学界 | RESBINNET：剩余二元神经网络

2017-11-07 机器海岸线

选自 arXiv

作者：Mohammad Ghasemzadeh, Mohammad Samragh & Farinaz Koushanfar 等

机器海岸线编译

参与：方建勇

RESBINNET: RESIDUAL BINARY NEURAL NETWORK

Mohammad Ghasemzadeh, Mohammad Samragh & Farinaz Koushanfar
University of California San Diego
California , USA
{mghasemzadeh,msamragh,farinaz}@ucsd.edu

论文链接：https://arxiv.org/pdf/1711.01243

摘要：最近在训练轻量级二元神经网络方面的努力提供了有前景的执行/记忆效率。 本文介绍了 ResBinNet，它是两种相互关联方法的组合，旨在解决二进制卷积神经网络收敛速度慢、精度有限的问题。 第一种方法称为残差二值化，学习一个神经网络层中的特征的多级二进制表示。 第二种称为温度调整的方法，逐渐将特定层的权重二值化。 这两种方法共同学习了一组软二值化参数，提高了二值神经网络的收敛速度和准确性。 我们通过实施原型硬件加速器来证实 ResBinNet 的可用性和可扩展性。 加速器可根据二值化特征的数值精度进行重新配置，从而在运行时间和推理精度之间进行权衡。

