基于误差模型的权重二值神经网络近似加速

朱新忠[1],程利甫[1,2],吴有余[2],林闽佳[1],胡汝豪[1]
(1.上海航天电子技术研究所,上海 201109;2.清华大学 集成电路学院,北京 100047)

摘　要: 针对智能识别系统精确度和硬件复杂度之间的均衡设计问题,提出了一种基于误差统计模型的权重二值神经网络近似加速方法。在提出了一种获得高精度轻量神经网络的权重二值化处理算法基础上,引入近似加法器、消除乘法器以进一步提高能效。最终提出了一种系统级误差统计模型用于系统评估和优化设计,该设计能够分析和预测权重二值神经网络近似加速系统的最终精度。结果表明:该模型可以准确地预测系统精度,与仿真结果对比,相对误差在 2.05%~3.07%。该模型预测用于指导相应软硬件的设计优化,可大幅提高设计的迭代速度。

关键词: 近似计算; 近似加法器; 高能效计算; 统计误差模型; 权重二值化神经网络
中图分类号: TN 47; TP 302.1　　**文献标志码:** A　　**DOI:** 10.19328/j.cnki.2096-8655.2021.04.004

Error Model Based Approximate Computing Design for Binarized Weight Neural Network System

ZHU Xinzhong[1], CHENG Lifu[1,2], WU Youyu[2], LIN Minjia[1], HU Ruhao[1]
(1.Shanghai Aerospace Electronic Technology Institute, Shanghai 201109, China;
2.Department of Microelectronics and Nanoelectronics, Tsinghua University, Beijing 100047, China)

Abstract: In order to solve the problem of design tradeoff between accuracy and hardware cost for smart recognition systems, a binarized weight neural network acceleration method based on approximate adders and a statistical error model is introduced. On the basis of a lossless quantization algorithm to train binarized weight neural networks, replacing multipliers with approximate adders are proposed to further improve the energy-efficiency. A statistical system-level error model is finally innovated to predict the system accuracy and guide the design optimization. The experimental results show that the proposed error model can accurately predict the system, and the relative error is only within the range from 2.05% to 3.07% compared with the simulated results. This indicates that the prediction model can be used to guide future hardware-software optimization and can greatly improve the iteration speed of the design.

Key words: approximate computing; approximate adder; high energy-efficiency computing; statistical error model; binarized weight neural network

0　引言

当前,在航天系统中需要进行很多的图像或语音识别工作,在处理较为简单的语音任务,尤其是指令任务时,对系统实时性、高能效的要求越来越高。而深度学习已被多媒体广泛用于处理应用程序,包括图像、视频和语音的识别和分类等,其所在硬件平台也在不断发展和演进。对于航天系统而言,神经网络也逐渐被采用到简单的分类任务之中,如关键词语音命令的识别(Keyword Spotting and Recognition, KWSR)。对于网络结构逐渐复杂

收稿日期: 2020-09-19; **修回日期:** 2020-12-04
基金项目: 上海市优秀技术带头人计划(19XD1431400)
作者简介: 朱新忠(1975—),男,硕士,研究员,主要研究方向为综合电子、有效载荷数据处理与存储的研发。
通信作者: 程利甫(1985—),男,博士研究生,高级工程师,主要研究方向为星载微系统、综合电子、有效载荷的研发和应用。

的深度神经网络来说,其加速所需要的硬件能耗随着网络规模的增加而迅速增加,因此,近年来近似计算和更简单的权重二值化神经网络(Binarized Weight Neural Network, BWNN)结构逐渐被引入到实时性要求高的识别加速过程中。KWSR往往应用在物联网、手机或其他基于电池的边缘智能设备中,由于功耗和面积非常敏感,因此,简化的多层深度神经网络广泛地被应用于处理输入数据,而这些技术成熟度较高,逐渐也被航天系统所采用。

