深度学习常用网络及模型综述

软件学院 罗华坤 2019011799

摘 要：深度学习概念于2006年提出，此前已有众多网络问世，例如多层感知机（MLP），卷积神经网络（CNN），循环神经网络（RNN）等。近年来，Transformer，GAN等模型也相继被提出，用于减少参数、生成数据、提升性能等。该文在网络及模型层面按时间顺序进行对比分析，整理深度学习常用网络及模型的异同、优势与不足之处，对进一步了解深度学习领域有一定的参考价值。

关键词：神经网络，CNN，RNN，Transformer

**Summary of commonly used networks and models for deep learning**

**Abstract:** The concept of deep learning was proposed in 2006, and many networks have been introduced before, such as multi-layer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), etc. In recent years, models such as Transformer and GAN have also been proposed to reduce parameters, generate data, and improve performance. This article compares and analyzes the network and model levels in chronological order, sorts out the similarities and differences between commonly used networks and models in deep learning, and has a certain reference value for further understanding of the field of deep learning.

**Key words:** neural network; CNN; RNN; Transformer

**一、引言**

如今，深度学习技术广泛应用到了生活的方方面面，从验证身份所需的人脸识别系统，到日常交流学习要用的翻译工具，都很大程度上使用了深度学习技术。

在深入学习该领域之前，对深度学习发展的大致了解是十分必要的。本文归纳对比分析各个常见网络模型，按照时间顺序，从最早的MLP，CNN, RNN，到近年提出的transformer，介绍异同、优势不足以及应用方向等，最后对该领域提出一些愿景。

二、多层感知机（MLP）

启发于神经科学的突触结构，人们创建了特征神经元，可以对上层数据线性组合，经过非线性变换，如sigmoid函数，后再输出到下一层。最后通过设计损失函数的方式来修正模型参数。

2.1 传统思路下的感知机

起初感知机只有一层，且没有引入非线性函数，因此只能处理线性可分问题，连最简单的异或都无法划分。后来尽管人们尝试增加层数，忽略庞大的计算量，如何更新网络中的参数成为了一个更大的问题。

2.2 改进后的多层感知机

为了解决更新多层网络中的参数问题，Hinton于1986年提出了反向传播算法（有所争议），其在后来所有网络中成为了必要环节。并且为了减少随着网络层数增加导致的过拟合问题，引入了正则项来平滑曲线，减少过拟合现象。尽管如此，简单的多层感知机仍无法适应随之而来更复杂的需求，因此人们进一步改进，将其变成一个模块融合使用。[1]

2.3 自编码器Autoencoder

为了实现数据降维及重建，人们引入了两个mlp模块，一个encoder，另一个为decoder。Encoder负责将数据压缩成一个低维向量，decoder则将该低维向量重构成高维数据。在该过程中，同样加入了正则项以及在输入增加扰动以防止过拟合，还引入了KL散度，使用交叉熵作为损失函数提高性能。

基于上述功能，其目前在特征提取、文本检索分类、数据去噪等领域发挥了重大作用。

2.4 存在的优势与不足之处

mlp模型较为简单，容易实现并很好地完成一些简单的任务，例如手写字体的识别等。但当数据大小变大，mlp的参数量与计算量都大幅增加，这也决定了mlp在应用中更多作为一个小的模块参与。

三、卷积神经网络CNN

同样受神经科学启发，LeCun等人于1988年提出了卷积神经网络，引入了离散的卷积操作与池化操作。每个神经元只能看到一定范围的数据并保持空间上的位置信息，越深层的神经元可见范围越大，实现了逐层提取特征的功能。

对比起mlp需要将数据展成一维向量，容易丢失空间位置信息，卷积网络保留了空间信息，因此对图像等信息的处理效果更好。相对于mlp，CNN的结构使得反向传播的计算更为复杂。

3.1 LeNet

LeNet作为首个被提出的模型使用了较大的卷积核提取图片特征，使用sigmoid作为激活函数。但由于sigmoid函数的饱和区过大，在饱和区内梯度接近0，容易发生梯度消失现象，因此效果不显著，ImageNet的top1正确率仅有50%，并被后来的支持向量机超越，深度学习领域开始走入寒冬。[2]

3.2 AlexNet

二十年后，人们在AlexNet中引入了relu函数，无论是在梯度性质还是在计算难度上都优于sigmoid函数。并且在训练时采用dropout操作防止数据过拟合，但最终的期望保持不变。同时使用数据增广操作，不仅增加了数据量也一定程度缓解了过拟合现象。[3]

