Analog AI BERT FactSheet

Overview

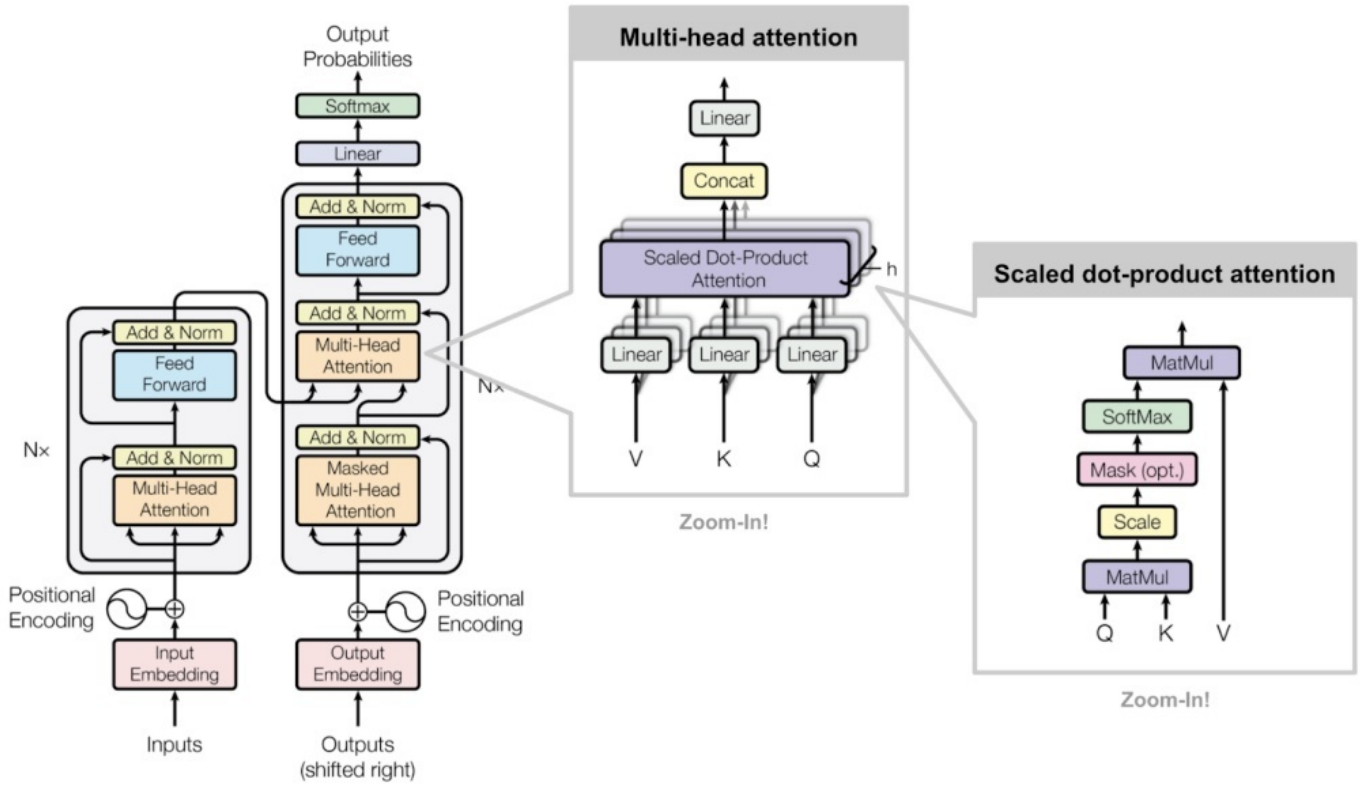
This document is a FactSheet accompanying the BERT model used in the IBM AI Hardware Center – with Analog AI Hardware. FactSheets aim at increasing trust in AI services through supplier's declarations of conformity and this FactSheet documents the process of training the Audio Classifier model as well as its expected results and appropriate use.

Purpose

BERT stands for Bidirectional Encoder Representations from Transformers. As the name suggests, this is based on the Transformer architecture. It is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context. Pre-training is performed for two NLP tasks, namely, (a) Masked Language Modeling and (b) Next Sentence Prediction. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks. These can include sentiment classification, abstract summarization, question answering and conversational response generation to give a few examples.

Model Architecture

The figure below shows the nominal architecture of BERT models – these consist of many layers of Encoder layers. BERT Base (Cased or uncased) model uses: 12 layers (transformer encoder blocks) with 12 attention heads, and a total of 110 million parameters (neural network weights). We have used the hugging face implementation of BERT (https://huggingface.co/docs/transformers/model_doc/bert) which is based on the "Attention is All You need" Google paper [1].



