题 目： 基于卷积二次神经网络算法的图像分类研究

学院： 数学与统计学院 专业： 信息与计算科学 学生姓名： 魏资珏 学号： 20271024

|  |  |  |  |
| --- | --- | --- | --- |
| 文献综述：  近年来图像分类领域的发展势头迅猛，成果斐然，由于在此领域计算机识别能力已经远超人类，经典的ILSVRC (ImageNet) 竞赛也在2017年落幕。2018年起，由WebVision竞赛（Challenge on Visual Understanding by Learning from Web Data）来代替，而在ILSVRC竞赛中崭露头角的几个神经网络模型AlexNet, ZFNet, VGGNet, GoogleNet, ResNet 都采用的是一阶神经元，在2017年Fenglei Fan等人发表了论文“A New Type of Neurons for Machine Learning”[1]，该文探讨了在神经网络中将内积替换为输入向量的二次函数的可行性，在此后二次神经网络的发展打开了新的世界。而本研究致力于探究同结构的一阶二阶神经网络算法与性能的区别。  在此对一阶神经网络中的**DNN（深度神经网络）和CNN（卷积神经网络）**详细介绍其**发展历程**：  DNN（深度神经网络）：DNN的发展起源于上世纪五六十年代的单层感知机，这是一种只包含输入层和输出层，以及一个隐藏层的神经网络。然而，早期的单层感知机对稍微复杂一些的函数都无能为力，直到上世纪八十年代，Hinton、Rumelhart等人发明了多层感知机，也就是具有多层隐藏层的感知机，才克服了这个问题。多层感知机能够使用sigmoid或tanh等连续函数模拟神经元对激励的响应，同时在训练算法上使用了反向传播BP算法。这种多层感知机就是现在所说的DNN。  CNN（卷积神经网络）：CNN的起源可以追溯到上世纪80年代，当时主要用于识别手写数字和字符。然而，直到2006年，Hinton利用预训练方法缓解了梯度消失的问题，使得深度神经网络变得可训练，将隐含层发展到7层，神经网络真正意义上有了“深度”，由此揭开了深度学习的浪潮，第三代神经网络开始正式兴起。  **深度神经网络（DNN）发展趋势：**1、网络架构的发展：随着时间的推移，研究者们不断提出新的DNN架构，以提高图像分类性能。包括但不限于ResNet、Inception、EfficientNet等。2、迁移学习和预训练：迁移学习成为一个重要的研究方向，通过在大规模数据集上进行预训练，然后在目标任务上进行微调，取得了显著的性能提升。3、自监督学习：自监督学习变得越来越受关注，尤其是在没有大规模标注数据的情况下，通过网络自己生成标签进行训练。4、硬件加速：为了应对深度网络的计算需求，硬件加速方案如GPU和TPU得到广泛应用。  **近年来卷积神经网络（CNN）发展趋势**：1、卷积核设计：针对不同的图像分类任务，研究者们设计了各种不同的卷积核，以更好地捕捉图像中的特征。2、空间金字塔池化：为了处理不同尺寸的目标，空间金字塔池化等技术被引入，使网络具有多尺度的感受野。3、注意力机制：引入注意力机制以增强模型对重要图像区域的关注，提高分类性能。4、实时性和轻量级模型：针对嵌入式系统和移动设备，研究者们着重设计实时性和轻量级的CNN模型，以在资源受限的环境中运行。  **二阶神经网络与一阶神经网络相比，具有以下优势**：1、非线性拟合能力：一阶神经网络（如多层感知器或前馈神经网络）只能学习到线性分类边界，对于非线性问题，需要增加隐层神经元数量或使用核技巧等方法，增加了模型复杂度。而二阶神经网络可以利用二阶导数信息，更好地逼近复杂的非线性函数，因此可以更高效地解决非线性问题。2、特征表达能力：二阶神经网络可以利用二阶导数信息，提取更丰富的特征表达。相比之下，一阶神经网络只能提取一阶特征表达。3、鲁棒性：由于二阶神经网络可以更好地拟合复杂的非线性函数，因此对于噪声和异常值等干扰因素具有较强的鲁棒性。而一阶神经网络对于噪声和异常值较为敏感。4、泛化能力：二阶神经网络可以利用二阶导数信息，更好地理解数据的内在结构和规律，从而在训练数据较少的情况下也能取得较好的泛化性能。而一阶神经网络可能需要更多的训练数据才能获得较好的泛化性能。  神经网络中的**二次深度神经网络（Quadratic Deep Neural Networks，简称QDNN）**具有比一阶DNN更好的非线性、逼近和学习能力。现有DNN的每个神经元都是由输入X和权重参数W的线性组合（即一阶多项式形式）表示的，而QDNN的神经元则是由输入X和权重参数W的二阶多项式表示，与线性神经元相比，QDNN的优势来自于二阶多项式形式的独特性：(1) 更强的非线性，从而提高了特征外延能力；(2) 更高的模型效率，因为QDNN可以使用较小的网络深度/宽度来近似多项式决策边界。  **现有的QDNN设计**存在一些理论和实际应用上的**缺陷**。相关研究人员提出了一种新的QDNN神经元架构设计，并进一步开发了QuadraLib[3]（一个QDNN库），旨在为QDNN的架构优化和设计探索提供支持，这种新架构旨在提高QDNN的非线性表示能力和学习性能，同时降低计算复杂性。QuadraLib为QDNN的设计、训练和部署提供了全面的工具和功能，包括但不限于模型构建、训练、优化、评估等。  **本论文研究重心**在**卷积二次神经网络**(**CQNN**[2])与**二次深度神经网络**（**QDNN**）的**理论模型和实验结果优化**，基于对现有二阶模型的学习与实验分析，比较一阶与二阶神经网络、二阶神经网络中神经元的不同设计的理论优势与缺陷。具体来说:进行实验对比分析优化现有QuadraLib，并尝试在不同图像分类任务中验证模型可行性,对现存二次神经网络的学习与方法进行总结对比，比较给出具体的性能差距，在实际运用中去比较分析并拓展思路尝试新的神经网络架构。 | | | |
