性。对它们的内部工作原理也只有部分了解。目前形式的深层网络似乎不太可能成为构建通用智能机器或理解大脑的最佳长期解决方案，但它们的许多方面可能仍将存在。目前，深度网络在特定类型的可视任务和特定的基准测试数据集上做得非常好。但与人类视觉系统相比，深层网络的通用性、灵活性和适应性要差得多。此外，像深网这样的方法在面对自然图像的巨大复杂性时可能会遇到根本性的困难，这可能导致组合爆炸。为了说明我们的主要观点，同时尽量减少参考文献，这篇论文对我们小组的工作略有偏向。

**关键词：**深度神经网络 计算机视觉 成功 限制 认知科学 神经科学

### 介绍

在过去的几年里，深度网络在计算机视觉和生物视觉系统研究方面取得了巨大进步。但作为这些领域的研究人员，我们发现自己对此有着复杂的感受。一方面，我们惊叹于他们的成功以及他们如何在一些现实世界的任务中取得惊人的成果，并且在学术环境中，他们在基准数据集上的表现几乎总是优于其他方法。但是，另一方面，我们意识到他们对围绕他们的炒作的局限和担忧。最近几篇论文（Darwiche，2018; Marcus，2018）从机器推理和认知科学的角度对深度网络进行了批评，论文认为尽管深度网络作为一种工具很有用，但它们需要与其他方法相结合才能实现人类智力。我们研究的本质意味着我们与许多学科（认知科学，计算机科学，应用数学，工程学，神经科学，物理学和放射学）的研究教师互动，而深度网络则是一个经常谈话的话题。我们发现自己花了一半时间批评深度网络的局限性，而另一半时间则赞扬他们并捍卫他们反对他们的批评者。当我们不是经常但自信的告诉你，深度网络永远不能够做xxx时，我们已经知道，他们可以。本意见书旨在提供一个关于视觉学习的深度网络的优势和劣势的平衡观点。

本文的组织如下。在第2节中，我们讨论了神经网络的历史及其繁荣与萧条的趋势。第3节描述了Deep Nets的一些成功，同时也提到了警告和细则。在第4节中，我们讨论了对Deep Nets内部运作的有限理解。第5节探讨了它们有助于构建生物视觉系统理论的潜力，以及它们与真实神经元和神经回路的有限关系。在第6节中，我们讨论了Deep Nets正在努力解决的挑战。第7节更具有推测性，并认为视觉研究人员试图模拟日益复杂的视觉任务，他们将面临深层网可能无法处理的组合爆炸。

### 历史Some History

我们正处于第三波神经网络方法中。前两次浪潮——上世纪50年代至60年代和80年代至90年代——引发了相当大的兴奋，但慢慢失去了动力。除了少数例外，对于机器(人工智能/机器学习)和理解生物视觉系统(神经科学、认知科学、心理学)来说，神经网络的整体表现令人失望。但第三次浪潮(2000年至今)之所以引人注目，是因为深度网络在许多重大基准问题上取得了巨大成功，并在工业上应用于现实世界的任务。应当承认，目前许多成功的神经网络的几乎所有基本思想都是在第二次浪潮中发展起来的。但是直到大数据集的可用性和强大的计算机(如gpu)的普及(2000年后才开始使用，并推动了第三次浪潮)，人们才意识到它们的优势。

这些神经网络波动的起伏反映了智力时尚的变化和其他方法的流行程度的变化。第二波神经网络在一定程度上是由传统人工智能的局限性所驱动的。在传统人工智能中，令人失望的结果和对过度承诺的指责导致了上世纪80年代中期的人工智能寒冬。第二波的衰落对应于支持向量机、核方法和相关方法的兴起。这要归功于那些神经网络研究人员，他们在有时很难发表神经网络论文的情况下，不顾挫折，通过波谷继续研究。钟摆现在又摆了起来，有时似乎很难发表任何与神经网络无关的东西。我们猜测，如果研究人员抵制时尚的吸引力，转而追求多样化的方法和技术，进展会更快。同样令人担忧的是，为学生开设的课程往往倾向于紧跟最新潮流，而忽视了旧的技术(直到它们被重新发现)。

目前神经网络的成功主要用于人工智能任务，他们在人脸识别（目前正在研究数千万人的数据集）和医学图像分析等任务方面取得了重大进展。神经网络越来越多地用于模拟思维和大脑，但它们与真实神经元和神经回路的关系应该谨慎对待。虽然人工神经网络受到生物学的启发，但必须承认真正的神经元要复杂得多，理解真正的神经回路仍然是神经科学最基本的挑战之一。

### 成功，精美印刷

计算机视觉社区一直对深度网络持怀疑态度，直到AlexNet（Krizhevsky等人，2012）在ImageNet（Deng et al。，2009）中对物体进行分类。此分类任务假设存在一个前景对象，该对象由有限的背景区域舍入，因此输入类似于图1中右下角图像中的一个红色框。AlexNet的成功刺激了视觉社区，导致各种深度网络架构在对象分类方面的表现越来越好，例如（Simonyan和Zisserman，2015; He等，2016; Liu等，2018）