在最近几年的发展中,多类深度神经网络被应用于KWSR或者相关的航天系统中,包括深度神经网络(Deep Neural Network, DNN)[1-2]、卷积神经网络(Convolutional Neural Network, CNN)[3-4]、基于长期和短期记忆的递归神经网络(Long-Short Term Memory-Recurrent Neural Network, LSTM-RNN)[5-6]、基于门控循环单元(Gate Recurrent Unit Network, GRUN)的神经网络[7]和卷积递归神经网络(Convolutional Recurrent Neural Network, CRNN)[8]。基于深度神经网络的KWSR提高了语音的鲁棒性,但是其所包含的大量参数和引入的运算会产生大量在存储和计算方面的硬件开销。对于多层神经网络的压缩而言,量化是最为常用的方法之一。因此,通过探索和分析不同神经网络结构和压缩方法,BWNN被发现可以用于实现超低功耗的KWSR[9-13]。其与传统神经网络的区别在于,传统的神经网络权重均为16 bit或者更高的位宽,而这一网络仅需要1 bit位宽的权重即可实现高精度的识别,即:BWNN将权重和隐藏层二值化,激活值设为+1或-1。这样的结构大大降低了存储压力和片上带宽压力,也因为1 bit的位宽,几乎将网络中的乘法运算消除,仅需要优化加法运算的硬件实现。

本文提出了一个面向BWNN的基于逐位量化的KWSR网络,针对KWSR中的近似加法器进行了优化设计。针对近似计算引入的误差,需要一个系统性的评估方法,本文提出了一种统计意义的误差分析模型,可用于预测近似系统对神经网络的加速效果。具体来说,使用本文的BWNN量化方法,对不同种类的神经网络进行二值化并测试其精度,从中选取最适合的网络结构进行量化。随后,通过提出的误差统计模型,本文使用建模为软件仿真的近似加法器进行神经网络加速的精度评估。通过与功能仿真结果进行比较,本文的误差统计模型预测精度很高,最终的系统误差预测和真实系统误差对比,相对偏差约在3%以内。

1 原理分析

对BWNN而言,一方面其权重占用的存储空间可以大大减少;另一方面可以使用位运算代替常规神经网络中的乘法操作,这样可以减少大多数乘法运算。总之,通过建立BWNN,只需要加法器就可执行几乎所有的操作,因此,我们后续对硬件的分析也集中在加法器模型上。

1.1 BWNN系统的训练方案设计

传统对神经网络进行二值化的方法是在获得定点神经网络后进行截断并微调,这样的方式会不可避免地降低识别精度。基于权重位宽均为1 bit的XNOR-Net的量化原理,本文提出了一种逐位量化的权重二值化方法。这一方法在网络的训练过程中介入,而非对最终的训练结果二值化,从而减少KWSR的准确率。

量化的具体方法如下:

$$Q_{\text{quantize}_k}(x_i)=\frac{1}{2^k-1}R_{\text{round}}\left(x_i\times(2^k-1)\right) \tag{1}$$

$$f(x)=\frac{T_{\tanh}(x)}{2\times\max(|T_{\tanh}(x)|)}+\frac{1}{2} \tag{2}$$

$$w_{\text{q}}=2\times Q_{\text{quantize}_k}\left(f\left(w_i\right)\right)-1 \tag{3}$$

式中:w_i为第i层神经网络的权重数值;k为目标的量化位宽数值;$Q_{\text{quantize}_k}(\cdot)$、$f(\cdot)$为量化函数和压缩函数;$w_{\text{q}}$为对应的权重量化数值结果。

因此,对于任意一层的神经网络层,均有对应的量化结果。

$$z_{\text{q}}=x_i\times w_{\text{q}}+b_i \tag{4}$$

$$z_{\text{q}}=x_i\times\left\{2\times Q_{\text{quantize}_k}\left(f\left(w_i\right)\right)-1\right\}+b_i \tag{5}$$

式中:x_i为当前神经网络层的输入;b_i为量化前的偏置量;z_{q}为当前神经网络层的输出数值。

本文所述的逐位量化算法流程如图1所示。

在第k比特位宽度($k>1$),输入层和批处理归一化(Batch Normalization, BN)层将同时量化。实际上,由于BN层包含数据压缩处理,激活函数tanh的量化可以被舍弃,因此,压缩函数$f_c(\cdot)$可以按以下方式优化:

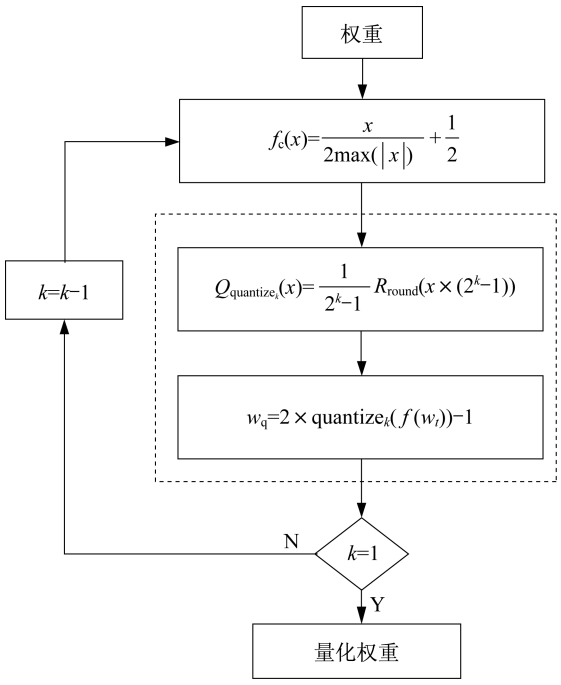

图 1　神经网络的逐位量化算法流程图

Fig.1　Flow chart of the bit-by-bit quantization method for neural networks

$$f_c(x)=\frac{x}{2\times\max(|x|)}+\frac{1}{2} \tag{6}$$

在整个量化过程中，权重是首先压缩为 0～1 之间的数据。压缩数据由式(1)和式(3)得出。随后，权重量化为无损定点[－1，1]之间的数字。为了使量化权重在训练过程中更接近理想值，在处理过程中采用逐位量化的方法进行反复训练。第一次训练量化的比特位宽度和训练后的权重会保存下来以用于重新训练，并且量化的位宽在下次训练中逐渐降低。例如，量化位宽可以是从 8 bit 宽度逐渐量化到 4 bit 宽度，然后 2 bit 宽度最终量化为 1 bit 宽度。这样渐进式权重训练的最有利之处在于速度快，且可以提高权重的训练效率和可靠性。

1.2　近似加法器的概率误差分析

1.2.1　误差评估量纲

为了使描述清晰，我们首先定义准确值为 L_{accu}，近似值作为 L_{appro}。

最大误差量纲(Maximum Error Magnitude，MEM)即最大误差，为准确值与近似值差值的绝对值，公式如下：

$$E_{\text{rror}_{\max}}=\max|L_{\text{accu}}-L_{\text{appro}}| \tag{7}$$

相对误差量纲(Relative Error Magnitude，REM)即相对误差，为准确值、近似值差值的绝对值和准确值绝对值的比，公式如下：

$$E_{\text{rror}_{\text{relate}}}=\left|\frac{L_{\text{accu}}-L_{\text{ppro}}}{L_{\text{accu}}}\right| \tag{8}$$

平均误差量纲(Average Error Magnitude，AEM)为绝对差大小介于精确值和近似值之间所有差值的平均数，平均误差满足如下公式：

$$E_{\text{rror}_{\text{avg}}}=\frac{\sum|L_{\text{accu}}-L_{\text{ppro}}|}{\text{数值个数}} \tag{9}$$

均方误差量纲(Mean Squared Error Magnitude，MSEM)为在所有可能的精确值与近似值之间的大小距离值上取平均，平方误差度量公式如下：

$$E_{\text{rror}_{\text{MSE}}}=\frac{\sum(L_{\text{accu}}-L_{\text{ppro}})^2}{\text{数值个数}} \tag{10}$$

1.2.2　低延迟近似加法器模型

基于文献[9]中的研究内容，代表基于块的通用模型加法器结构如图 2 所示。输入位分为多个不相交或重叠的子加法器。每个子加法器产生相应输入的输出部分和，同时使用前面子加法器的输出进位来生成结果。

文献[6]中提出的误差模型如下：

$$\begin{aligned}\Pr[E]&=\Pr\left[E_2\vee\cdots\vee E_L\right]=\\&\sum_{i=2}^{L}\Pr\left[E_i\right]-\sum_{2\leqslant i\leqslant j\leqslant L}\Pr\left[E_i\wedge E_j\right]+\\&\sum_{i=2}\Pr\left[E_i\wedge E_j\right]-\cdots+(-1)^L\Pr\left[\bigwedge_{i=2}^{L}E_i\right]\end{aligned} \tag{11}$$