| 研究方案：  **卷积神经网络（CNN）和深度神经网络（DNN）的理论基础**：  1、神经网络理论：神经网络是模拟人类大脑神经元的一种计算模型，由多个神经元组成，通过调整权重和阈值来学习输入数据的特征。卷积神经网络和深度神经网络都是神经网络的一种，具有多层结构和连接性。  2、反向传播算法：反向传播算法是一种监督学习算法，通过不断调整神经网络的权重和阈值，使得输出结果与真实结果之间的误差最小化。卷积神经网络和深度神经网络都使用反向传播算法进行训练和优化。  3、激活函数：激活函数用于引入非线性因素，使得神经网络能够更好地学习和模拟复杂的非线性数据。卷积神经网络和深度神经网络中常用的激活函数包括sigmoid、tanh、ReLU等。  4、深度学习理论：深度学习是一种机器学习技术，通过构建深度神经网络来自动提取数据的特征并进行分类或回归等任务。深度神经网络是深度学习的一种实现，而卷积神经网络则是深度学习中一类具有代表性的算法。  **而在本研究中，理论基础**是对现有的各种QDNN模型的分析，将目前的QDNN模型分为四种类型：   |  |  |  | | --- | --- | --- | | 类型 | 神经元格式 | 说明 | | T1 |  | 每个大小为 d 的输入 X 将进行出与 d × d 全秩权重矩阵的乘积 | | T2 |  | 二阶项是通过直接平方每个输入 X | | T3 |  | 二阶项来自一阶神经元的平方 | | T4 |  | 二阶项由两个一阶神经元的不同权重参数的哈达玛乘积计算得出 |   以往的QDNN作品大多侧重于从理论上证明单个二次元神经元具有更强的非线性和逼近能力，因此它们只将QDNN应用于一些非常简单的学习任务，且网络结构较小。通过实际应用来研究这些现有QDNN的组合模式，发现有几个突出问题：1、近似能力不足；2、计算复杂性较高；3、收敛性能不佳；4、实施可行性问题；5、结构设计问题。所以有了新的QDNN神经元设计，同时为了优化以上问题，研究人员开发了QuadraLib，本研究将对生成的QDNN进行分析。  **需要解决的问题**：   1. 二阶神经网络的基本搭建 2. 优化在一阶神经元组合的过程中产生的哈达玛乘积计算复杂度问题以及组合的实现性。 3. 对于反向传播过程中不同神经元结构带来的梯度变化问题的优化与分析 4. 梯度消失问题，新神经元二次层是，其中反向传播过程中梯度是 5. 内存效率需要提高，在大多数DNN库中，反向传播过程中的梯度计算由自动微分（Auto-Differentiation，AD）[10]支持，会保存太多不必要的中间参数，而符号微分（Symbolic-Differentiation，SD）可以在反向传播之前推导出每个参数部分梯度的符号表达式。将SD与AD算法混合应用的可行性问题   **现有**QuadraLib**缺点**;1、模型设计复杂性：由于QDNN中的二次神经元具有更高的计算能力，设计QDNN模型时需要更多的考虑。例如，为了降低模型的计算成本并避免潜在的梯度消失或模型退化问题，需要调整模型的深度和宽度。由于二次项会产生极端值，批标准化层对QDNN非常重要，用于调节输出激活值。2、缺乏标准化：尽管有一些关于QDNN的研究工作，但该领域尚未形成标准化。不同的研究工作可能采用不同的参数和层类型，这使得比较不同QDNN模型的性能变得困难。3、缺乏广泛的应用场景：由于QDNN相对较新，它们在许多常见任务上的性能尚未得到充分验证。要证明QDNN的广泛适用性，还需要更多的实验和应用案例。4、可解释性挑战：与传统的线性神经网络相比，QDNN的内部工作机制更加复杂。这使得解释QDNN的决策过程变得更加困难。为了使QDNN在实际应用中得到更广泛的接受，需要开发更有效的可解释性方法。  **研究方法**：  1、解析法：通过分析程序的源代码、算法和数据结构等，来研究程序的性能、安全性和可维护性等方面。在本程序中去手推核心算法，详解所有的神经元设计函数（二元层结构），以及自动生成器的理论过程。从理论神经元结构优化的角度，提出了一种新的二次神经元。对于实用模型构建和训练选择 QuadraLib，其中的自动生成器可以生成最优的QDNN 模型结构。解析法可以帮助我们深入了解程序的实现细节，发现潜在的问题和改进点。  2、实验案例研究法：通过实际操作和试验来研究程序的性能、功能、安全等方面，并收集相关数据进行分析和评估。进行实验来评估 QuadraLib 中设计组件的有效性，包括新的二次神经元设计和基于混合BP的二次优化器。然后在几个有代表性的深度学习任务（图像分类、物体检测、基于GAN的图像生成）上进行整体实验，以证明 QuadraLib在各种任务中的通用性和可扩展性。实验法可以帮助我们了解程序的实际情况，发现潜在的问题和改进方向。  3、比较法： 在理论性能层面与实际应用层面，对原QuadraLib论文中提到的一些案例进行比较分析，将新提出的神经元格式与原有神经元格式和一阶形式进行分析比较。  **研究步骤**：   1. 搜寻相关资料及文献并分析理解 2. 初步理解理论源码，并进行二阶神经网络的复现与构建，输出各个网络参数进行整合分析 3. 分析比较所有二阶神经网络性能（理论层面）并尝试新神经元设计 4. 整合实验比较分析结果，完成可视化与结构化展示内容 5. 尝试解决现有QuadraLib缺陷，尝试结合其他二阶神经网络来优化 6. 论文写作   **预期成果**：  1.对现有的所有二阶神经网络进行实验分析，并综合比较理论性能与应用实践性能，进行试验比较，给出总结论文。  2.修正QuadraLib中的缺点并进行部分改进，结合现有网络来给出新型设计，并将修正后的结果编写保存为完整文件。 | | | |