深度网络也迅速适应其他视觉任务，例如物体检测，其中图像包含一个或多个物体，背景更大，例如PASCAL挑战（Everingham等，2010）。对于这项任务，深度网络通过初始阶段进行扩充，该阶段为对象的可能位置和大小提出建议，然后应用深度网络对提案进行分类（当前方法将所有方案和对象组合在一起，称为end-to-end）。这些方法优于以前的最佳方法，可变形零件模型（Felzenszwalb等，2010），用于PASCAL物体检测挑战（PASCAL是ImageNet之前的主要物体检测和分类挑战）。其他深度网络架构也在其他经典任务中提供了巨大的性能提升，如边缘检测，语义分割，遮挡检测（带边界所有权的边缘检测），对称轴检测。人体关节检测，人体分割，双目立体，3D深度估计也发生了重大增加，单个图像和场景分类。其中一些任务如图1所示。

但是虽然Deep Nets非常有效，几乎总是超越替代技术，但它们不是通用的，它们的成功来自以下三个限制。

首先，Deep Nets专为特定的视觉任务而设计。大多数Deep Nets都是针对单个任务而设计的，而针对一个任务设计的Deep Net将不适合另一个任务。例如，为ImageNet上的物体分类而设计的Deep Net无法在利兹运动数据集（LSD）上执行人体解析（即检测人体关节）。然而，有一些例外和“转移学习”，有时可以使在一项任务上训练的深度网络适应密切相关的任务，前提是注释数据可用于该任务（参见第6.1节）。直观地说，这是因为Deep Net学到的特征捕获了对这两个任务都有用的图像结构。此外，研究人员最近开发了Deep Nets，例如UberNet（Kokkinos，2017），它可以使用相同的网络执行多达四项任务。但是，总的来说，有一个不同的Deep Net架构的动物园正在设计用于特定的任务，包括网络级联和网络的几个不同级别的监督。

其次，在基准标记数据集上表现良好的Deep Nets可能会在数据集之外的真实世界图像上严重失败。这是因为真实世界图像的集合无限大，因此任何数据集都难以代表现实世界的复杂性，无论多大，都是如此。这是一个重要的问题，我们将在第7节中回到这里。目前，我们只是重新标记所有数据集都存在偏差。 这些偏差在早期视觉数据集中特别明显，并且研究人员通过利用背景背景迅速学会利用它们（例如，在Caltech101中检测鱼很容易，因为它们是唯一背景为水的物体）。比较研究表明，在某些数据集上表现良好的方法往往无法推广到其他数据集（Tor ralba和Efros，2011）。 尽管使用了大数据集和Deep Nets，但这些问题仍然存在。例如，即使对于ImageNet，背景仍然存在问题（Zhu等，2017）。如果数据集包含来自有限观察条件的对象，则也会出现偏差，例如，如图2所示，如果从训练数据集中表示的视点中显示，训练用于检测ImageNet上的沙发的Deep Net可能无法识别它们。特别是，深网对数据集中经常出现的事件有偏见。但是在现实世界的应用中，这些偏差特别成问题，因为它们可能对应于视觉系统的故障可能导致可怕后果的情况，例如，用于训练自动驾驶车辆的数据集几乎从不包括坐在路上的婴儿。 类似地，数据集通常倾向于低估已知会导致算法失败的危险因素，例如双目立体声的镜面反射。 我们将在第6.3节回到这个例子。

第三，几乎所有Deep Nets都需要注释数据进行培训和测试。 这会使视觉研究人员偏向于那些容易进行注释的视觉任务。例如，对象检测的注释仅需要在对象周围指定紧密的边界框。 但是对于其他视觉任务，例如检测人类的关节，注释要困难得多，而对于某些任务来说，几乎是不可能的。有一些方法可以减少第6.1节中讨论的监督需求，并且还有可能使用合成刺激（由计算机图形引擎生成），使得groundtruth可用于所有视觉任务。但是，现实的合成刺激是有限的，视觉社区不愿依赖它直到它们变得足够逼真。

总之，Deep Nets是一套不断根据特定视觉任务的需求进行改进和开发的工具。他们几乎都依赖于完全监督的数据，我们稍后将讨论这些警告，并且他们的表现可能无法推广到他们接受过训练的数据集之外的图像。由于现实世界图像的无限复杂性，数据集偏差对于视觉尤其成问题，我们将在第7节中讨论。

### 了解深度网络

很难描述Deep Nets可以做什么以及理解他们的内部运作。理论结果表明，如果有足够数量的隐藏单元，多层感知器和深网可以代表任何输入输出函数（Hornik等，1989）。但是，正如任何已经证明这种类型的定理的人都很清楚（Xu et al。，1994），在渐近极限中保持的理论结果具有有限的效用。更有价值的是有限数量的隐藏单元和有限的训练数据的结果，但很难看出对于像Deep Nets这样复杂的系统可以获得什么意义上的理论结果。

在更直观的层面上，至少在应用于视觉任务时，似乎可以粗略地了解Deep Nets。Deep Nets的层次结构类似于视觉文本的经典模型，例如NeoCognition（Fukushima和Miyake，1982）和HMax（Riesenhuber和Poggio，1999），并且捕获了激发这些模型的许多直觉。深网包含特征表示，其中较低级别的那些具有有限大小的感知字段并且对模式的精确位置敏感。但是当我们提升层次结构时，接收字段变得更大并且对特定模式更敏感，而不太关心它们的确切位置。