式中：E_i 为二进制变量，当第 i 个子加法器错误时，$E_i=1$，否则 $E_i=0$。考虑任何第 i 个加法器，当 $2\leqslant i\leqslant L$，$E_i=1$，会有

$$\Pr\left[P_i;\ N\right]=\Pr\left[\bigwedge_{i=2}^{L}A\oplus B=1\right]=\frac{1}{2} \tag{12}$$

$$\Pr\left[G_i;\ K\right]=\Pr\left[A_{1-K}+B_{1-K}\geqslant 2^K\right]=\frac{1}{2}-\frac{1}{2^{K+1}} \tag{13}$$

式中：$A_{1-K}+B_{1-K}$ 为没有输入到第 i 个子加法器的较低比特位置；$\Pr\left[P_i;\ N\right]$ 为之前的子加法器生成的进位数值参与计算；$\Pr\left[G_i;\ K\right]$ 为之前的较低有效位在第 i 个子加法器产生一个进位；N 为加法器的位宽；K 为产生进位的低比特数据位宽；$\oplus$ 为异或运算符；P_i 为第 i 个子加法器产生了进位这一事件；G_i 为第 i 个子加法器的低比特位产生进位这一事件。

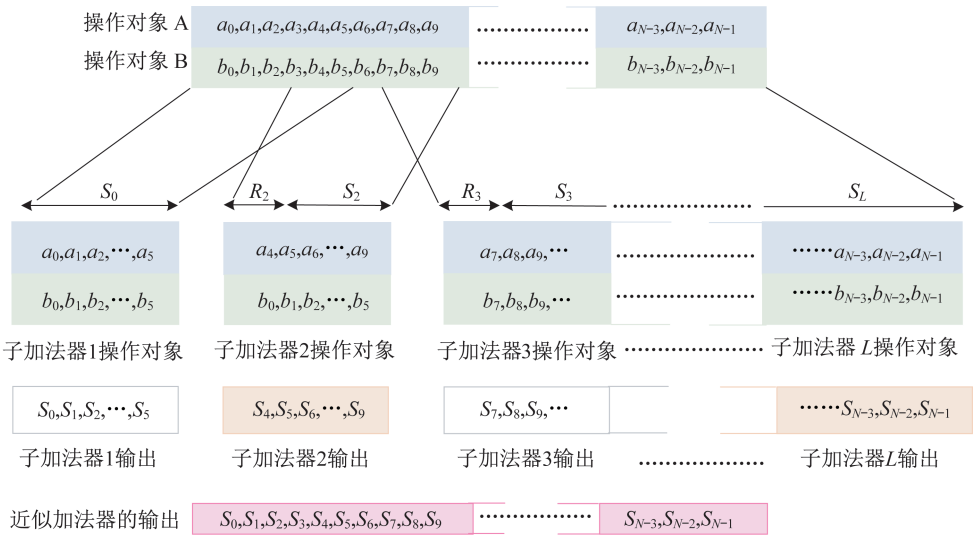

图 2　基于块的通用模型加法器结构[6]

Fig.2　Structure of generic block-based approximate adder [6]

1.3　近似加法器阵列的误差统计模型

对于由近似加法器组成的近似计算阵列，可分为两种不同的情况：如果加法器用于不同的计算源，例如不同的神经元，其误差统计模型则为单独考虑，近似加法器阵列的误差模型是所有加法器的最大误差，即 MEM 值；如果近似加法器形成一个累加结构，将阵列的误差模型视为所有加法器的平均误差，即 AEM 值。

2　实现方案

在这一部分，进行了以下实验。首先，本文为 KWSR 系统设计了各种网络，见表 1，它们由不同的层组成；然后，将所有网络通过前述的方法进行二值化，对于模型验证，使用 Matlab 模拟近似加法器的功能行为并获得 BWNN 的准确性；最后，将误差模型引入 BWNN 以获得模型输出精度，两者精度在本文末尾进行了比较。

