**ACFNet: An Adaptive Channel Feature Information Fusion Method**

**Abstract**

在本文中，我们将Swish激活函数以及模型重参数定义法等性能提升技巧引入到了轻量级神经网络中，针对轻量级神经网络ShuffleNet V2的通道混洗结构进行了改进，设计出更合理的通道混合卷积结构，并在其中加入了通道注意力机制模块。在ImageNet标准数据集上的实验表明，该创新网络结构的检测性能明显优于ShuffleNet V2，在相同的实验条件下，比ShuffleNet V2在ImageNet分类任务上的top-1错误率要低1.7%，且在GPU设备上的模型推理速度与ShuffleNet V2相当，而在0.5x和1.0x低倍率的模型下，模型推理速度要更优于ShuffleNet V2。

**Introduction**

在实际应用的场景中，嵌入式设备、移动端设备等，往往因其本身资源有限，导致一些性能强劲的大型神经网络对于这样的任务与需求也显得无能无力。不过在过去几年中，轻量级卷积神经网络的研究越来越受到人们的重视，其中最具有代表型的轻量级网络就有MobeilNet系列和ShuffleNet系列。前者使用深度可分离卷积来构建轻型深度神经网络，极大地降低了网络模型的参数量和运算量，后者则利用逐点组卷积和通道混洗的结构，在保证模型的检测精度同时，大大地降低模型的计算量。在ShuffleNetV2中，作者认为多分支特征信息提取结构会较大地增加模型的计算成本，不符合高效的轻量级网络设计原则。但在之前Google提出的Inception网络结构中证实了多分支结构对于提取特征信息的有效性，因此在2021年的RepVGG论文中，作者提出了重参数定义的方法，并通过该方法实现多分支结构等价转换为单分支结构，且转换前后模型检测精度不受影响。所以我们也尝试将重参数定义方法引入到轻量级神经网络中，在保持模型的计算量不变的情况下提高模型的检测精度。另外我们还注意到，ShuffleNetV2中的通道混洗结构采用的是人工固定设计模式，即分组卷积中不同组的特征信息的交互方式是按照人为规定的固定方式进行。但我们认为人为固定的信息交互规则不一定是最好的，通过模型自适应地训练或许可以得到更好的信息交互方式。并且我们认为不同通道上的特征信息其携带的信息权重是不一样，所以我们在不同组信息交互的过程中还引入了通道注意力机制，我们没有直接利用SENet中的通道注意力机制模块，因为它会大大地增加模型的计算量和参数量，而是对SENet中的通道注意力机制模块进行了结构裁剪，使得它能更好地利用到轻量级神经网络中，同时也更便于使用重参数定义法进行结构转换。我们对ShuffleNetV2结构中的激活函数也进行了替换，将原先的ReLU激活函数全部替换为Swish激活函数，通过实验数据表明，Swish激活函数在ImageNet大型数据集上，相比ReLU确实更具有优势。

**Related Work**

在过去几年中，轻量级卷积神经网络的研究越来越受到人们的重视，其中最具有代表性的轻量级网络有MobeilNet系列[1]和ShuffleNet系列[2]。前者使用深度可分离卷积来构建轻型深度神经网络，极大地降低了网络模型的参数量和运算量，后者则利用逐点组卷积和通道混洗的结构，在保证模型的检测精度同时，大大地降低模型的计算量。

**通道混洗** 在ShuffleNet系列中，通道混洗的概念首次被提出，它的主要目的是改善分组卷积中不同组之间的信息闭塞问题，通过对不同组的特征信息在通道顺序上进行重新交错排列，实现不同组之间的信息交互。但美中不足的是，这样的通道混洗操作是按照人为固定规则来实施的而非网络自适应地进行信息交互 。

**注意力机制** 在最近几年，注意力机制的研究热度持续高涨，诸如SENet等[3]一大批高性能的注意力机制网络结构相继被提出，但遗憾地是它们都并没有被广泛应用到轻量级神经网络上，其主要原因是由于注意力机制给模型带来了高昂的计算成本，而这样的计算成本对于轻量级神经网络来说是致命的。

**多分支卷积结构** 在GoogLeNet系列[4]论文中，多分支卷积结构被广泛采用，并证实其对于提高模型检测精度的有效性，但ShuffleNetV2提出的四项高效结构设计原则中指出，多分支卷积结构会造成网络结构碎片化，可能会降低网络运行效率，引入了额外的开销，因此在目前为止的轻量级神经网络中，多分支卷积结构仍未得到广泛采用。

而在我们的网络结构中，我们很好地解决以上存在的问题，其主要的解决思路来着于RepVGG系列[5] 论文中提出的重参数定义方法，通过重参数定义法我们可以进行巧妙地结构转换，在保持模型推理速度不变的同时，显著地提高模型检测精度。