通过研究深网络的内部过滤器/特征的活动可以部分地理解这一点（Zeiler和Fergus，2014; Yosinski等，2015）。特别是，如果深度网被训练用于场景分类，则一些卷积层过滤器大致对应于在场景中经常出现的对象，而如果深度网被训练用于对象检测，则一些特征大致对应于对象的部分（ Zhou et al。，2015）。在对受限制的物体子集（例如，车辆）的详细研究中，研究人员（Wang等人，2015）发现了特征向量的规则活动模式，称为视觉概念，其大致对应于对象的语义部分（具有敏感性）查看点），见图3。但我们承认，尽管这些研究令人鼓舞，但它们仍然具有相当的压力，缺乏真正理解的精确性（例如，这些研究尚未启用研究人员）以不受限制的方式学习对象和对象部分的模型。

这表明了以下粗网概念图。卷积级别表示不同抽象级别的强度模式的多样性。最低级别表示局部图像模式，而高级别表示较大的模式，这些模式对于强度模式的细节是不变的。从相关的角度来看，权重向量表示图像模式的模板的字典。Deep Net的最终“eision层”通常难以解释，但它们基于较低层所代表的模板进行决策是合理的。这种“模板”对Deep Nets的解释表明它们非常有效地学习和表示各种各样的图像模式，并在它们之间进行插值，但不能超出它们在训练数据集中看到的模式。其他研究表明，Deep Nets在建模视觉属性方面效率较低，视觉属性完全由几何体指定，特别是如果输入由对应于边界边缘存在或不存在的二进制值模式组成。Deep Nets是否可以学习能够摧毁不同视觉特性的特征是一个悬而未决的问题，正如我们将在后面的第7节中所述，最终将是处理真实图像的完整复杂性所必需的。

### 深度网络和生物视觉

Deep Nets可以为研究生物视觉系统提供很多东西，尤其是认知科学，神经科学和心理学等学科，旨在了解大脑和大脑。他们可以通过利用大数据的可用性来帮助开发和测试计算理论，同时通过将深网中的人工神经元与大脑中的真实神经元相关联来提高理解大脑的可能性。但它们对于建模神经回路和人类认知能力也有很大的局限性。

#### 5.1利用大数据

Deep Nets和其他机器学习技术的使用可以帮助开发利用大数据的思维和大脑理论。 这可以通过三种方式完成。首先，Deep Nets可以帮助开发处理真实世界图像的巨大复杂性的理论。其次，它们可以用来部分地学习人类和其他动物通过发展和经验获得的视觉世界的知识。第三，它们使理论能够在复杂的刺激上进行测试，并与其他理论进行比较。我们现在将依次解决这些问题

历史上，生物视觉系统的研究很大程度上依赖于简单的合成刺激。由于自然图像刺激的复杂性意味着通过系统地改变实验参数来进行受控的科学实验是非常困难的，因此这些研究已经产生了许多重要的发现并且是他的必要条件。这也遵循完善的分而治之的科学策略，旨在通过将复杂现象分解为更容易理解的块来理解。但研究简化刺激的愿景有深度网和大数据可以帮助解决的局限性。作为20世纪80年代发现的计算机视觉研究人员，简化合成刺激的发现虽然有时提供了动机和良好的起点，但如果可以扩展到现实刺激，通常需要进行大量修改。计算机视觉研究人员不得不离开他们的合成刺激的舒适区域并解决视觉的基本挑战：即视觉系统如何处理现实世界图像的复杂性和重要性，并实现转换进入眼睛的光线的奇迹，或者相机，进入三维物理世界的解释。在需要解决这些问题的驱动下，计算机视觉研究人员开发了大量的数学和计算技术，并使用像Deep Nets这样需要大量注释数据集的工具，从数据中学习理论的重要性。通过预先设定对视觉刺激的实验反应，例如人体在行为实验中的表现，神经元的反应或fMRI活性，可以将相同的技术直接应用于研究生物视觉。

大数据和挖掘数据的学习方法对于视觉尤其重要，因为像Gregory和Marr这样的领导视觉科学家已经开始思考，视觉系统需要以自然和生态约束的形式了解世界。 在格里奥里语中，“错误”不仅仅是对刺激的被动接受，而是一个涉及记忆和其他内部过程的积极过程。  
  
  
换句话说，人类和其他动物的视觉系统利用了通过发展和经验获得的大量知识。 像Deep Nets这样的大数据方法为视觉科学家通过研究现实世界图像的属性来部分学习这些知识提供了一种替代方法。

最后，大数据集的使用对于测试视觉理论也非常重要，因为它们能够与替代理论进行详细比较。他们很容易拒绝利用小数据集和简化刺激所固有的偏见的“理论”。总之，使用Deep Nets和大数据使生物视觉研究人员能够开发和测试可在现实视觉领域中工作的理论，并解决视觉的基本挑战。