表 1　深度神经网络的实现方案与对比

Tab.1　Implementation schemes and specifications of DNN models

模型类型	模型参数	乘法运算量/MOPS	加法运算量/MOPS	计算精度
CNN1	4CONV＋1FC	0	8.4	87.4%(4 bit)
CNN2	4CONV＋1FC	0	9.3	87.9%(4 bit)
LSTM1	LSTM(118)	3.6	3.6	90.0%(4 bit) N/A(1 bit)
LSTM2	LSTM(214)	10.2	10.2	91.0%(4 bit) N/A(1 bit)
GRU1	GRU(154)	2.6	2.6	91.2%(4 bit) N/A(1 bit)
GRU2	GRU(250)	10.2	10.2	91.8%(4 bit) N/A(1 bit)
CRNN1	1CONV＋2GRU＋1FC	3.2	3.2	91.11%(4 bit) N/A(1 bit)
CRNN2	1CONV＋2GRU＋1FC	9.3	9.3	92.26%(4 bit) N/A(1 bit)

2.1　BWNN 的设计

使用 Google 语音命令数据集(Google Speech Command Dataset，GSCD)作为训练集和验证集。GSCD 中有 10.5 万组 1 s 长的音频数据，集中包含

35个关键字的片段。当训练神经网络时，我们将卷积层的权重和完全连接的层量化至1 bit位宽。BWNN模型经过培训，可以将音频识别分类为10个关键字之一，“YES”“NO”“UP”“DOWN”“RIGHT”“LEFT”“ON”“OFF”“STOP”“GO”“沉默”（即不说任何话）和“未知”（即所说词语不在10个关键词以内）。

表1总结了所测试的神经网络的层次类型、计算要求和准确性，其中部分结构来源于文献[5-6, 14-16]中采用GSCD进行KWSR命令的网络。这些网络架构的权重都进行了二值化测试。其中缩写的含义：CONV为卷积层，FC为全连接层，LSTM为LSTM单元的个数，GRU为门递归单元个数。

表中可见，LSTM、GRU和CRNN都比CNN的精度更高，但是它们在二值化后无法得到收敛的结果，即无法进行二值化。而为了提高语音识别的鲁棒性并降低电路的功耗，可以适当牺牲网络的识别精度，同时还需要控制识别精度高于85%。因此，CNN是适用于BWNN的结构。

2.2　二值化权重神经网络的实现

由于权重在整个KWSR系统中加载后，计算过程将不会更改，而数据将在整个操作过程中不断变化，需要减少权重占用的存储和数据刷新速度以进一步降低功耗。因此，需要根据网络规模来评估和优化网络卷积核、全连接层的权重第1和第2卷积层的输出。本文将3×3卷积核用于卷积运算，同时减少数据位宽并防止数据溢出。而最后卷积层的输出结果直接影响第一全连接层的权重大小。因此，本文减少了卷积层的卷积核数量，并增加卷积核的步幅以减少输出结果的大小。

3　验证结果与分析

基于上述的方法，在实现了4个BWNN之后，评估了错误在软件仿真结果和模型输出结果之间，见表2。对于提出的4个BWNN，仿真系统精度和预测系统精度之间的相对差异约为2%～3%。结果表明，采用的误差统计模型可以预测本文所叙述的计算系统的精度。其中4个网络的拓扑结构阐述如下：BWNN 1～4均由4层卷积、1层全连接（30个神经元）组成，卷积层参数（通道数、卷积核的三维尺寸、步长）见表1。通过4种不同的卷积网络设计，可以应对不同复杂度的分类场景，针对不同步长、不同通道数均进行了验证，体现了模型的精确性和普遍适用性。

表2　系统误差的预测和实测对比

Tab.2　Predicted and simulated accuracies

网络类型	网络层参数	仿真系统精度/%	模型预测精度/%	精度相对误差/%
BWNN1	卷积1(48,3,3,2,1) 卷积1(28,3,3,1,1) 卷积1(28,3,3,1,1) 卷积1(18,3,3,1,1) 全连接层(30)	91.6	88.8	3.05
BWNN2	卷积1(42,3,3,2,1) 卷积1(34,3,3,1,1) 卷积1(26,3,3,1,2) 卷积1(18,3,3,2,1) 全连接层(30)	91.3	88.5	3.07
BWNN3	卷积1(28,3,3,2,1) 卷积1(24,3,3,1,1) 卷积1(16,3,3,1,2) 卷积1(12,3,3,2,1) 全连接层(30)	87.8	86.0	2.05
BWNN4	卷积1(20,3,3,2,1) 卷积1(24,3,3,1,1) 卷积1(16,3,3,1,2) 卷积1(12,3,3,2,1) 全连接层(30)	87.3	85.7	2.29