1.通道注意力机制的重参数定义

卷积神经网络通过增加模型深度，即卷积通道数来获取更多的特征信息，但在诸多特征信息通道中，并不是每一个通道都具有丰富的特征信息，也许有多个通道特征信息都是类似的，也许很多通道特征信息也是冗余的，但人们无法事先指定到底哪个通道上的特征信息是重要的，因此通道注意力机制结构的研究可以使得模型自适应地对重要的特征信息给予更多的权重，其中典型的就是SENet，在SENet中同一批次的不同数据都拥有各自不同的通道权重参数，但这样带来了极大的模型计算量，显然不适应在轻量级网络中使用。因此我们尝试模型同一批次的所有数据都共用同一个通道权重参数，这样做有一个明显的好处就是可以将这个通道权重结构通过重参数定义为一个1x1的分组卷积结构，极大地降低了模型的计算量，同时又弥补了轻量级网络中通道注意力机制的缺失

2.通道混洗的重参数定义

ShuffleNet中提出了通道混洗的结构来解决传统分组卷积带来的不同组之间信息闭塞的问题。在该网络中，一组特征信息被分为n份等通道长度的特征信息，然后分别将n组特征信息在通道顺序上，按彼此穿插的顺序重新进行排序组成新的一组特征信息，来达到不同组之间信息交互的目的。同时V2模型中提出了分组卷积的高效设计原则，最终他们将n的取值为2，但是这样会将同一组的特征信息进行强制性打乱，使得同一组的特征信息过于分散化，同时这样不同组之间的通道顺序排列是人为固定的，缺乏模型通过不断训练来自适应选择的能力。因此我们引入了一个1x1的普通卷积，将它与指定的特征信息进行卷积操作，可以实现将原先的特征通道按照反向通道顺序排列的功能，再将原先的特征信息与卷积操作过后的特征信息进行相加操作，即可实现在不打乱一组特征信息顺序的同时，使得一组特征信息的前后两部分实现信息交互的目的，同时在1x1卷积过程中，还将引入可训练的通道注意力权重参数，让模型可以自适应的训练两组特征间信息交互的权重参数，使得不同组之间的信息交互方式变得更加合理有效。

3.卷积操作之间的统一化重参数定义

前面我们引入了1x1的分组卷积来实现轻量级网络中的通道注意力机制，同时我们又引入了1x1的普通卷积来实现对一组特征通道的反向排序，而他们的操作对象均是相同的特征信息对象，因此可以通过重参数定义的方式，在推理过程中，将1x1分组卷积融合到1x1普通卷积里，最终通道注意力机制和通道混洗操作就可以通过一个重参数定义后的1x1卷积操作来实现，可以极大降低模型的运算量，同时提高模型的检测精度。

最后我们还将原先ShuffleNet中的激活函数，由ReLU更换为目前各大神经网络中更为常用的SiLU激活函数，实验数据证明这样的更换是合理且有效的

**Method**

在SENet结构中采用的通道注意力模块对模型训练过程中的同一批次的不同数据都设置了对应的权重参数A1，这样的设计高效但是会大大降低模型运算速度，因此我们进行结构裁剪，使得同一批次的数据都共用同一组权重参数，而不同批次数据则使用不同的通道权重A1，这样的设计会降低通道注意力模块的有效性，但同时也能降低模型的运算量，裁剪后的结构可以等效理解为对原特征图V执行了1x1分组卷积，例如V=[v1,v2...vc]，其中vc表示第C个特征通道，经过1x1分组卷积后，U=[a1\*v1,a2\*v2...ac\*vc]，其中的权重参数a初始值均设置为1.0，并让其参与到模型的训练过程中，直至训练为一组合理的权重参数A1~，就行CANet中操作的那样。

我们针对分组卷积中不同组的信息交互问题设计了一种通道混洗卷积的结构，它的本质是一种人为设计的1x1的逐点卷积核，在通道数为C的1x1卷积核中几乎所有的卷积参数都为0，但是我们会将最后一个特征通道上的卷积参数设置为1.0，将这个卷积核与原特征图V进行卷积，就会得到原特征图的最后一个特征通道数据，这样的卷积核共有C个，最后就会得到一个与原特征图在特征通道上顺序刚好相反的新特征图U，将原特征图与新特征图执行元素相加操作，就达到了卷积信息交互的目的。与此同时我们还在这样的逐点卷积过程中引入了通道权重参数B1，并将权重值依次赋值到对应的卷积核通道上。同理，B1中所有权重初始值我们设置为0.0，然后让其参与到模型的训练过程中，直至训练为一组合理的权重参数B1~，这样的信息交互过程是通过模型自适应训练进行的，而非固定地执行元素相加操作，例如V=[v1,v2...vc]，其中vc表示第C个特征通道，经过1x1逐点卷积后，U=[a1\*v1+b1\*vc,a2\*v2+b2\*vc-1...ac\*vc+bc\*v1]。