#### 5.2真正的神经元和神经回路

从神经科学的角度来看，Deep Nets已被用于预测大脑活动，例如fMRI和其他非侵入性测量，并且有越来越多的例子（Cichy等，2016; Wen等，2017）。它们还被应用于预测神经反应，如通过电生理学测量的，特别是用于预测腹侧流中神经元的反应（Yamins等，2014）。 这些都是Deep Nets能够从数据中学习并处理真实刺激的复杂性的例子。但就理解神经科学而言，最好将其视为一个起点。灵长类动物的腹侧流非常复杂，有证据表明物体和部件的三维结构（Yamane et al。，2008），并且与组合物的经典物体识别理论有关（Biederman，1987） ）在许多方面与标准的深网不同。更一般地，灵长类视觉系统必须执行第3节中列出的所有视觉任务，即边缘检测，双目立体，语义分割，对象分类，场景分类和3D深度估计。 视觉社区为这些任务开发了一系列不同的深网，因此极不可能，例如，对ImageNet上的对象分类进行过训练的Deep Net能够解释灵长类视觉系统的丰富性。

还应该强调的是，虽然Deep Nets在前馈人员中自下而上地进行计算，但是有大量证据表明大脑中存在自上而下的过程（Lee和Mumford，2003），特别是受到自上而下的关注（Gregoriou）等人，2014）。研究人员还发现了皮层回路（Mc Manus等，2011），它们实现了空间相互作用（尽管可能采用自下而上和自上而下的方式）。 这些类型的现象需要其他数学模型，也许是第7节中描述的组合模型。

但是，更重要的是，必须承认深网中使用的人工神经元与大脑中的真实神经元之间存在巨大差异。神经元的人工模型充其量是现实神经元的极大简化，如体外真实神经元的研究所示（Poirazi和Mel，2001）。 Neuroscien tists已经发现有超过一百种不同类型的神经元，并且存在巨大的形态学差异，可以利用它们进行计算（Seung，2012）。对神经回路也缺乏详细的了解。例如，C-elegans的接线图已经有30多年的历史了，但是它仍然只是有限地理解它如何作为神经电路发挥作用（正如O. Hobert所说，接线图就像一条道路地图，告诉你汽车可以开车的地方，但不会告诉你汽车实际开车的时间和地点。理解神经回路还需要了解它们的动力学以及如何根据一系列可能的机制（如快速变化的突触）进行改变（Von Der Malsburg，1994）。了解真实的神经元和真实的神经回路是一项令人着迷的科学挑战和激动人心的工程进步（Boy den et al。，2005）以及巨大数据集的可用性和分析它们的工具意味着必将取得进展。但这些都是极具挑战性的科学任务。总之，真实神经电路和深网中的人工电路之间的跳跃仍然很大，很可能真正的神经电路最终会变得更复杂。

#### 5.3认知能力：深层网和科学理解

很明显，Deep Nets和其他机器学习技术对视觉科学家非常有帮助，但他们怀疑它们是否足以捕捉生物视觉系统的复杂性。人类视觉系统比Deep Nets或其他系统表现得更好  
AI视觉系统，几乎可用于所有视觉任务。少数例外情况是关于进化和经验使人类处于不利地位的情况。例如，人工智能系统可以通过识别数百万张面孔的数量来超越人类，前提是它们在合理的光照条件和有限的遮挡情况下从前面看到，但直到最近，大多数人在其一生中从未见过超过几千人。在阅读计算机断层扫描（CT）图像时，AI系统也可能比普通放射科医生更好，但即使是最专业的放射科医生也只能看到相当少的CT扫描（并且AI系统可以直接访问三个CT扫描中的三维数据，而放射科医师只能查看二维切片）。在每种情况下，人类处于不利地位，因为他们无法访问并因此无法利用大量注释的大数据，这使得Deep Nets能够在这些任务中做得很好。但是，表现优于人类的Deep Nets的真实例子非常罕见（十分之一是由于Deep Nets过度拟合进行研究的数据集）。 此外，人类可以形成各种各样的视觉任务，而当前的AI系统通常专注于单个任务。

此外，认知科学的研究表明，人类视觉系统可以在当前深网无法比拟的抽象层次上工作。这可以通过视觉类比的人类能力来说明，其中一些仅取决于视觉相似性，但其他依赖于部分和子部分的概念，而其他部分包括功能的概念。正如我们将在第7节中论证的那样，这反映了当前机器学习方法的局限性以及当前技术（如Deep Nets）将会出现问题的建议。从另一个角度来看，也可以认为视觉科学的目标是发现潜在的原则。从这个角度来看，一个用无法解释的Deep Net解释现象的模型不会非常令人满意。这是一个值得商榷的问题，合理的人可以不同意这个问题。但我们怀疑AI的进展还需要可解释的模型，部分原因在于实用的工程原理，这对于调试以及性能和安全保障是必要的。