最重要的是，由于硬件领域的发展，在AlexNet首次使用了GPU加速，开启了GPU并行计算的先河。

3.3 GoogLeNet

在模型泛化领域，Telgarsky与Zhou Lu等人证明了神经网络越深越宽效果越好。[4][5]因此GoogLeNet选择了更小的卷积核，在减少参数量的同时增加了网络的深度。不仅如此，其还将一个卷积操作改写成了多路并行的卷积操作，增加了网络每层的宽度，还有利于GPU的并行计算。在该规模的模型中GoogLeNet的参数更少，但在ImageNet上却有了大幅度的提升并且可以投入工业使用。[6]

似乎上述结论已经确立了今后CNN要朝着更深更宽的道路上发展，然而实验发现，当层数增加到一定数目后，训练集中错误率不降反升，即发生了模型的退化。尽管人们提出批归一化操作来解决，但也仅仅解决了由于深度增加造成的梯度消失问题。

3.4 ResNet

HeKaiMing等人于2016年引入了残差网络，认为残差学习更为容易、神经网络不擅长拟合恒等变换，因此将上层的输入与下层的输出直接相连，形成shortcut，中间设计Residual模块。这一操作也使得更深的网络成为可能，将发生退化的层数从50多增加到了152层。而且在ImageNet竞赛上也首次低过人类的错误率。[7]

3.5 DenseNet

由于ResNet仅在特定区域使用shortCut，不难想到如果采用全连接的方式后效果会更好。随后HuangG等人因此提出了DenseNet，在具体的全连接实现上采取了矩阵连接的方式，而不是像ResNet在输出处相加，于是大幅减少了参数的使用，更加强调了底层数据，imageNet的正确率再次得到提升。同时，为了控制全连接层导致的通道数爆炸，引入了过渡层来降低通道数。[8]

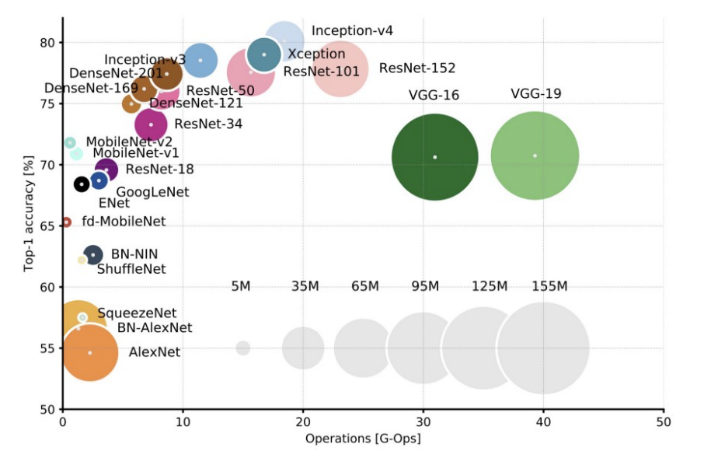


图 1 各模型ImageNet的top1正确率[9]

3.6 应用、优势与不足

从上述例子中可以看出，CNN的发展有多条路径，提高深度、增加通路与模块，以及后序提出的spatial transformer、Fast-CNN还增加了空间变换、框选检测的能力。不同的方向允许CNN用在更加广泛的领域，现在CNN还被用于做图片的风格化、超分辨以及生成工作。[10]

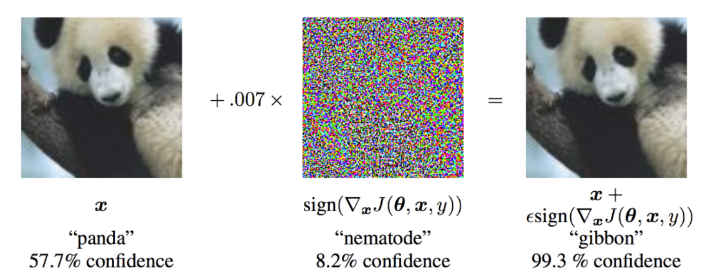
CNN采用的卷积核使得人们无需关注像素与像素之间的关系但仍保留了内部空间连接，这让CNN在图片分类检测领域表现效果优越。然而，这也为CNN蒙上了一层黑布，隐藏了内部机理。

