中国科学院软件研究所学术年会’2022
暨计算机科学国家重点实验室开放周
学术论文

基于弹性连接的二值神经网络

AAAI 2022

Elastic-Link for Binarized Neural Network

胡杰、吴梓恒、谭俊恺、吴恩华等

联系方式：胡杰，13121684025，hujie@ios.ac.cn

背景

通过将神经网络的特征和参数传入sign函数进行二值化（+1，-1），可以使得标准的**浮点数的数值计算**被**位与或**操作替代。在资源受限的移动设备或者嵌入式设备上，对神经网络进行二值化后，能够大幅**减少神经网络在推理时所需的计算和存储开销**。很多学者已经对此做了深入的探索，但是他们都刻意避开了对含有1x1卷积网络的探索。已有二值化方法在遇到1x1卷积时性能均有较大程度地下降，但是1x1卷积又是一个在神经网络中极其常见且重要的模块。因此解决1x1卷积的二值化性能问题可以使得二值化技术运用在所有卷积神经网络中，提高了二值化技术的普适性。

弹性连接

我们提出一种通用的模块命名为“弹性连接”（Elastic-Link，简称EL）来解决带有1x1卷积的神经网络在二值化后精度大幅下降的问题。我们把二值化卷积前的浮点数特征通过残差连接与二值卷积后的特征进行加和，从而减少网络二值化过程中的信息损失。然而1x1卷积的使用往往伴随着特征通道数目的改变，即输入通道数与输出通道数不一致。EL模块对于通道数增加的情况，直接在通道维度上复制通道特征来补齐通道数；对于通道数减少的情况，我们把输入通道进行不交叠分组（每组的通道数等于输出通道数）求和。同时为了让特征分布不产生较大偏移，我们还增加了一个可学习参数 $\boldsymbol{\gamma}$ 来进行加权。EL模块结构示意图见图1。