总之，Deep Nets和其他利用大数据的技术是思维和大脑研究人员应该知道如何使用而不是滥用的工具。 但同样清楚的是，目前的Deep Nets未能捕捉到一些最有趣的现象，例如人类执行抽象和执行类比推理的能力（尽管Deep Nets可能会用作构建此类理论的构建块）。 尽管如此，生物学和人工视觉模型之间的密切关系对两个学科都是有益的。 人工智能的研究人员开发了大量的技术工具，如Deep Nets，可以让他们的模型应用于自然图像的复杂性，并在严格的现实条件下进行测试。 视觉科学家可以挑战计算机视觉研究人员开发理论，这些理论在使用比当前计算机更低功率的数量级时，在具有挑战性的情况下可以表现得比人类更好或更好。

### 挑战

本节介绍Deep Nets当前面临的一些挑战以及解决这些挑战的尝试。 其中一些挑战正在逐步被克服，而其他挑战，如对非本地攻击的敏感性，可能需要更多根本性的变化，我们将在第7节中讨论。

#### 6.1放宽对全面监督的需求

Deep Nets的一个缺点是它们通常需要大量带注释的训练数据，这限制了它们在大数据可用的情况下的使用。但情况并非总是如此。特别是，“转移学习”表明，在某些视觉任务的注释数据集上学习的Deep Nets的特征有时可以转移到新的数据集和相关的任务，从而使用更少的数据学习，有时可以用更少的监督。例如，如前所述，Deep Nets首次成功用于ImageNet上的对象分类，但在较小的PASCAL数据集上对象检测失败。这可能是因为PASCAL不足以训练Deep Net，但ImageNet（ImageNet几乎比PASCAL大两个数量级）。但研究人员很快意识到，通过在ImageNet上训练的Deep Net的权重来初始化Deep Net的权重，可以在PASCAL上训练Deep Net用于对象检测和语义分段（Girshick et al。，2014; Long et al。，2015; Chen et al。，2018）。这也引入了一种生成提案的机制，见图1（右下）。

这种在另一个领域学习的Deep Net知识转移的能力与孩子学习的方式直观相关。 一个孩子最初对其他幼小动物的学习速度相当慢，但在关键时期，孩子的学习加速得非常快（Smith和Gasser，2005）。 从模板的“模板”来看，这可能发生，因为在孩子学会识别足够的物体之后他/她可能有足够的构建块（即深层网络过滤器）以能够用a表示新的物体。 现有模板的字典。 如果是这样，可能只需要几个新对象的例子来进行几次学习。

已经为Deep Nets展示了新的对象类别的少量学习，前提是它们首先在大量对象类别上进行了训练（Mao等，2015; Vinyals等，2016; Qiao等，2018）。 另一种策略是训练Deep Net以在对象类别集上学习相似性（技术上是连体网络），从而获得新对象的相似性度量。 例如，Lin等人。 （2017）训练了一个连体网络来学习ShapeNet中对象的相似性（Chang et al。，2015），然后将这种相似性度量用于Tufa数据集中的聚类对象（Salakhutdinov等，2012）。 其他一些镜头学习任务也可以通过使用Deep Nets的功能来完成，这些功能是为某些其他任务而训练的，作为对对象视觉模式进行建模的方法。

最近，有关无监督学习的研究表明，无需详细监督即可学习光学流动和运动结构，而只需要能量函数模型（Ren et al。，2017; Zhou et al。，2017）。 像第三波中的许多神经网络一样，一些基本思想可以在第二波的模糊论文中找到（Smirnakis和Yuille，1995）。 在某些情况下，这甚至可以从单个图像开始学习深度。 其他形式的无监督学习表明，Deep Net的特征可以通过区分scram bled和unscrambles图像（Doersch等，2015），或通过随时间跟踪对象来学习（Wang和Gupta，2015）。

其他研究表明，Deep Nets可以利用大量无监督或弱监督的数据，前提是他们有足够的注释数据。 例如，使用图像来训练对象检测，其中仅知道图像中的对象的名称但是它们的位置和大小是未知的。 这被称为弱监督学习，它可以被视为缺失/隐藏数据问题，可以通过多实例学习（MIL）或期望最大化（EM）等方法来解决。 这些类型的方法的性能通常通过使用少量完全监督的训练数据来改进，这有助于EM或MIL算法收敛到良好的解决方案，例如，参见Papandreou等人。（2015年）。

#### 6.2防御对抗性的例子

Deep Nets的另一个限制来自于研究表明它们可以通过图像的不可察觉的修改成功地进行攻击，从而导致Deep Nets在对象分类方面犯下重大错误（Szegedy et al。，2014），对象检测和语义分割（Xie et al。，2017）（见图4和图5）。这个问题部分地出现是因为数据集是有限的并且仅包含所有可能图像的最终部分。因此，在列车运行图像中任意接近无限多的图像，因此有可能Deep Network会对其中一些图像进行错误分类。研究人员已经证明，他们可以通过白盒攻击，深度网的细节已知，或黑盒攻击，而不是黑盒攻击，找到这样的图像。但现在有防御这些攻击的策略。一种策略是将这些“攻击图像”视为额外的训练数据，称为“对抗性训练”（Goodfellow等，2015; Madry等，2017）。最近的另一种选择（Xie et al。，2018）是在图像中引入小的随机扰动，利用“攻击图像”非常不稳定的假设，因此小的随机扰动将抵抗它们（诚然，Athalye et al。（2018）成功规避了这种防御）。应该承认，可以针对任何视觉算法安装对抗性攻击，并且成功攻击大多数其他视觉算法要容易得多。

