**学位论文评阅后修改情况说明**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **姓名** | **阮晓峰** | **学号** | **201718014628053** | **指导教师** | **胡卫明** |
| **论文题目** | **深度卷积神经网络压缩与加速方法研究** | | | | |
| **评阅修改意见及修改情况说明：**  （**列举评阅书中指出的修改意见，逐条说明做了哪些对应的修改，要求具体详实）** 评阅人1指出的修改意见及作者回复 **问题一：与该论文题目相近或相同的博士论文已经有好几篇了，建议修改论文题目，凸显出本文的方法特色，如：基于剪枝的.....方法。**  **回复： 在本文中，我们从模型压缩的方法、过程以及特征学习三个方面，提出了3种有效的方法，我们所提出的方法不仅仅局限于某一种模型压缩方法。**  **问题二：第3章关于模型压缩方法的对比实验中（如表3.5，3.6等），所提方法与对比方法在压缩模型的参数量上并没有优势，甚至计算量更大，需要对实验室结果进行解释。**  **回复：在第3章实验部分，表3.5和3.6表头中，模型的参数量和计算量使用参数下降率和*FLOPs*下降率来衡量，具体计算公式在实验设置评价指标中已给出，其中*FLOPs*下降率越大，模型计算量越小。在表3.5中，对于ResNet34，尽管ASFP 达到了最高的精度，但是压缩率略低，EDP 获得了最高的压缩率，精度也可以接受。对于ResNet50，EDP 获得了最高的精度，同时压缩率仅次于FPGM 和NISP，远远超过了ThiNet、SFP 和HRank。对于ResNet101，更深的网络，EDP 达到了最高的压缩比，和Baseline 比较，Top-1精度提高0.46%、Top-5 精度提高0.14%。在表3.6中，使用EDP压缩后的模型计算量比0.75x MobileNet v2和AMC略低，同时精度也高于对比的方法。** | | | | | |
| 评阅人2指出的修改意见及作者回复 **问题一：论文3~5章分别提出了三种模型压缩与加速方法，建议对所提出方法进行比较分析。**  **回复：我们在第6章《总结与展望》中对三种方法进行了总结，并在论文修改版本中修改。**  **问题二：公式（5.4）与公式（5.9）使用KL散度，而对压缩模型输出预测使用了对数函数，存在歧义，建议去掉对数函数，或将KL散度在基准模型不变的条件下，进一步明确为交叉熵损失。**  **回复：按照评阅人的意见，在论文修改版本中，已经去掉输出预测的对数函数。**  **问题三：论文第5章所提方法包含多个（论文设置为2个）子压缩模型，算法3及正文部分，未涉及训练完成后多个子模型的处理方法，建议给出对子模型的具体选择或合并操作说明。**  **回复：按照评阅人的意见，在论文修改版本中，我们在实验设置部分添加了具体的说明。具体表述为“实验结果为所有压缩子模型对应结果的平均值”。**  **问题四：建议对图表中的符号含义进行简单说明，英文缩写第一次出现建议给出全称。**  **回复：按照评阅人的意见，在论文修改版本中，我们对一些图表中的符号含义进行了说明，同时对第一次出现的英文缩写给出了全称。**  **问题五：存在少量笔误。**  **回复：按照评阅人的意见，我们梳理了全文，并在论文修改版本中更正。** 评阅人3指出的修改意见及作者回复 **问题一：如果参考原文献的图表，建议作者在图表的标题中给出被引文献。**  **回复：按照评阅人的意见，我们梳理了全文，并在图表的标题中给出了被引文献。** 评阅人4指出的修改意见及作者回复 **问题一：根据图3.7的实验结果，EDP低秩分解和通道剪枝部分的性能表现各自优于对比方法SVD和Slimming，其理论基础是什么？请补充阐述和分析。**  **回复：我们的方法各自优于对比的方法，原因有两点：（1）稀疏正则方式不同。在我们的方法中，采用一种解耦的方式，显示地施加正则，即通过对系数矩阵的行和列施加稀疏正则，可以实现通道稀疏和低秩分解；而对比的SVD方法通过对权重矩阵的核范数施加正则；我们的通道稀疏，由于卷积层分解为两层，有更多的空间表示学习稀疏，避免了slimming直接稀疏通道带来的信息丢失；（2）训练方式不同。我们在训练过程中，采用早停的方式，仅仅在训练期间的一部分施加正则约束，而对比的SVD和Slimming在整个过程施加约束，这样会影响压缩后模型的性能。**  **问题二：DPFPS不需要预训练模型，可以直接学习一个结构化稀疏的网络，实验结果表明了其有效性。但是，为什么剪枝后的模型不需要Fine-tuning依然能保持良好的性能，理论分析不够充分。**  **回复：在以前的结构化稀疏方法中，通过施加稀疏正则的方式训练网络后，模型的稀疏率与预设剪枝率很难匹配。在较大的预设剪枝率下，模型的稀疏率往往达不到预设剪枝率。在剪枝过程中，一些非零参数被移除，导致压缩后的模型精度下降明显。因此，为了恢复精度，需要微调剪枝后的模型，这样会带来一些超参数，比如学习率、微调轮次、多次迭代次数等。