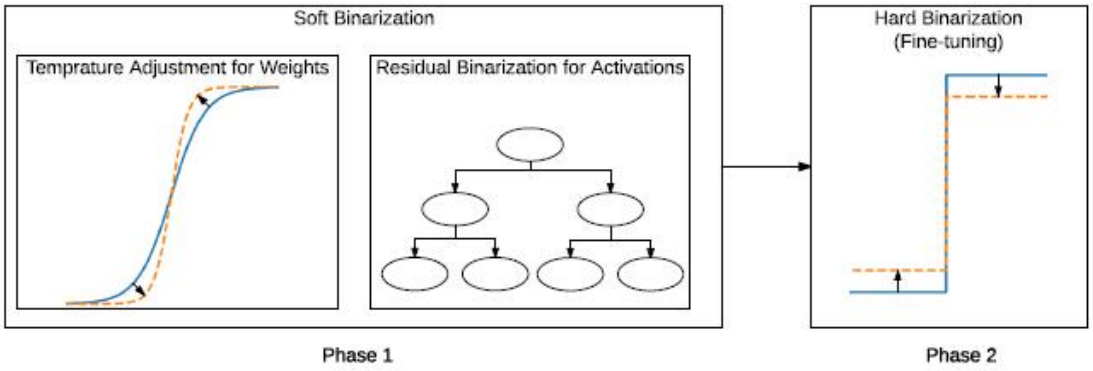

图 1：ResBinNet 二进制训练的全局流程。 残差二值化学习特征映射的多级表示。 温度调整在可训练的权重上执行变量变化，并在训练阶段逐渐将其推向二进制值。

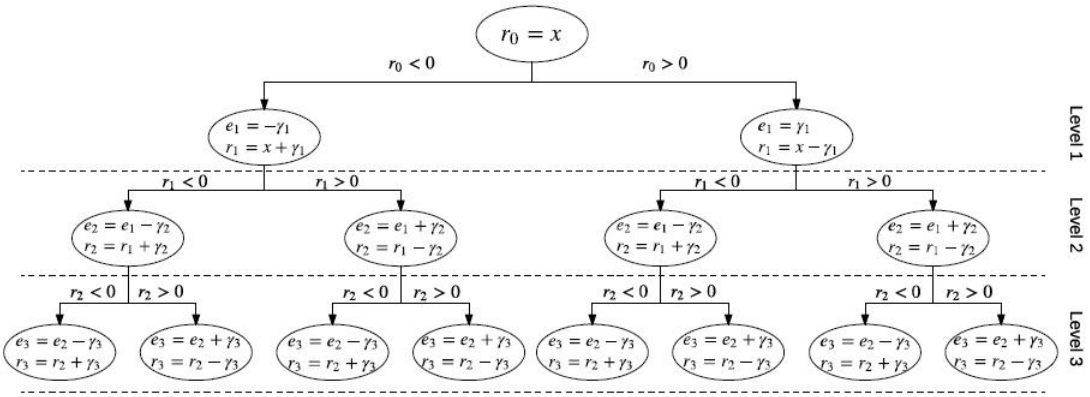

图 2：用于计算 3 级残差二进制估计的示意流 e。 随着层次越来越深，估算变得更加准确。

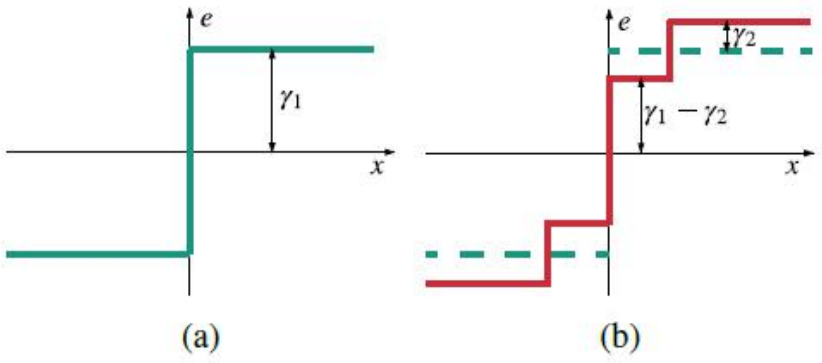

图 3：二进制激活函数的例证。 （a）传统的 1 级二值化。 （b）两级残留二值化。

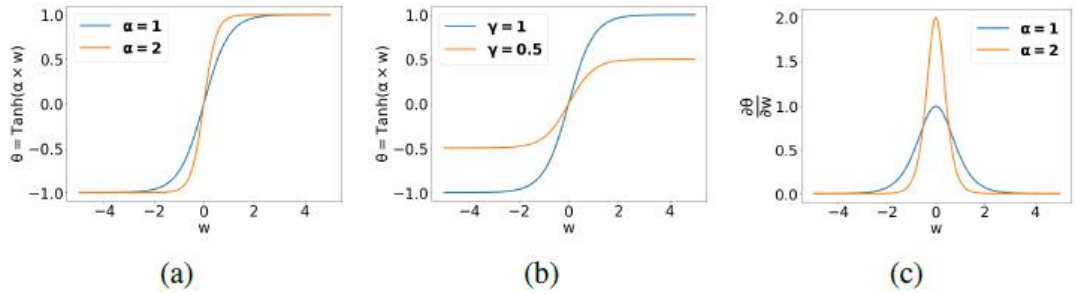

图 4：Tanh 非线性变量的一个例子。 （a）温度的影响

参数：更高的值提供更好的软二进制估计。 （b）边界的影响

参数：是每个权重矩阵 W 的可训练值。 （c）温度的影响

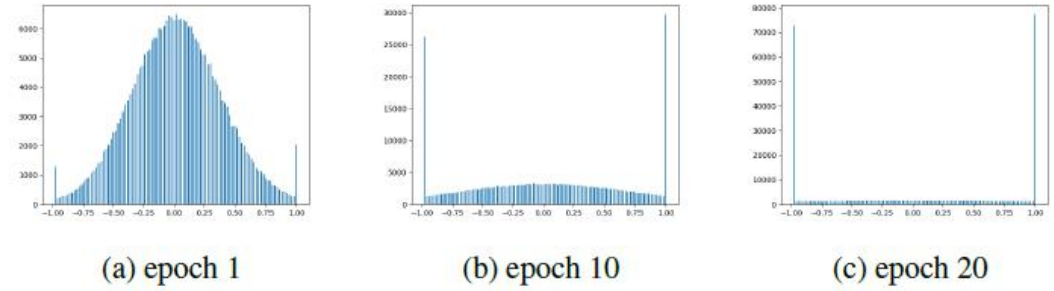

图 5：训练期间神经网络某一层中元素的直方图。

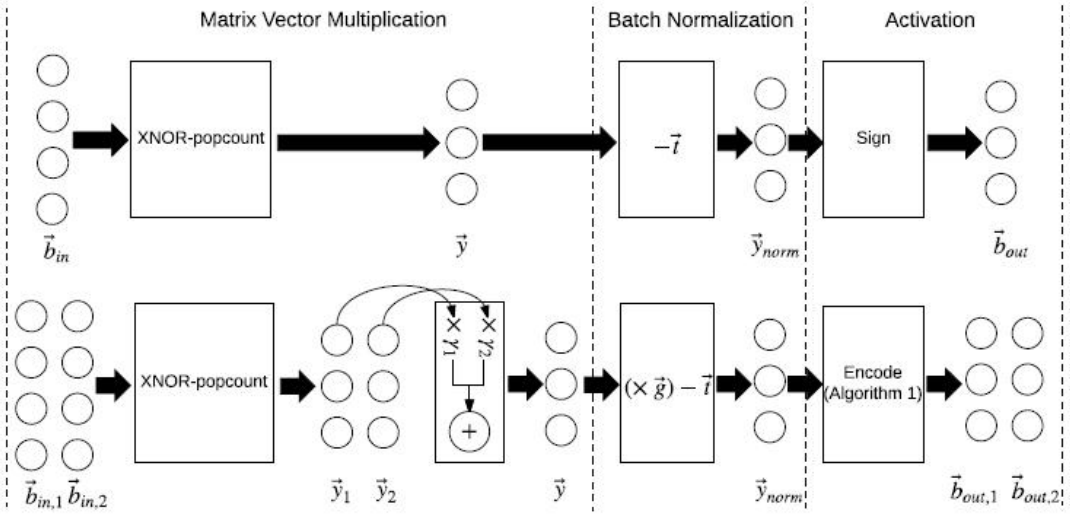

图 6：基线（顶部）和我们的修改（底部）二进制 CNN 层的硬件体系结构。

Benchmark	CNN Architecture
MNIST	$784\ (input) - D256 - BN - RB - D256 - BN - RB - D256 - BN - RB - D10 - BN - Softmax$
CIFAR10 & SVHN	$3 \times 32 \times 32\ (input) - C64 - BN - RB - C64 - BN - RB - MP - C128 - BN - RB - C128 - BN - RB - MP - C256 - BN - RB - C256 - BN - RB - D512 - BN - RB - D512 - BN - RB - D10 - BN - Softmax$