图 2 对抗生成样本FGSM[11]

尽管声称灵感来源于神经科学，对抗性样本的出现说明了卷积神经网络仍不是以人类的角度来观测一张图片且存在一定的不安全性。此外，池化层也破坏了整体与局部的关联，经常出现将人脸的眼镜与嘴巴对调仍能识别成功的例子。

四、 循环神经网络RNN

现代社会每日流通着大量的文本信息，如何理解甚至预测这些文本成了一大需求。尽管可以采用一维的卷积核来实现，但CNN在这种问题的表现往往不太好。这与文本序列长度不确定、上下文的依赖关系或近或远等原因有关。因此一种具有时序性的网络被首次提出，即RNN。

4.1 基本的RNN（vanilla RNNs）

模型按照序列的长度分为一个个小单元，每个单元输入包括此时的序列输入和上个单元的隐藏信息，放入一个mlp层训练得到此刻的输出与传到下层的隐藏信息。和之前的网络一样，RNN也经常面临着过拟合问题。

由于训练输入的都是真实的数据，而在测试时输入采用的是上个单元的输出，这种训练方式（teaching forcing）往往会过拟合，从而实际表现不佳。于是人们引入curriculum learning，即从真实输入逐渐转化到测试模式，来缓解过拟合现象。

4.2 LSTM

分析RNN的反向传播算法可以发现，需要在时间维度上进行连乘，因此长序列十分容易发生梯度消失现象，无法利用过去长时间的信息。而LSTM引入了记忆细胞，在隐藏信息传递中建立shortcut。

LSTM包括记忆细胞与三个控制门，分别包括旧记忆的遗忘、新记忆的添加、生成下一个隐藏状态，经过一系列的公式运算，在输入、隐藏状态、记忆信息之间建立联系，减少梯度的连乘，解决了梯度消失问题。[12]

4.3 GRU及更多类似网络

GRU其实是对LSTM的一点改进，包括取消了记忆细胞部分，直接利用三个门对隐藏状态进行重置、输入操作，使得整个模型的参数更少、运算速度更快。[13]

不仅如此，为了适应文本的上下文有关性、增强隐藏信息的表达能力，分层RNN，双向RNN都被相继提出，使得更多信息在单元与单元之间流通，预测能力更强。

4.4 Seq2Seq

机器翻译是文本序列分析一个重要问题。为了解决这一类问题，Seq2Seq模型被提出。类似autoencoder，Seq2Seq同样包含encoder与decoder。但与AE不同的是，encoder与decoder是两个RNN的模块。Encoder将输入统一编码成一个隐藏状态，再输入decoder通过一个个单元输出对应的翻译。[14]

但不难发现，一个隐藏状态难以完全表达全文各词之间的关系，因此需要更多的信息。后来attention机制得以运用，将encoder每个单元增加输出，最后再与decoder的输出做内积，经过softmax函数后得到最终结果。如今也常被用于自动回复。

4.5 应用、优势与不足

RNN在文本预测、机器翻译等领域有着优异表现，在数据处理后还可以做语音识别、合成音乐等工作。RNN运用了上下文的隐藏关系，将隐藏关系交给机器去学习。并且如今也与CNN联合产生了许多有趣的应用，例如输入图片输出对应描述，处理视频这种更加复杂的数据等。

然而，和CNN类似，这种隐藏信息的传递仍是十分不透明的，且当序列更长时，表现也不足。

五、 深度学习新模块——Transformer

启发于attention机制，google提出了一种全新的模块——transformer，采用全attention的机制替代了原来的LSTM或GRU来解决序列问题。同样地，由encoder与decoder构成。[15]

5.1 encoder

Encoder的输入需要先将原始输入进行编码至一定维度，再添加位置信息。而encoder内部包括一个multi-head attention模块与前馈网络部分，中间引入了残差操作与归一化部分。、

多头attention模块通过三个矩阵将输入变为三个向量Q（Query）、K（key）、V（value），利用self-attention机制，对每个Q进行一次attention操作后输出结果。经过前馈网络后输出K, V给decoder部分。

5.2 decoder

Decoder层也是接受类似encoder层的输入，也要经过embedding与positional embedding。但与encoder不同的是，为了防止过拟合，此处引入了masked multi-head attention。即每个词只能看到前文。接着输出再与encoder的输入结合放入multi-head attention模块，最后得到结果。

