关于动态神经网络的简单介绍

胡海

**摘要**：随着当前计算的数据越来越庞大，其对卷积网络的计算能力提出很大考验。较传统的轻量级卷积神经网络由于其较低的计算预算限制了卷积神经网络的深度(卷积层数)和宽度(通道数)，导致其表示能力有限，从而导致性能下降。为解决传统卷积神经网络在计算能力和表示能力上的不足，已提出了一系列的优化方法，包括动态卷积、条件参数化卷积、动态神经网络、动态区域感应卷积来提高神经网络的表示能力以及计算能力。但由于深度神经网络模型变得越来越庞大和复杂，由于计算需求的增加，维持模型规模的增长越来越具有挑战性，又对动态神经网络进行了优化以满足对计算能力的需求。本文就神经网络进行简单的介绍。

**关键词**：卷积神经网络、动态卷积、优化方法

**A brief introduction to dynamic neural networks**

Hai Hu

Abstract: As the data currently being computed becomes more and more massive, it puts the computational power of convolutional networks to a great test. The more traditional lightweight convolutional neural networks have limited representation capability due to their lower computational budget limiting the depth (number of convolutional layers) and width (number of channels) of the convolutional neural network, leading to performance degradation. To address the shortcomings of traditional convolutional neural networks in terms of computational and representational capabilities, a series of optimisation methods have been proposed, including dynamic convolution, conditional parameterised convolution, dynamic neural networks, and dynamic region-sensing convolution to improve the representational as well as computational capabilities of neural networks. However, as deep neural network models have become larger and more complex, it has become increasingly challenging to maintain the growth in model size due to increased computational demands, and additional optimisations have been made to dynamic neural networks to meet the demand for computational power. This paper provides a brief introduction to neural networks.

Key words: convolutional neural network; dynamic convolution; optimisation methods

1.引言

随着社会的发展及研究的越来越深入，数据量越来越庞大，对卷积神经网络的计算能力提出了更高的要求。轻量级卷积神经网络由于其较低的计算预算限制了卷积神经网络的深度和宽度，导致其表示能力下降和性能降低。随计算机的发展越来越快，计算机发展的分支也越来越多，计算机视觉就是其中一个重要的部分，深度神经网络在计算机视觉中起着重要作用。而卷积层是深度神经网络的基本组成成分之一，其基本假设为相同的卷积核应用于数据集中的每个例子。为了增加模型的容量，模型开发人员通常会添加更多的卷积层或增加现有卷积的大小（内核高度/宽度、输入/输出通道的数量）。在任何一种情况下，额外容量的计算成本都与卷积输入的大小成比例地增加，基于此种情况一旦数据量庞大，其计算成本也就越庞大。此外神经网络在语义分割上也有应用，经典的人工设计网络打算在单个网络体系结构中表示所有实例，这缺乏对现实环境中各种规模分布的适应性。

为解决上述神经网络的计算能力及表达能力上的不足，提出了动态卷积神经网络的概念，在计算能力方面，动态卷积可以在不增加网络深度或宽度的情况下增加模型的复杂性，动态卷积不是每层使用一个卷积核，而是基于多个并行卷积核的关注动态聚合。它不会增加网络的深度或宽度，而是通过注意力聚合多个卷积核来增加模型能力，因此其不会使计算成本大幅度提高。在语义分割方面，动态路由在推理过程中生成与数据相关的转发路径，这意味着特定的网络架构随输入而变化。该方法可以将不同尺度的实例(或背景)分配到相应的解析阶段进行自定义特征转换。不同尺度分布的输入图像会选择不同的路径进行特征变换。已有一些动态网络研究通过修剪通道来实现高效的目标识别。与之不同的是，本研究侧重于语义表示，旨在缓解规模差异，提高网络效率。

2.发展现状

动态卷积解决了轻量级卷积神经网络因计算预算不足所造成的表示能力下降及性降低的问题。这是一种新的设计，可以在不增加网络深度或宽度的情况下增加模型的复杂性。动态卷积不是每层使用一个卷积核，而是基于多个并行卷积核的关注动态聚合，这些卷积核是输入相关的。装配多个核不仅由于核尺寸小而计算效率高，而且由于这些核通过注意力以非线性方式聚集而具有更强的表示能力。通过简单地对最先进的架构MobileNetV3-Small使用动态卷积，ImageNet分类的前1精度提高了2.9%，仅增加了4%的FLOPs。动态卷积神经网络也更难训练，它需要所有卷积核的联合优化和多层的注意力。为此提出了有效联合优化的两个关键：（1）约束输出注意力，以方便输出注意力模型的学习；（2）在训练早期将注意力平坦化，以方便卷积核的学习。

