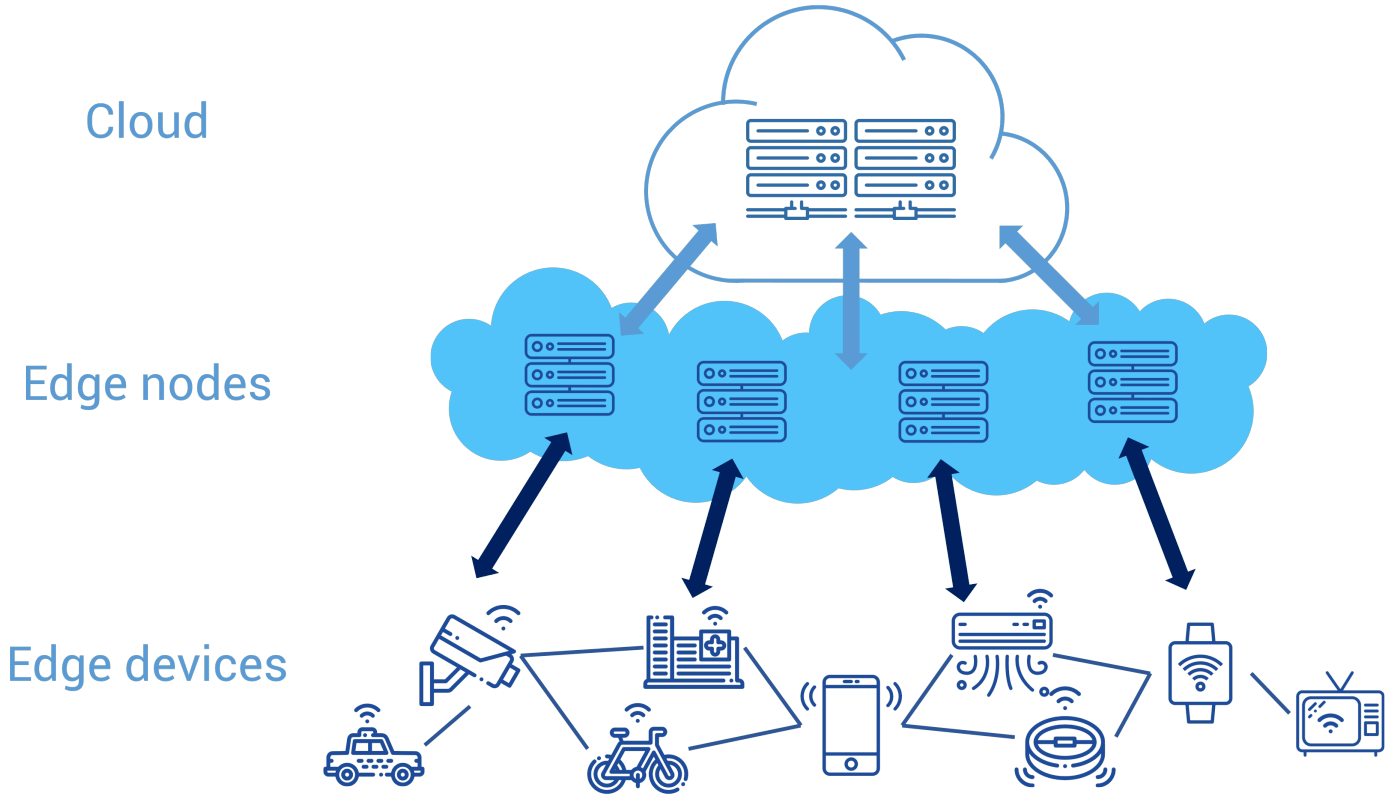
POSIT™ Arithmetic for Edge Computing



Arithmetic is a key component and is ubiquitous in today's digital world, ranging from embedded to high performance computing systems. With machine learning at the fore in a wide range of application domains from wearables to automotive to avionics to weather prediction, sufficiently accurate yet low-cost arithmetic, low power is the need for the day.

Andre Guntoro,
Robert Bosch (Germany)

Introduction

Computer arithmetic is ubiquitous in applications ranging from an embedded domain such as smartphones to high-performance computing (HPC) applications like weather modeling. Specifically, in embedded systems, since the platforms have performance limitations due to limited power and area budgets, using appropriate arithmetic is desirable.

Challenges

The constraints on power and area footprints combined with demand for high performance in applications like edge computing in internet-of-things (IoT) have created tremendous challenges for computer architects. Over the years in response to this challenge, engineers have developed hardware-efficient implementations of computer arithmetic but they are facing memory bottlenecks as communication bottlenecks over the network.

Approach

1) *Task accuracy*: We retrain the Deep Neural Networks (DNNs) over 5 epochs with 10 different multipliers. The results is presented in Figure. 1. For ResNet20, we observe a uniform accuracy recovery, reaching the defined accuracy tolerance in 70% of the cases, which corresponds to multiplier with energy saving of up to 56.8%. In case of keyword spotting, the accuracy tolerance is reached in all cases, with maximum energy saving of 68%, although the accuracy slightly decreases for approximate multipliers with MRE higher than 2.4%.