#### 6.3解决对上下文的过度敏感问题

对Deep Nets的一个更严峻的挑战是他们对环境的敏感度过高。 图6显示了将吉他放入丛林中的猴子照片中的效果。 这导致Deep Net错误地将猴子识别为人类，并将吉他误解为鸟类，大概是因为猴子比人类更不可能携带吉他，鸟类比吉他更可能在猴子附近的丛林中（Wang） et al。，2018）。 最近的工作给出了很多关于深网过度敏感的例子，例如将大象放在房间里（Rosenfeld等，2018）。

这种对上下文的过度敏感也可以追溯到有限的数据集大小。 对于任何对象，数据集中只会出现有限数量的上下文，因此Deep Net将偏向于它们。 例如，在早期的图像字幕数据集中，观察到长颈鹿只发生在树上，因此生成的标题未提及没有树的图像中的长颈鹿，即使它们是最主要的对象。

当我们考虑Deep Nets的当前限制时，观察到有限大小的数据集是一个共同的主题。回想一下，我们已经提到了如何使用syntic数据，参见图2，以显示在ImageNet上训练的Deep Nets无法识别来自某些视点的对象。合成数据的一个优点是它使我们原则上能够生成无限量的图像，从而系统地探索诸如视点和材料属性等各种因素的影响，例如，参见Qiu和Yuille（2016）; Al corn等。 （2018）。类似地，合成数据可用于系统地改变立体声视觉的危险因素（已知会引起立体算法失败的镜面反射因素;参见图7），使研究人员能够表征立体算法对这些因素的敏感性（Zhang et al。，2018）。因此，合成数据集提供了生成尽可能多的数据的可能性，以便系统地研究深网对诸如视点和光能传递等有害因素的敏感性，这些因素实际上已经到达（提供的合成数据集足够逼真，可以真实地代表真实世界图像）。

捕捉各种各样的背景的困难，以及探索大范围的滋扰因素的需要，对于像Deep Nets这样的数据驱动方法来说是非常有问题的。 似乎确保网络可以处理所有这些问题将需要任意大的数据集，这对训练和测试数据集提出了巨大的挑战。 我们接下来将讨论这些问题。

### 组合爆炸：当大数据集不够时

本节认为，视觉研究人员在应对现实世界数据的复杂性时会面临组合爆炸，以便开发能够在现实世界中对复杂视觉任务进行稳健工作的算法。 在这种情况下，大数据集将不够大，并且开发算法和测试它们将需要新颖的方法。

#### 7.1组合爆炸

Deep Nets在大型数据集上进行训练和评估，这些数据集旨在代表现实世界。但是，正如前面所讨论的，Deep Nets可能无法对他们接受训练的数据集之外的图像进行操作，可能会在数据集中很少发生的罕见事件中出错（但可能会产生灾难性后果，例如跑过婴儿或未能检测到癌症肿瘤），并且对广告攻击和背景变化也很敏感。这些问题都不一定是Deep Nets成功的必然决定因素，而且对于某些视觉领域和任务来说，它们肯定可以克服。但我们认为，这些是一个问题的早期预警信号，随着视觉研究人员尝试使用Deep Nets在无约束的主管中解决日益复杂的视觉任务。也就是说，为了处理现实世界图像的组合复杂性，数据集必须变得指数级大，这显然是不切实际的。

要理解这种组合复杂性，请考虑以下思想实验。想象一下，通过从对象字典中选择对象并将它们放置在不同的配置中来构建可视场景。这显然可以用指数的方式完成。即使对于单个对象的图像，我们也可以获得类似的复杂度，因为它可以以指数的方式被部分地遮挡。我们还可以通过无数种方式更改对象的上下文。虽然人类善于适应视觉环境的变化，但深网更敏感，如图6所示。我们注意到，这种组合爆炸可能不会发生在某些视觉任务中，而深网可能对医疗非常成功。图像应用，因为在上下文中变化相对较小（例如，胰腺总是非常接近十二指肠）。但是对于许多现实世界的应用，尤其是那些使人类在视频中与世界互动的应用，似乎没有指数大的数据集就无法捕捉现实世界的复杂性。

这对当前训练和测试视觉算法的方法带来了巨大挑战。这些方法是由机器学习研究人员开发的，以确保算法捕获数据的基础结构，而不仅仅是记忆训练数据。他们假设训练和测试数据是从一些未知概率分布中随机抽取的样本。但关键的是，数据集需要足够大才能代表数据的非分布式。有趣的是，据我们所知，机器学习基础的研究人员从未直接解决过这个问题。相反，他们专注于理论结果，称为可能近似正确（PAC）定理，它给出机器学习算法已经学习基础数据结构的概率的界限，其关键见解是训练数据的数量必须比学习算法在看到数据之前可以考虑的假设集大得多（Valiant，1984; Vapnik，1998; Poggio和Smale，2003）。但是，无论如何，如果图像集是组合大的，那么在有限数量的随机抽取样本上训练和测试模型的标准范例变得不可行。 这迫使我们解决两个新问题：（I）我们如何在有限大小的数据集上训练算法，以便它们能够在捕获现实世界的组合复杂性所需的真正庞大的数据集上表现良好？ （II）如果我们只能在有限的子集上测试它们，我们如何有效地测试这些算法以确保它们在这些庞大的数据集中工作？