以上介绍的1x1分组卷积和1x1逐点卷积对于模型有明显地性能提升作用，但遗憾地是，它们的设计过程以及运算过程是较为复杂的，对于轻量级神经网络的影响不容忽视。因此我们引入了重参数定义法，在模型训练阶段应用多分支结构，引入通道权重参数进行训练，初始化1x1逐点卷积核，并进行通道混洗卷积操作。当模型训练结束后，通道权重参数以及卷积核参数都已固定，因此在模型推理阶段通过重参数定义进行模型结构等效替换，极大程度地提高模型的推理速度，简化模型推理结构，保证模型推理精度。

**Experiments**

1. 数据集的选择即实验条件的设定

我们采用的硬件资源与ShuffleNetV2中的硬件资源有所差异，我们采用了单个RTX3060 英伟达显卡，卷积库是CUDNN 8.0。因为硬件资源有限，所以我们没有直接引用ShuffleNet中的模型训练设置和超参数设置，其中主要的原因在于显存不足对训练批次数的限制。因此在0.5x和1.0x倍率的模型训练时，batchSize由1024均更改为256，1.5x和2.0x倍率的模型训练时，batchSize由1024分别更改为200和150，除此之外的超参数和模型训练设置我们都与ShuffleNetV2原论文保持一致。我们选择ImageNet 2012分类数据集来进行模型评估，我们在ImageNet验证集上比较了top-1和top-5分类错误率。 从256×输入图像中裁剪224×224中心视图并评估分类准确性。为了公平地进行比较，我们采用了ShuffleNetV2的开源代码，对其不同倍率下的模型训练参数进行调整，然后在我们现有的硬件资源上进行训练，确保两者模型使用完全相同的训练设置与训练环境，让比较数据更具有可靠性。

1. 与ShuffleNetV2的实验条件数据对比（包括精度和速度两部分）

图1总结了我们的改进模型与ShuffleNetV2模型下在精度和速度上的直接度量结果，其中还包括训练阶段的相关数据以及经过重参数定义后的推理阶段相关数据：

从以上数据中可以看出，不管在哪一种倍率的模型结构下，我们改进的模型在精度上对于ShuffleNet都有较为明显的提升。其中0.5倍率模型下提升了1.6%和1.1%；在1.0倍率模型下提升了1.6%和1.0%；在1.5x倍率模型下提升了1.7%和1.0%；在2.0x倍率模型下提升了1.7%和1.1%，虽然在训练阶段，我们的模型参数量和计算量均大于ShuffleNet，但在经过重参数定义后的推理阶段，我们的模型参数量和计算量得到很大程度地提升，基本都与ShuffleNet保持相当的水平。同时我们发现，在重参数定义后的推理阶段，在0.5x和1.0x低倍率模型的比较中，我们模型在GPU上的推理速度是明显高出ShuffleNet的，这说明利用重参数定义法对我们设计的模型结构进行优化是富有成效的，但是在1.5x和2.0x高倍率模型下，即使使用了重参数定义法进行了优化，我们模型在GPU上的推理速度依旧低于ShuffleNet，我们分析认为，重参数定义法的优化提升已经无法抵消高倍率模型下，因我们改进模型结构中而增加模型参数带来的负面影响。这是一个值得关注的问题，也将会是我们接下来需要去进一步研究并完善的地方。但总体上看，我们对于ShuffleNet进行的模型改进是合理且有效地，在保持模型推理速度不变的情况下，明显地提升了模型的检测精度。（推理比较的设置参数还需要添加一下）

1. 消融实验数据介绍（主要对精度的影响）

接下来我们将讨论在改进模型中一些结构的有效性，我们采用了1.0x倍率的模型进行了消融实验，图2总结了我们进行的消融实验相关数据：

从实验数据中可以看出，将网络中的激活函数由ReLU替换为SiLU后，模型的检测精度有了较为明显的提升，同时模型的推理速度并未出现明显的降低。之后，我们将网络中的3x3单分支卷积结构替换为3x3+1x1的双分支结构，结果表明，多分支确实可以提取到更多的通道特征信息，有利于模型检测精度的提升，但同时由于新引入了模型参数，模型的运算速度会受到一定程度地影响，但模型经过重参数定义后，将多分支结构等效替换为单分支结构，避免了模型推理速度的降低。接下来我们去掉多分支结构，引入通道注意力机制，模型的检测精度由32.2提升到31.9。同理，经过重参数定义后，避免了模型推理速度的降低，实验数据证明，我们所进行的结构改进都能有效地提升模型的检测精度，同时又能通过巧妙地重参数定义来改善这样的模型改造带来的不利影响。

**Conclusion**

我们通过对以往一些轻量级网络的学习与理解，发现了其中的一些问题点所在，并在ShuffleNet的基础上提出了一些新的模型结构，通过全面的实验证明了我们这样的结构设计是合理且有效的，我们巧妙地利用重参数定义法实现了分组卷积信息交互卷积结构与通道注意力机制结构的结合，极大避免了新设计结构带来的弊端。同时我们也发现了一些新的问题，即如何通过重参数定义法能够完全等效地转换SENet中提出的高性能通道注意力机制结构，是未来需要继续重点去研究的课题。

**References**

**参考文献**