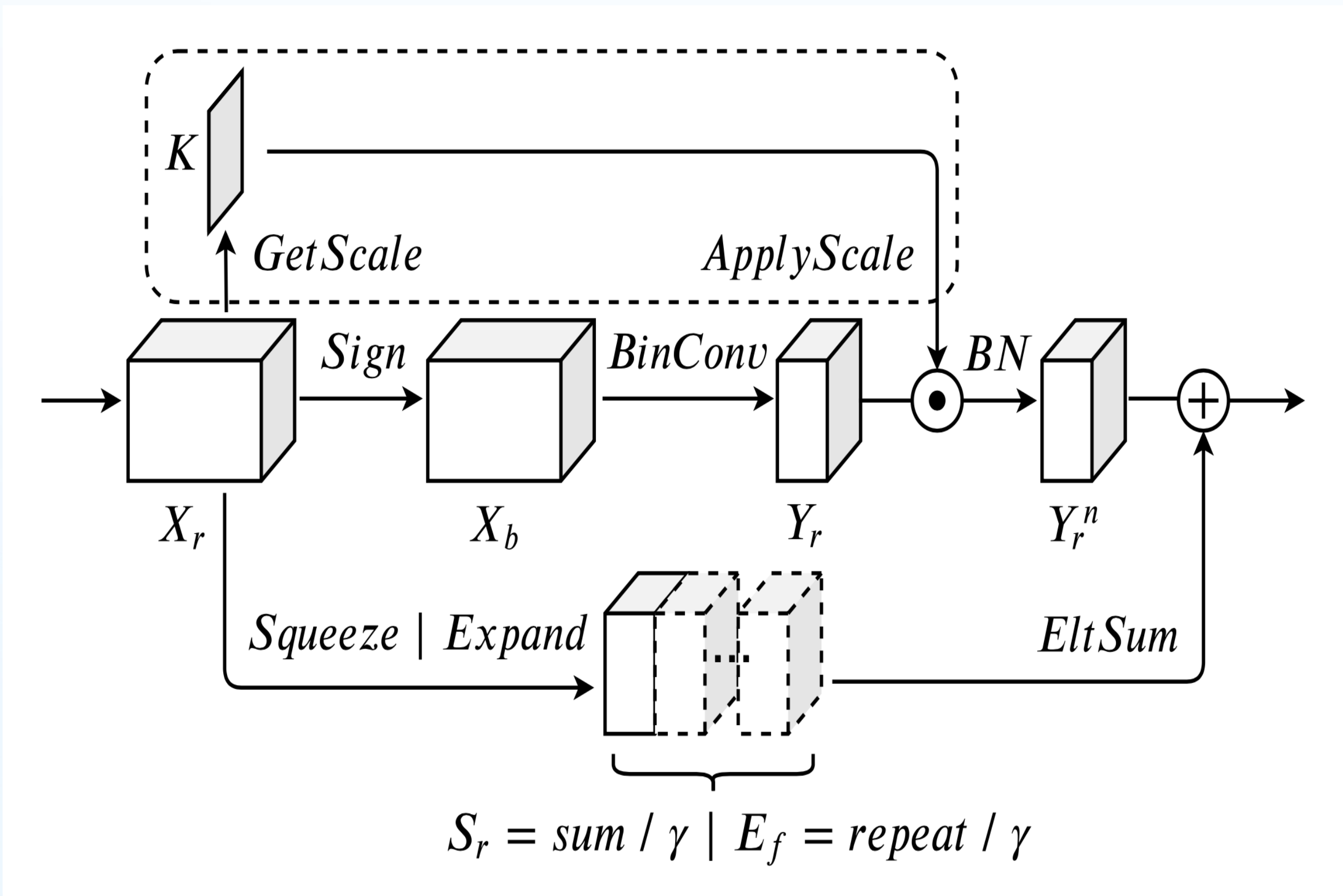

图1 弹性连接结构示意图

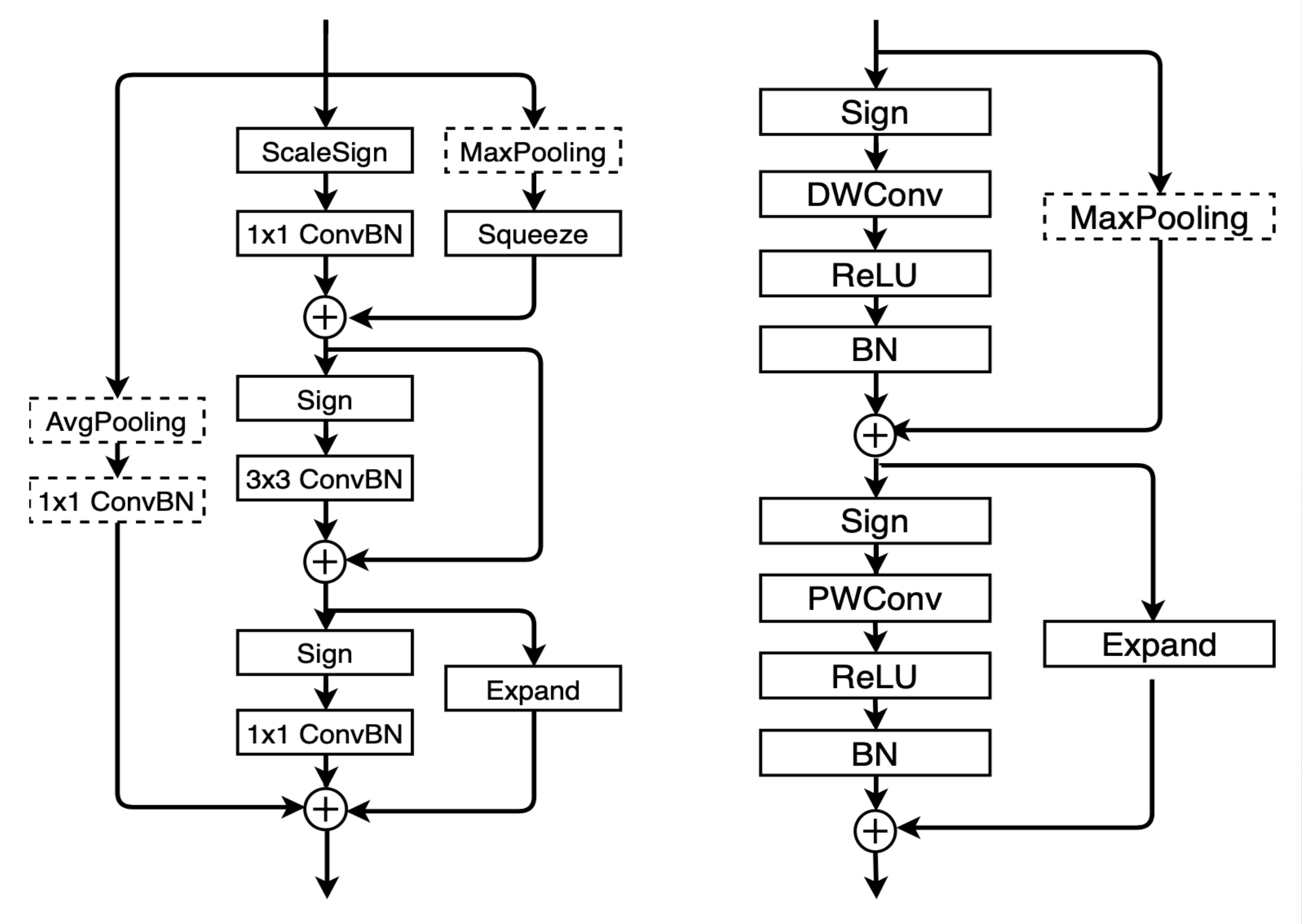

图2 EL在ResNet和MobileNet二值化模型中的应用

实验和结论

在ResNet和MobileNet两个比较有代表性的模型上验证了EL的效果（结构见图2）。从表1可以看出，EL方法在ImageNet1k图像分类任务上均取得了最好的结果。

	Bottleneck Block		Efficient Block	Basic Block	
	RN26	RN50	MobileNet	RN18	RN34
Full-Precision	72.5	75.9	70.6	69.3	71.5
XNOR (Rastegari et al. 2016)	52.1	54.2	*Not Converge*	51.2	53.2
ABCNet (Lin, Zhao, and Pan 2017)	45.2	52.9	*Not Converge*	42.7	52.4
TBN (Wan et al. 2018)	-	-	-	55.6	58.2
Bi-Real (Liu et al. 2018)	57.8	62.7	*Not Converge*	56.4	62.2
BinaryE (Bethge et al. 2019)	57.9	61.2	*Not Converge*	56.7	59.5
CI-Net (Wang et al. 2019)	-	-	-	56.7	62.4
XNOR++ (Bulat and Tzimiropoulos 2019)	-	-	-	57.1	-
GBCN (Liu et al. 2019)	-	-	-	57.8	-
MoBiNet (Phan et al. 2020)	-	-	54.4	-	-
EL (Ours)	64.0	65.6	56.4	60.1	63.2
Real-to-Bin (Brais, Bulat, and Tzimiropoulos 2020)	64.8	65.9	54.8	65.4	66.1
EL† (Ours)	**67.1**	**68.9**	**61.2**	**65.7**	**66.5**

表1 EL对比当前STOA的模型结果，取得了最优性能

EL是一个通用的模块，它可以嵌入在所有的卷积网络中。同时它也可以兼容其它二值网络的各类训练技巧，使得二值网络的精度往前迈进了一个台阶。