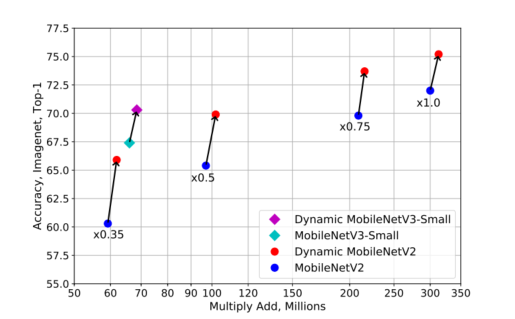


图1计算成本和ImageNet分类精度之间的权衡

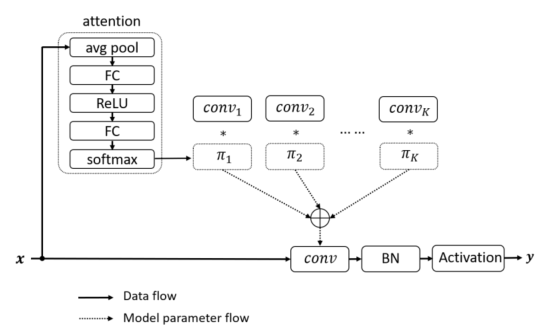


图2动态卷积层

为提高深度卷积神经网络在计算机视觉方面的性能及表现能力，提出了条件参数化卷积的概念，它为每个例子学习专门的卷积核。一般来说，深度卷积神经网络在计算机视觉任务上性能的提高很大程度来自于增加模型大小和扩展到越来越大的数据集的能力。但是，在当前增加增加模型容量的方法在计算上是昂贵的。部署性能好的推理模型可能会消耗大量的世数据中心容量，并且对于具有严格延迟限制的应用程序来说往往是不可行的。用条件参数化卷积代替普通卷积使我们能够增加网络的大小和容量，同时保持高效的推理。我们证明，具有条件参数化卷积的缩放网络提高了几种现有卷积神经网络架构在分类和检测任务上的性能和推理成本权衡。在ImageNet分类方面，我们将条件参数化卷积方法应用于EfficientNet-B0，仅需413M次乘法运算即可实现78.3%的准确率。在条件计算模型中这是通过对每个例子仅激活整个网络的一部分来实现，一种条件计算方法使用强化学习或进化方法来学习离散路由函数。

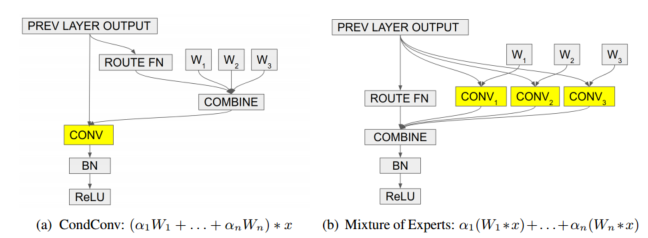


图3 （a）条件参数化卷积架构图 （b）混合方法图

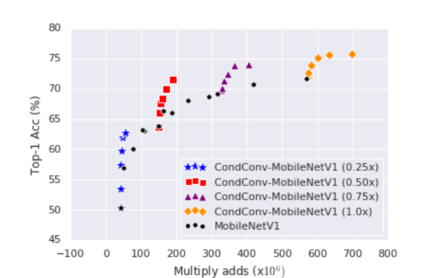


图4 不同宽度模型性能图

为使卷积神经网络在语义分割中发挥更好的性能及表示能力，提出了一种概念性的新方法来缓解语义表示中的尺度差异即动态路由。该框架生成与数据相关的路径，以适应每张图像的尺度分布。为此，提出了一种可微门控函数，称为软条件门，用于动态选择尺度变换路径。此外，通过对门控函数进行预算约束，可以以端到端方式进一步降低计算成本。我们进一步放宽了网络级路由空间，以支持每个forward中的多径传播和跳过连接，带来了可观的网络容量。动态路由在推理过程中生成与数据相关的转发路径，这意味着特定的网络架构随输入而变化。该方法可以将不同尺度的实例(或背景)分配到相应的解析阶段进行自定义特征转换。不同尺度分布的输入图像会选择不同的路径进行特征变换。已有一些动态网络研究通过删除块或修剪通道来实现高效的目标识别。使用动态路由，可以较容易的实例化语义分割。