表 1：评估基准网络架构。 C64 表示 64 输出通道的 3 3 卷积，M P 表示 2 2 最大池，BN 表示批量归一化，D512 表示具有 512 个输出的稠密层。 剩余二值化使用 RB 显示。。

Benchmark	Binarynet (Courbariaux et al. (2016))			FINN (Umuroglu et al. (2017))			ResBinNet				
	# Epochs	Accuracy	Size (Mbits)	# Epochs	Accuracy	Size (Mbits)	# Epochs	Size (Mbits)	Accuracy (1-level)	Accuracy (2-level)	Accuracy (3-level)
CIFAR-10	500	89.85%	5.24	NA	80.1%	1.5	50+1	1.5	76%	83.5%	84.6%
SVHN	200	97.47%	5.24	NA	94.9%	1.5	10+1	1.5	95.2%	96.9%	97.1%
MNIST	1000	99.04%	52.7	NA	95.83%	0.3	30+1	0.3	97.3%	97.9%	98.1%

表 2：模型大小的比较，训练时期的数量和准确性。

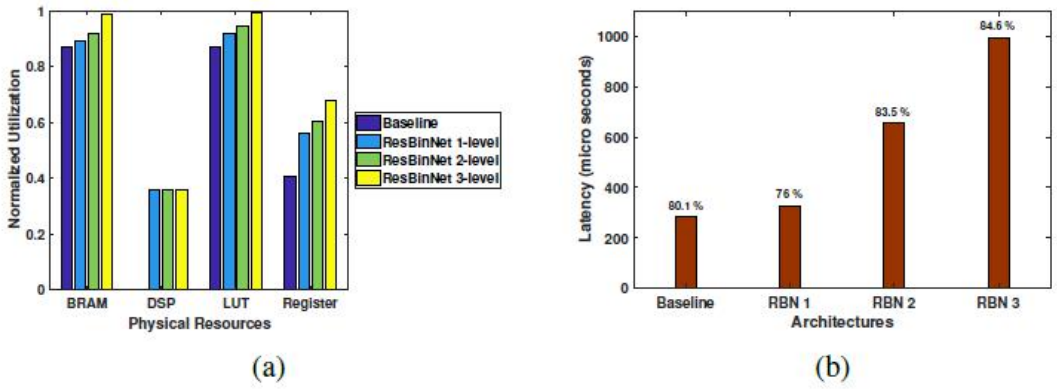

图 7：(a)在 Xilinx ZC706 评估套件上实现的具有不同残差水平的 ResBinNet 与基线设计(Umuroglu 等(2017))的资源利用开销。 （b ）ResBinNet 提供的不同残差水平下的延迟精度折衷。

✂----------