从计算机图形学的角度考虑这些问题是有帮助的。 直接（参见图8）指定具有13个参数的计算机程序，其可以从不同视点，在不同照明下以及在有限数量的背景场景中渲染单个对象的图像。 如果我们为每个参数允许1,000个不同的值，我们总共获得1039个不同的图像，比任何现有数据集大1030个数量级。 该程序可以扩展到包括大量视觉场景中的多个对象，并且原则上，我们可以指定具有有限但非常大量的参数的模型，这些参数可以生成组合大量的真实图像， 近似于现实世界。 但是，虽然这提供了一种可能生成所有真实世界图像的方法，但它并没有解决如何在这些数据集上训练和测试模型的问题。

#### 7.2克服组合复杂性的模型

像目前形式的Deep Nets这样的方法似乎不太可能处理组合式爆炸。 数据集可能永远不会大到足以训练或测试它们。 在这里，我们描绘了我们认为相关的想法类型。 我们可以从面向并克服这些挑战的人类视觉系统中获得一些指导。 人类每年看到大约109张图像（假设每秒30张图像）这是很大的，但不是组合的。 但是，超过一个关键时代的人类可以从少数例子中学习，感知三维结构，处理抽象，在有用时可以利用上下文，但在没有时却忽略它。 最近的实验（Ullman等，2016）表明人类可以明确地解释图像，只要它们高于临界尺寸（取决于图像内容），并且不需要额外的背景。

组合性可能是解决方案的一部分。这是一个普遍的原则，可以被诗意地描述为信仰的一个体现，即世界是可知的，人们可以将事物分开，预先加入它们，并且可以随心所欲地重新组合它们。关键的假设是结构是根据一组语法规则，从更基本的子结构中分层次地组成。这表明子结构和语法可以从有限数量的数据中学习，但会推广到com二进制情境。与Deep Nets不同，组合模型需要结构化表示，这些结构表示其结构和子结构，使其能够使用相同的表示来执行多个任务（例如，检测对象，对象部分和对象边界）（Chen et al。，2007） ）（有人认为Deep Nets是成分，但这是一个非常不同的意义）。组合模型能够推断出他们所见过的数据，推理系统，进行干预，进行诊断，并根据相同的知识结构转换许多不同的问题（Pearl，2009）。引用斯图尔特·杰曼的话说，世界是成分的，也可能是神的前身，因为否则上帝似乎有必要手工制造人类的智慧（Geman，2007）。

组合性与模式理论和合成分析密切相关（Grenander，1993; Mumford，1994; Tu等，2003; Zhu和Mumford，2006; Mum ford和Desolneux，2010）。它可以通过玩具世界的例子来说明，如图9所示，其中图像是根据元素组成的基本词汇表创建的。这三个小组显示了从左到右复杂程度越来越高的微观世界。对于每个微观世界，存在一种语法，其指定由元素组成构成的可能图像。在左侧面板中，基本组件是不重叠的字母，因此在解释图像时很容易。中间和右侧面板由更复杂的语法生成，这些语法可以是彼此严重遮挡的不同字体，条形和片段。解释这些图像要困难得多，并且似乎需要这样一种观念，即字母由基本部分组成，它们可以以各种字体出现，以及“取而代之”的概念。（解释一封信的部分内容因为被另一封信封闭而丢失）。

图9中的第三个微观世界是组合大数据集的一个例子，因为图像是通过从字典中选择对象并随机放置它们同时允许遮挡来进行的。这个微观世界与CAPTCHA基本相同，可以用来区分人类和机器人。有趣的是，关于CAPTCHAs的工作（George et al。，2017）表明，在基本代币的组成和分解几何和外观方面代表对象的成分模型可以在这些类型的数据集上表现良好。他们的推理算法包括自下而上和自上而下的处理（Tu et al。，2003），它使算法能够“解释”字母的缺失部分并强加解释的“全局一致性”以消除歧义。直观地说，部分探测器为字母提供自下而上的建议，这些建议可以在上下游阶段进行验证或拒绝。相比之下，Deep Nets在这些数据集上表现更差。大概是因为，与成分模型不同，它们无法捕捉到域和外推的低位生成结构 - 在训练数据集之外。由于微观世界是组合大的，因此不可能在深度网上训练足够的数据以保证整个数据集的良好性能。其他理论研究，例如Yuille和Mottaghi（2016），表明组合模型非常适合通过共享零件和使用分层抽象来处理复杂的事物。

其他非视觉示例说明了相同的观点。 最近的一个例子是研究人员（Santoro等，2018）试图训练标准的深网进行智商测试。 该任务需要在3×3网格中的8个给定图像内找到有意义的规则/模式的组合（可能存在干扰物），并且目标是填充最后丢失的图像。 毫不奇怪，Deep Nets并不能很好地概括。 对于自然语言应用，神经模块网络（Andreas et al。，2016）比静态的固定结构深网更有前途，因为动态架构布局可能足够灵活，可以捕获一些有意义的组合。 事实上，我们最近验证了各个模块在联合训练后确实执行了他们想要的功能（例如AND，OR，Filter（红色）等）（Liu et al。，2019）。