2) *Effects of data augmentation*: Approximate computational units introduce unwanted noise in DNN operations through approximation errors. Such noise, in controlled amounts, acts as a regularizer when training approximate DNNs. Thus, we propose DNN retraining without data augmentation, as this is also another form of regularization through input alteration. By performing data augmentation, the DNN approximation error is then harder to compensate. We compare the results of training with and without data augmentation in Figure. 1. For image classification, we randomly flip the training samples, and for keyword spotting, we add background noise with a volume of 10% to the initial time series. As observed, data augmentation worsens the accuracy degradation in approximate DNNs, specially for speech recognition tasks.

3) *Conclusions*: The paradigm of approximate computing delivers promising results for optimizing energy consumption of perception tasks. In this work, we present an unprecedented analysis of hardware-oriented approximate computing for speech recognition tasks, and highlight the role of data augmentation and regularization through approximation error through experiments in DNNs for image recognition and keyword spotting.

Benefits

Figure. 2 summarizes the decimal accuracy of 16-bit representations as a function of the magnitude (log base 10) of the absolute value of the value represented by the format. Fixed-point (integer) format is the simplest and fastest format, but has very unbalanced accuracy about low magnitudes and a very restricted dynamic range. The float values have flat accuracy except for the subnormal number range on the left, where accuracy tapers to zero. For the most common values in the range of about 0.01 to 100, POSITs have higher accuracy than IEEE floats and bfloats, but less accuracy outside this dynamic range. For all precisions, float accuracy forms a trapezoidal shape; fixed point accuracy looks like a triangular ramp upward; and POSIT accuracy is an isosceles triangle centered at magnitude zero. Depending on the applications, POSITs often maximize information-per-bit in the Shannon sense, compared to the other formats.

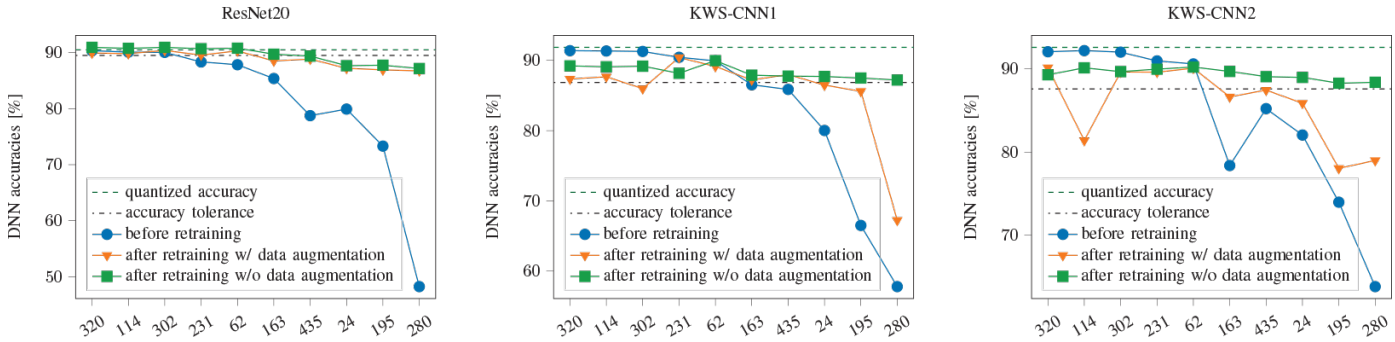


Figure 1: Task accuracy with 10 different approximate multipliers on 3 DNNs

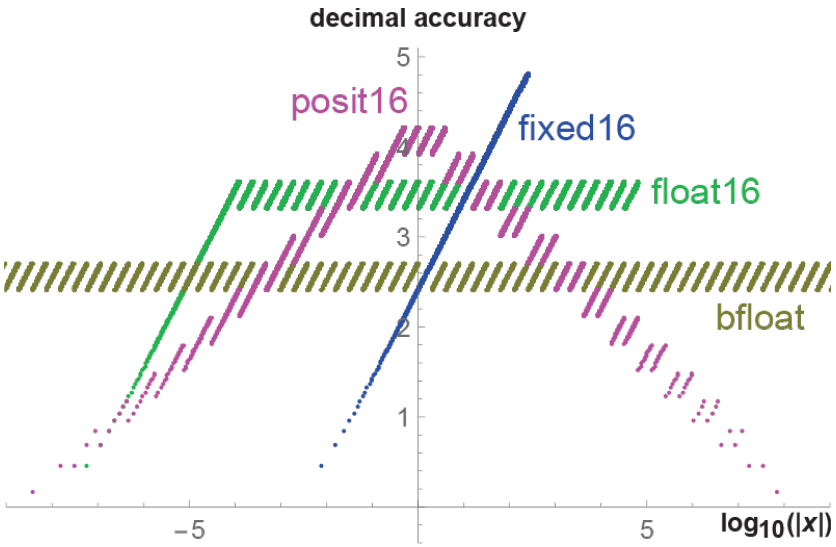


Figure 2: Accuracy as a function of bit string for 16-bit formats

Company Description

Robert Bosch Research and Technology center conducts research on next generation edge computing, automotive and IoT technologies. Its operations are divided into four business sectors: Mobility Solutions, Industrial Technology, Consumer Goods, and Energy and Building Technology. As a leading IoT provider, Bosch offers innovative solutions for smart homes, Industry 4.0, and connected mobility. Bosch is pursuing a vision of mobility that is sustainable, safe, and exciting.