4 结束语

本文提出了系统的误差统计模型,可用于BWNN在近似加法器的近似加速系统中。本文为KWSR提出了二进制加权神经网络的量化方法,参考了近似加法器的基本误差模型并针对KWSR系统进行了优化。此外,对面向10个命令词识别的网络进行实验,并将其二值化为BWNN。通过使用误差统计模型,可以预测BWNN的系统精度。通过比较仿真结果和模型预测的系统精度,本文提出的方法可以实现3%以内的精度预测相对损失。这一工作对后续航天系统中KWSR的系统设计,提供了有力的工具。

参考文献

[1] CHEN G G, PARADA C, HEIGOLD G. Small-footprint keyword spotting using deep neural networks [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2014: 4087-4091.

[2] 李索,张支勉,王海鹏.基于深度学习算法的极化合成孔径雷达通用分类器设计[J].上海航天,2018,35(3):1-7.

[3] SAINATH T N, PARADA C. Convolutional neural networks for small-footprint keyword spotting [C]// Sixteenth Annual Conference of the International Speech Communication Association. 2015: 1478-1482.

[4] 张锐,王兆魁.基于深度学习的空间站舱内服务机器人视觉跟踪[J].上海航天,2018,35(5):1-9.

[5] SUN M, RAJU A, TUCKER G, et al. Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting [C]// IEEE Spoken Language Technology Workshop (SLT). Washington D. C., USA: IEEE Press, 2016: 474-480.

[6] 赵霞,白雨,倪颖婷,等.基于深度学习的语义分割算法综述[J].上海航天, 2019,36(5):71-82.

[7] ZHANG Y D, NAVEEN S, LAI L Z, et al. Hello edge: keyword spotting on microcontrollers [DB/OL]. (2018-02-14) [2020-08-27]. https://arxiv. org/pdf/1711.07128.pdf.

[8] ARIK S O, KLIEGL M, CHILD R, et al. Convolutional recurrent neural networks for small-footprint keyword spotting [DB/OL]. (2017-03-15) [2020-08-27]. https://arxiv. org/ftp/arxiv/papers/1703/1703.05390.pdf.

[9] REDA S, MUHAMMAD S. Approximate circuits: methodologies and CAD [M]. Switzerland: Springer, 2019: 236-242.

[10] SIMONS T, LEE D J. A review of binarized neural networks [J]. Electronics, 2019, 8(6): 661.

[11] LI, J J, WANG Y, LIU B S, et al. Simulate-the-hardware: training accurate binarized neural networks for low-precision neural accelerators [C]// Proceedings of the 24th Asia and South Pacific Design Automation Conference. Washington D. C., USA: IEEE Press, 2019: 323-328.

[12] LIU B, SUN Y H, CAI H, et al. An ultra-low power keyword-spotting accelerator using circuit-architecture-system co-design and self-adaptive approximate computing based BWN [C]// Proceedings of the 2020 on Great Lakes Symposium on VLSI. Washington D. C., USA: IEEE Press, 2020: 193-198.

[13] LIU B, CAI H, WANG Z, et al. A 22 nm, 10.8 μW/15.1 μW dual computing modes high power-performance-area efficiency domained background noise aware keyword-spotting processor [J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2020, 67(12): 4733-4746.

[14] LIU B, CAI H, GONG Y, et al. Binarized weight neural-network inspired ultra-low power speech recognition processor with time-domain based digital-analog mixed approximate computing [C]// IEEE International Symposium on Circuits and Systems, Virtual. Washington D. C., USA: IEEE Press, 2020: 1-5.

[15] YANG M H, YEH C H, ZHOU Y Y, et al. A 1 μW voice activity detector using analog feature extraction and digital deep neural network [C]// IEEE International Solid-State Circuits Conference. Washington D. C., USA: IEEE Press 2018: 346-348.

[16] WARDEN P. Speech commands: a public dataset for single-word speech recognition [DB/OL]. (2018-04-09) [2020-08-27]. https://arxiv. org/pdf/1804.03209.pdf.