| 主要参考文献：   1. Fan F, Cong W, Wang G. A new type of neurons for machine learning[J]. International Journal for Numerical Methods in Biomedical Engineering, 2017, 34(2). DOI:10.1002/ cnm.2920. 2. Mantini P, Shah S K. CQNN: Convolutional Quadratic Neural Networks[C]. Milan, Italy. 2020 25th International Conference on Pattern Recognition. IEEE, 2021: 9819-9826. DOI:10.1109/ICPR48806.2021.9413207. 3. Xu Z, Yu F, Xiong J, et al. QuadraLib: A Performant Quadratic Neural Network Library for Architecture Optimization and Design Exploration[C]. Proceedings of Machine Learning and Systems 4, 2022, 4: 503-514. DOI:10.48550/arXiv. 2204.01701. 4. Xu Z, Xiong J, Yu F, et al. Efficient Neural Network Implementation with Quadratic Neuron[J]. arXiv preprint, 2020, abs/2011.10813. DOI:10.48550/arXiv.2011.10813. 5. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint, 2014, abs/1409.1556. DOI:10.48550/arXiv.1409.1556. 6. Fan F, Wang G. Universal Approximation with Quadratic Deep Networks[J]. Neural Networks, 2020, 124: 383-392. DOI: 10.48550/arXiv.1808.00098. 7. Cheung K F, Leung C S. Rotational quadratic function neural networks[C]. Singapore. 1991 IEEE International Joint Conference on Neural Networks. IEEE, 1991: 869-874 vol. 1. DOI:10.1109 /IJCNN.1991.170509. 8. Chrysos G G, Moschoglou S, Bouritsas G, et al. Deep Polynomial Neural Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(8): 4021-4034. DOI:10.1109/CVPR42600.2020.00735. 9. Chen H, Wang J, Qi Q, et al. Bilinear CNN Models for Food Recognition[C]. Sydney, NSW, Australia. 2017 International Conference on Digital Image Computing: Techniques and Applications. IEEE, 2017: 1-6. DOI:10.1109/DICTA.2017.8227411. 10. Jiang Y, Yang F, Zhu H, et al. Nonlinear CNN: improving CNNs with quadratic convolutions[J]. Neural Computing and Applications, 2020, 32: 8507-8516. DOI:10.1007/s00521-019-04316-4. 11. Baydin A G, Pearlmutter B A, Radul A A, et al. Automatic Differentiation in Machine Learning: a Survey[J]. Journal of Marchine Learning Research, 2015, 18: 1-43. DOI:10.48550/arXiv.1502.05767. 