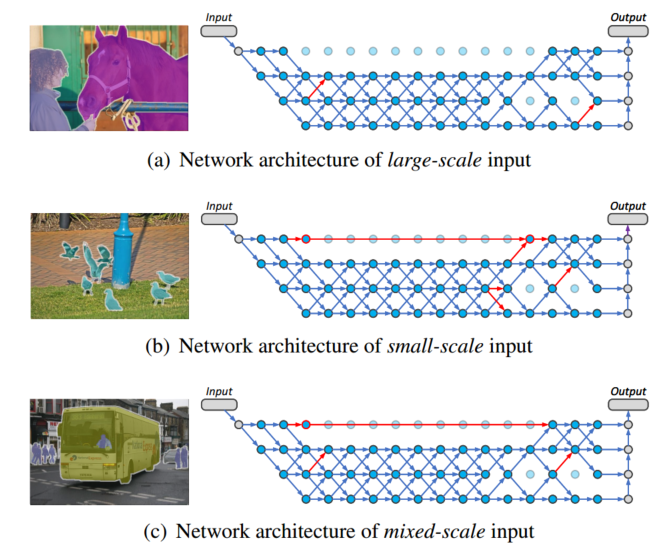


图5 不同规模输入动态路由的转发路径 （a）大规模输入 （b）小规模输入 （c）混合规模输入

随着深度神经网络模型变得越来越庞大和复杂，由于计算需求的增加，维持模型规模的增长越来越具有挑战性。其局限性是整个网络的静态激活，而不管输入是什么，这比人类大脑可以动态地、零星地激活与感知信息相关的神经元的效率要低得多。因此，机器学习研究人员已经做出了许多努力来设计动态神经网络，这些神经网络可以在推理期间将输入馈送到不同的子神经结构或大型模型的参数中。而动态神经网络显示出良好的特性，包括效率、适应性、通用性和可解释性。

不幸的是，现有的深度学习(DL)框架还不能有效地运行动态神经网络。它们的优化主要集中在静态神经网络上，其算子的执行顺序对所有输入都是确定的。在一般程序的编译器中，利用程序的运行时特性来动态优化其执行已经被广泛研究。通过分析动态运行时的概况，发现许多动态神经网络由于其分支激活的不均匀分布而具有相似的机会，可以进一步利用这些机会进行动态优化。现有的研究提出了用Brainstorm（与人大脑皮层相似的方式）来优化动态神经网络。

Brainstorm: 用于优化动态神经网络的深度学习框架，它通过统一动态应该如何表达来弥合差距。头脑风暴提出(1)Cell，关键的数据抽象，允许模型开发人员表达动态存在的数据粒度（数据粒度是指数据仓库中数据的详细程度和精度）;(2)Router，一个统一的接口，让模型开发人员表达应该如何动态分配Cell。头脑风暴处理路由行动的有效执行。这种设计允许Brainstorm以正确的粒度收集细粒度数据流的概要文件。可追溯性进一步为动态神经网络的动态优化开辟了新的空间，使其执行专门化到运行时动态分布。广泛的评估表明，脑力风暴带来了11.7倍的加速(平均3.29倍)或导致42%的内存消耗减少流行的动态神经网络与提出的动态优化。

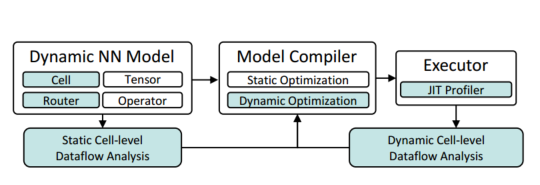


图6 Brainstorm架构图

**参考文献**

[1] Cui, W., Han, Z., Ouyang, L., Wang, Y., Zheng, N., Ma, L., ... & Guo, M. (2023). Optimizing dynamic neural networks with brainstorm. In 17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23) (pp. 797-815).

[2] Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., & Liu, Z. (2020). Dynamic convolution: Attention over convolution kernels. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11030-11039).

[3] Yang, B., Bender, G., Le, Q. V., & Ngiam, J. (2019). Condconv: Conditionally parameterized convolutions for efficient inference. Advances in neural information processing systems, 32.

[4] Zhou, Y., Ren, T., Zhu, C., Sun, X., Liu, J., Ding, X., ... & Ji, R. (2021). Trar: Routing the attention spans in transformer for visual question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2074-2084).

[5] Li, Y., Song, L., Chen, Y., Li, Z., Zhang, X., Wang, X., & Sun, J. (2020). Learning dynamic routing for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8553-8562).

[6] Chen, J., Wang, X., Guo, Z., Zhang, X., & Sun, J. (2021). Dynamic region-aware convolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8064-8073).

附录

问题1：动态神经网络与静态神经网络的区别

答：静态神经网络是一种基于固定输入和输出的神经网络结构。在这种网络中，输入和输出的数量以及结构在整个训练和推理过程中都是固定的。网络的拓扑结构在训练期间保持不变，权重是唯一在学习中调整的参数。这种网络通常用于处理静态的、固定大小的数据。

动态神经网络是一种允许输入和输出的数量在模型运行时变化的神经网络结构。这意味着网络在处理序列数据或可变大小的输入时能够适应不同的长度和结构。与静态神经网络不同，动态神经网络的结构可以根据输入数据的特性动态地改变。这种灵活性使得它们在处理序列数据（如时间序列、自然语言文本）时非常有效。

两者的区别主要体现在输入输出结构、处理数据类型和应用场景上。在输入输出结构上静态神经网络的输入输出结构是固定的，而动态神经网络可以适应可变的输入输出结构；在处理数据类型上静态神经网络适用于处理静态的、固定大小的数据，而动态神经网络更适合处理动态的、序列型的数据；在应用场景上静态神经网络通常用于图像分类、固定大小的特征提取等任务，而动态神经网络更适合于序列建模，例如自然语言处理、语音识别等领域。