5.3 存在的优势与不足之处

在multi-head attention以及前馈网络这两个模块中，每个向量都可以拆分进行并行化操作，因此计算效率比传统的RNN要高得多。并且发现，由于encoder的输入到每一个decoder中，因此容易建立更加长久的依赖。

并且由于attention本身的解释性相较于传统的RNN强，因此transformer也具有较强的可解释性，这一点从过程的可视化可以看出。该优势也被计算机视觉领域运用，其中VIT模型（2021）就是一个典型。VIT将图片切块，看成文本序列编码输入，发现在大数据下效果超过SOTA。Transformer首次连接nlp与cv问题，其在cv领域的研究运用如今十分火热。[16]

但研究显示，transformer并非图灵完备，因此其并不擅长预测简单重复的文本，其表现甚至比传统的还要差。Embedding这步不仅丢失了位置信息，后加入的位置信息嵌入也只是权宜之计，还缺少类似CNN具有的“平移不变性”。

六、 deep Learning领域未来窥探

除了上述提到的众多网络之外，还有图神经网络（GNN）、生成对抗网络（GAN）、深度强化学习等。在庆祝深度学习繁荣发展之余，我们应当看到中间的一些隐忧与不足。

6.1 正确率不足

此处的正确率并非指的是在实验室中测试得到的正确率，而是真实投入生产中的得到的。现实世界中的问题往往更加复杂，影响因素众多，实验得到的模型在现实中正确率很难达到工业级别要求。且现实问题影响严重，自动驾驶的一个错误很有可能导致数条生命的逝去，医疗诊断的一次误判延误的是病人的治疗时间。

6.2 落地能力不足

尽管如今各大公司动用所有显卡来训练一个模型，但最终这些模型都会到每个用户的手中。如此大的参数并不是每个设备都能储存的，而且计算量也不是每个设备能够承受的，因此需要更加精简的模型才能用于实践中去。

6.3 小样本学习能力不足

模型的训练需要大量的数据，对于某些领域来说数据十分稀缺，因此小样本学习成为了研究热门。但如今的小样本学习与机器学习的先验知识要求冲突，因为即使是人的学习过程也是综合了以往的知识才能达到小样本的学习，所以该领域需要进一步的研究探索。

6.4 缺乏逻辑推理能力

不难发现深度学习做的比较好的都是分类、回归、预测领域，然而在逻辑推理领域深度学习仍无重要成果。这与深度学习内部机理有关，如今所谓的逻辑推理也只是浅层上的概率选择，并非深层的逻辑思考，这也是强人工智能与弱人工智能的区别。完善这一点可以进一步达成强人工智能。

七、 总结

上文对深度学习的主要网络与模型进行了介绍，在内部对每个网络按照时间顺序比较异同。还简要分析介绍了近期深度学习新出的模型transformer。在最后提出了对目前领域的发展的愿景，尽管目前该领域仍存在着一些不足，其未来的发展潜力及潜在价值仍是不可估量的。

参考文献（References）

1. Rumelhart D E, Hinton G E, Williams R J. Learning Representations by Back-Propagating Errors[J]. Nature (London), 1986, 323(6088): 533–536. DOI:10.1038/323533a0.
2. Lecun Y, Bottou L, Bengio Y, et al. Gradient-Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324. DOI:10.1109/5.726791.
3. Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Communications of the ACM, 2017, 60(6): 84–90. DOI:10.1145/3065386.
4. Telgarsky M. Benefits of Depth in Neural Networks[J]. 2016. .
5. Lu Z, Pu H, Wang F, et al. The Expressive Power of Neural Networks: A View from the Width[J]. 2017.
6. Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions[J]. 2014.
7. Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 770–778. DOI:10.1109/CVPR.2016.90.
8. Huang G, Liu Z, Van Der Maaten L, et al. Densely Connected Convolutional Networks[J]. 2016.
9. Canziani A, Paszke A, Culurciello E. An Analysis of Deep Neural Network Models for Practical Applications[J]. 2016. .
10. Girshick R. Fast R-CNN[J]. 2015.
11. Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[J]. 2014. .
12. Hochreiter S, Schmidhuber J. Long short-term memory.[J]. Neural Computation, 1997, 9(8):1735-1780.
13. Chung J, Gulcehre C, Cho K等. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling[J]. 2014.
14. Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks[J]. 2014.
15. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. 2017.
16. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. 2020.