在我们所提出的动态渐进式稀疏算法中，仅仅在预期稀疏的参数上施加结构化稀疏正则，模型训练结束后，网络的稀疏率与预设剪枝率相同，因此直接移除权值为零的参数，网络的性能不会下降，即不需要通过Fine-tuning来恢复精度。**  **问题三：第五章小结中认为“基于模型特征学习增强训练的动态剪枝方法，在训练过程中，该方法不需要数据集类别标签”。但是该方法利用了基准模型输出的类别信息指导压缩模型学习，由于基准模型具有较高的精度，其输出的类别信息相当于数据集标签；另一方面，如果基准模型本身是用带标签的数据集训练所获得的，那么压缩模型为什么不采用数据集标签呢，基于何种应用场景的实际需要？**  **回复：（1）我们充分利用预训练模型的表征信息来指导压缩模型的特征学习，缓解了对数据标签的强依赖性。在一些应用场景中，比如医学图像分析任务中，往往含有大量的无标签数据。（2）我们拓展了算法的应用场景范围。**  **问题四：本论文第三、四、五章分别从不同的角度研究了3种深度卷积神经网络压缩与加速方法，文中实验结果表明3种方法都在基本保持模型精度的前提下，获得了良好的模型压缩率。那么，这3种方法是否可以同时作用，以获得模型压缩率的进一步提升？**  **回复：在本论文提出的三种算法中，我们都采用了松弛的稀疏优化方式，通过学习稀疏的手段，自适应地获得轻量化的网络结构。在第三章，我们通过在原始网络中构造一个可压缩的模块，采用通用的稀疏正则方式，可以将分解与剪枝方法有效统一；在第四章，我们改进了通用的稀疏正则方式，提出了一种动态渐进式稀疏正则方式，可以从头开始学习一个结构化稀疏的网络。在第五章中，我们充分利用模型本身的特征表征信息，将提出的动态稀疏正则方式进行了扩展，应用到了无类别标签数据场景下的模型压缩与加速任务中。通过实验结果可以看出，在CIFAR-10下，使用DPFPS算法，我们可以减少VGG-Small模型93.32%的参数和70.85%的计算量，精度为93.52%；在第五章中，我们可以减少VGG-Small模型95.00%的参数和77.71%的计算量，精度为93.80%。这些结果说明，第四章提出的动态稀疏算法在第五章中是起作用的，并且通过利用第五章中预训练模型的表征信息，进一步提升了模型压缩性能。第三章我们重点突出的是低秩和剪枝两种方法有效统一，而第四章和第五章，重点是研究单纯的剪枝方法。正如前面所述，我们都采用了松弛的稀疏优化方式，我们的动态稀疏算法也可以用在第三章中。** 评阅人5指出的修改意见及作者回复 **问题一：公式（3.3）和（4.1）下面的“网络前向函数”的提法有混淆，建议改用“转移函数”或“激励函数”。**  **回复：按照评阅人的意见，在论文修改版本中，我们将“网络前向函数”改为“通过网络所有层（包括softmax层）变换后，对应输入数据的输出值”。**  **问题二： 建议把公式（3.4）和（3.8）第二个式子下面的“𝑖.𝑒.”换为“其中”。**  **回复： 按照评阅人的意见，在论文修改版本中，我们将“𝑖.𝑒.”改为“其中”。**  **问题三： 公式（3.13）的右边多了“× (1⋅1)”。**  **回复： 此处（1⋅1）表示卷积核大小，不是具体乘法计算，不进行修改。**  **问题四： 图 5.2的图题“提出方法框架图”不妥，建议修改为“方法框架图”。**  **回复： 按照评阅人的意见，在论文修改版本中，我们将“提出方法框架图”改为“我们的方法框架图”。**  **问题五： 论文中完全一致的段落较多。例如，在“摘要”和“总结与展望”部分。**  **回复： 按照评阅人的意见，在论文修改版本中，我们修改了《总结与展望》的第一段。具体表述为：**  **“近年来，深度学习算法在人工智能领域有了飞速发展。然而，主流的深度学习模型往往需要大量的计算开销和内存开销，难以直接部署到智能移动设备中。相关研究[77] 已经表明神经网络包含大量的冗余参数。在这种背景下，模型压缩与加速方法出现，在较小的精度损失下，能够有效地减少模型的参数以及提升模型的计算效率，从而使得深度学习模型在移动端的部署成为可能。本文围绕目前深度神经网络压缩与加速存在的若干问题，主要从压缩方法、压缩过程、特征学习3 个方面进行了系统地、深入地分析，同时提出了三种有效的模型压缩与加速方法。在这三种方法中，我们都采用了松弛的稀疏优化方式，通过学习稀疏的手段，自适应地获得轻量化的网络结构。首先，我们通过在原始网络中构造一个可压缩的模块，采用通用的稀疏正则方式，可以将分解与剪枝方法有效统一；其次，我们改进了通用的稀疏正则方式，提出了一种动态渐进式稀疏正则方式，可以从头开始学习一个结构化稀疏的网络。最后，我们充分利用预训练模型本身的特征表征信息，将提出的动态稀疏正则方式进行了扩展，应用到了无类别标签数据场景下的模型压缩与加速任务中。”**  **作者签字：**  **日期：** | | | | | |
| **导师意见**  **导师签字：**  **日期：** | | | | | |

注：双面打印，本人、导师签字。不要超过8page