组合模型具有许多理想的理论属性，例如可解释性，以及生成的能力，因此可以从中进行抽样。 这意味着，原则上，他们知道关于对象（或任何正在建模的实体）的一切，这使得它们比像Deep Nets这样的黑盒子方法更容易诊断，因此更难以愚弄。 但是学习组合模型很难，因为它需要学习构建块和语法（甚至语法的本质也是有争议的）。 然而，从边缘等基本的基本标记开始学习分层词典方面取得了一些有限的成功（Zhu et al。，2010）：见图10。

组成模型的当前限制是，为了通过合成进行分析，它们需要具有对象和场景结构的生成模型。 将分布放在图像上是一项挑战，除了面部，字母和常规文本等少数例外（Tu et al。，2003）。 但是从两个方向取得了可喜的进展。 首先，计算机图形模型变得越来越逼真，视觉外观可以大致考虑到几何，纹理和照明。 回想一下，1039幅图像（图8）仅由13个参数生成。 另外，Deep Nets也被应用于使用生成对抗网络（GAN）生成图像。从综合分析的角度来看，GAN的结果令人失望，尽管最近有条件的GAN的工作显示出了希望。

更重要的是，处理组合式爆炸需要学习3D世界的因果模型以及这些模型如何生成图像。 对人类婴儿的研究表明，他们通过制作预测环境结构的因果模型来学习，包括天真的物理学。 这种因果理解使得能够从有限数量的数据中学习并对新情况进行真正的推广。 这是对牛顿法的对比，它通过最少量的自由参数给出了因果理解，太阳系的托勒密模型提供了非常准确的预测，但需要大量的数据来确定其细节（即本轮）。

#### 7.3数据是组合时的测试模型

如果我们只能在有限数量的数据上测试它们，我们如何测试视觉算法来处理现实世界的复杂性？ 如果我们具有结构良好的模型，例如，如上所述的组成模型，那么我们可以利用模型的结构来确定它们的失效模式。 当然，这类似于通过系统地识别其弱点来测试复杂工程（例如，飞机）或软件结构的方式。 这更像是游戏理论，而不是决策理论（主要关注平均损失，这是机器学习理论的基础），因为它建议关注最坏的情况，而不是平均情况。 如果目标是为自动驾驶汽车开发视觉算法，或者在医学图像中诊断癌症，那么这是有道理的，其中算法的失败会产生重大后果。

如果可以识别视觉任务的故障模式并且维度较低，则可以这样做。 例如，如前面第6.3节所述，研究人员已经分离出导致立体算法失败的危险因素，包括镜面反射和无纹理区域。 在这种情况下，可以开发计算机图形系统地改变这些危险因素，以确定哪些算法对它们有抵抗力（Zhang et al。，2018）。 简而言之，我们可以沿着这些特定的尺寸对这些算法进行压力测试。

但是对于大多数视觉任务来说，很难识别出可以进行隔离和进一步测试的少量危险因素。 相反，我们应该将对抗性攻击的概念概括为包括非本地结构。 一个简单的可能性是允许其他更复杂的操作，这些操作会导致图像或场景发生合理的变化，例如，通过遮挡或改变被观察物体的物理特性（Zeng et al。，2017），但不会显着影响人类 知觉。

### 结论

这个观点的动机是关于深网与许多不同学科的研究人员的讨论。 我们试图取得平衡，承认深网的巨大成功，但不会被围绕他们的流行激情所淹没。 我们经常使用我们自己小组的工作来说明我们的一些要点，并向其他作者道歉，他们的作品我们会在该领域的更多学术评论中引用。 我们关注的几个问题与最近对Deep Nets的批评中提到的问题相似（Darwiche，2018; Marcus，2018）。

几年前，Aude Oliva和第一作者合作组织了一次由NSF赞助的计算机视觉前沿研讨会（MIT CSAIL，2011年8月21日至24日）。会议鼓励坦诚交换意见，特别是对于深网对计算机视觉的潜力存在巨大分歧。但几年后，正如Yann LeCun预测的那样，每个人都在使用Deep Nets。他们的成功是前所未有的，并且帮助视觉变得更加广为人知，大大增加了学术界和工业界之间的互动，导致视觉技术应用于广泛的学科，并产生许多其他重要的后果。但是，尽管取得了成功，但在实现一般目的人工智能和理解生物视觉系统的目标之前，仍然存在巨大的挑战。特别是当研究人员在现实世界条件下解决复杂的视觉任务时，处理组合爆炸。虽然Deep Nets和其他大数据方法肯定会成为解决方案的一部分，但我们相信我们还需要可以在其成功和见解的基础上建立互补的方法

致谢这项工作得到了大脑，思想和机器中心（CBMM）的支持，由NSF STC奖CCF-1231216和ONR N00014-15-1-2356资助。 我们感谢Kyle Rawlins和Tal Linzen提供反馈。