[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc.
https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Parameters and Inference Attributes

Total Number of weights/ops/Multiply Accumulate (MAC)

- 110 million parameters/weights
- Total number of operations will depend on the length of the input sequence (SL) as well as the batch size
- Analog AI accelerator prototypes are being designed for optimum performance for SL up to 128 and low batch sizes (1-4)
- For a SL of 32 (Batch Size = 4) total number of ops is approximately. 21.84B (out of which 21.74B are operation performed using Analog Non-Volatile Memory (NVM) arrays or tiles). Total execution time is approximately, 0.19msec

Reuse Factor for each node/layer

- The same weights are not re-used in this dataflow.

NVM Array Tile Size

- Each array tile has 512x512 synaptic weights. Synaptic weights are represented as a differential pair of conductances.

Power & Chip Area Estimates

- Average Power (Chip level): 3.13W

- Effective area: 780.4 sq.mm

Training Methodology

The model was trained using IBM's AiMOS supercomputer in a Hardware-aware (HWA) fashion before being deployed on Analog Accelerator hardware for Inference [2]. Hardware-aware training makes the model resilient to non-idealities of NVM devices as well as peripheral circuitry and leads to improved inference accuracy with Analog compute. Specifically, this is done by starting with a floating point (FP) trained baseline model and re-training it further with the addition of device & circuit non-idealities in the forward pass only. The backward pass and weight update are similar to traditional FP training. These can include inputs such as programming (or weight) noise, read noise, conductance drift, ADC (or output) noise, input & output bounds.

The following parameters were used for HWA training of BERT-base model:

- Weight Noise (Std-dev): 1.0
- Read Noise: 0.0175
- Analog-Digital Converter (ADC)/Output noise: 0.04
- Drift model from Phase Change Memory (PCM) devices
- Input & Output bounds: 3.0/10.0
- Learning rate: 5E-5 to 1.5E-4

This hardware-aware training produced a model with a validation accuracy of 80.6% (GLUE average) (compared to state-of-the-art at 81.2% from BERT paper).

[2] 2020 Nature Communications. Accurate deep neural network inference using computational phase-change memory

Inference Methodology

- Number of non-idealities can impact the weight programming & stability such as programming noise, read noise, conductance drift and temperature dependent variations. These parameters can be dependent on the weight (& conductance) distributions and are thus model & network dependent.
- Some of these non-idealities are introduced during the hardware-aware training methodology described above to make the model more resilient and to improve inference accuracy.
- Conductance variations post programming such as due to drift and temperature are mitigated by algorithmic corrections to minimize the impact on inference accuracy.

Training Data

The model was trained on the GLUE benchmark dataset (8 tasks: RTE, MRPC, STS-B, CoLA, SST-2, QNLI, QQP, MNLI). It has a size from 2.5k (RTE) to 393k (MNLI). All tasks are classification, except STS-B, which is a regression task. MNLI has 3 classes and all others have 2 classes. Sequence length goes up to 128 for our purposes.

Characteristics of the training hardware and characteristics of the inference hardware

AiMOS supercomputer and Cognitive Computing Cluster at IBM Research (CCC) was used. Batch Size is optimized for GPU Memory.
For both training of the model and inference accuracy evaluations, a computing node with a single NVIDIA V100 GPU was used on the CCC cluster.

Test Data

The model was tested on the GLUE benchmark dataset (8 tasks: RTE, MRPC, STS-B, CoLA, SST-2, QNLI, QQP, MNLI). It has a size from 2.5k (RTE) to 393k (MNLI). All tasks are classification, except STS-B, which is a regression task. MNLI has 3 classes and all others have 2 classes. Sequence length goes up to 128.

Performance Metrics

- *Accuracy*: Most tasks use accuracy; MRPC and QQP use both accuracy and F1 score; CoLA uses Matthews corr.; STS-B uses Person/Spearman corr.

Task names	RTE	MRPC	STS-B	CoLA	SST-2	QNLI	QQP	MNLI	Avg.
Data set size	2.5k	3.7k	5.7k	8.5k	67k	105k	364k	393k	
Metrics *		(f1)	(Spearma)	(Matthew')					
BERT paper (digital)	66.40%	88.90%	85.80%	52.10%	93.50%	90.50%	89.20%	83.40%	81.23%
Analog AI	64.00%	88.62%	84.52%	55.65%	91.03%	89.01%	89.63%	82.12%	80.57%
* Metrics is accuracy, unless specified, e.g., f1, Spearma Co, or Matthew's Co.									

- ***Classification Time***: 0.19msec assumes a Sequence Length of 32
- ***Energy Efficiency***: 6589 (Inferences/s/W)
- ***How often do you have to reprogram the weights/refresh the model:*** Due to non-idealities such as conductance drift and retention loss, it is expected that models would need to be refreshed (re-programmed) on Analog HW every 4 weeks.