12. Brooks D, Schwander O, Frédéric Barbaresco, et al. Second-Order Networks in PyTorch[C]. Geometric Science of Information: 4th International Conference, 2019: 751-758. DOI:10.1007/978-3-030-26980-7\_78. 13. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. The 32nd International Conference on Machine Learning, 2015: 448-456. DOI:10.48550/arXiv.1502. 03167. 14. Bae W, Yoo J, Ye J C. Beyond Deep Residual Learning for Image Restoration: Persistent Homology-Guided Manifold Simplification[C]. Honolulu, HI, USA. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2017: 145-153. DOI: 10.1109/CVPRW. 2017.152. 15. Bu J, Karpatne A. Quadratic residual networks: A new class of neural networks for solving forward and inverse problems in physics involving PDEs[C]. Proceedings of the 2021 SIAM International Conference on Data Mining, 2021: 675-683. DOI: 10.1137/ 1.9781611976700.76. 16. Salehinejad H, Sankar S, Barfett J, et al. Recent advances in recurrent neural networks[J]. arXiv preprint, 2017, abs/1801.01078. DOI:10.48550/arXiv.1801.01078. 17. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 2016: 770-778. DOI:10.1109/CVPR.2016.90. 18. N. DeClaris, M. Su. A novel class of neural networks with quadratic junctions[C]. Charlottesville, VA, USA. Conference Proceedings 1991 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 1991: 1557-1562 vol.3, DOI: 10.1109/ ICSMC.1991.169910. 19. Fan F, Shan H, Gjesteby L, et al. Quadratic neural networks for CT metal artifact reduction[C]. Developments in X-Ray Tomography XII. SPIE, 2019, 11113: 195-201. DOI:10.1117/12.2530363. 20. Milenkovic S, Obradovic Z, Litovski V. Annealing based dynamic learning in second-order neural networks[C]. Proceedings of International Conference on Neural Networks. IEEE, 1996, 1: 458-463. DOI:10.1109/ICNN.1996.548936. | | | |
| 毕业设计（论文）进度安排： | | | |
| 序号 | 毕业设计（论文）各阶段内容 | 时间安排 | 备注 |
| 1 | 搜寻相关资料文献 | 2023.12.1-12.20 |  |
| 2 | 初步尝试复现代码 | 2023.12.20-2024.1.4 |  |
| 3 | 理解全部算法并输出其中参数，进行整合分析 | 2024.1.5-2024.2 |  |
| 4 | 分析比较所有二阶神经网络性能（理论层面）并尝试新神经元设计 | 2024.2-2024.4 |  |
| 5 | 初步论文构思与起稿 | 2024.3 |  |
| 6 | 整合实验比较分析结果，完成可视化与结构化展示内容 | 2024.3-2024.4 |  |
| 7 | 完善论文内容，再次梳理检查实验数据与过程 | 2024.4-2024.5 |  |
| 8 | 论文相似性检测和修改，定稿 | 2024.5.1-2024.5.19 |  |
| 指导教师意见：  查阅资料全面，研究方案和计划进度可行，同意按学生提出的计划进行。 | | | |

指导教师（审核签名）：  审核日期： 2024 年 2 月 22 日