Yan-min Qian, Xu Xiang, 2019. Binary neural networks for speech recognition. *Frontiers of Information Technology & Electronic Engineering*, 20(5):701-715. https://doi.org/10.1631/FITEE.1800469

Binary neural networks for speech recognition

Key words: Speech recognition; Binary neural networks; Binary matrix multiplication; Knowledge distillation; Population count

Corresponding author: Yan-min Qian
E-mail: yanminqian@sjtu.edu.cn
ORCID: http://orcid.org/0000-0002-0314-3790

Motivation

- Deep neural network (DNN) based acoustic models are better than traditional GMM based models. However, a DNN based model is difficult to directly deploy on low-power embedded devices due to its high computational cost.
- The most computationally expensive operation in common DNN models is matrix multiplication, which can be sped up largely using low-precision data types, for example, 1-bit data type (binary matrix).
- The performance (speed and accuracy) of binary DNN acoustic models for speech recognition on various hardware platforms (CPU, GPU) needs to be investigated and improved.

Main idea

- For DNNs and CNNs used as the acoustic models in speech recognition, to speed up the inference, an extremely low precision data type (binary values) is adopted.
- Bit operations can be facilitated to speed up binary matrix multiplication on various hardware platforms.
- The accuracy of binary models can be improved using knowledge distillation with floating-point models.

Method

1. Train the binary version acoustic models using back-propagation with the designed activation functions and algorithms.
2. Implement fast binary matrix multiplication on CPU and GPU based on bit operations.
3. Improve the accuracy of binary version acoustic models by knowledge distillation with the corresponding floating-point models.

Major results

- Binary matrix multiplication can run 5~7 times faster than floating-point matrix multiplication

Table 1 Speed comparison on an Intel i3-4150 CPU (single thread)

Size	FMM	BMM	Speedup
16, 2048, 2048	34.5	249.0	7.2×
2048, 2048, 2048	91.2	263.5	2.9×
TPP	112.0	448.0	4.0×

FMM: floating-point matrix multiplication; BMM: binary matrix multiplication; TPP: theoretical peak performance

Table 2 Speed comparison on a HiSilicon Kirin 950 CPU (single thread)

Size	FMM	BMM	Speedup
16, 2048, 2048	3.8	25.4	6.7×
2048, 2048, 2048	12.0	29.1	2.4×
TPP	18.4	≈ 39.5	2.1×

FMM: floating-point matrix multiplication; BMM: binary matrix multiplication; TPP: theoretical peak performance

Table 3 Speed comparison on NVIDIA Tesla P100 GPU

Size	FMM	BMM	Speedup
16, 2048, 2048	1.3×10^3	7.0×10^3	5.4×
2048, 2048, 2048	7.6×10^3	30.6×10^3	4.0×
TPP	9.3×10^3	49.6×10^3	5.3×

FMM: floating-point matrix multiplication; BMM: binary matrix multiplication; TPP: theoretical peak performance

Major results (Cont'd)

- Binary DNN or CNN model can keep the performance degradation below 15% of knowledge distillation

Table 8 WER comparison of different BDNN models with the knowledge distillation from the same FPDNN on Switchboard telephone speech recognition task

Student	Teacher	WER (%)			
		Hub5'00		RT03S	
		SWB	CH	FSH	SWB
FPDNN	–	15.6	27.9	20.7	30.2
BDNN-2048	FPDNN	19.7	33.0	25.1	35.1
BDNN-3072	FPDNN	18.6	31.5	23.6	33.6

WER: word error rate; FPDNN: full precision DNN; BDNN-2048/-3072: binary DNN with 2048/3072 hidden units

Table 9 WER comparison of different BCNN models with the knowledge distillation from the FPVDCNN on Switchboard telephone speech recognition task

Model	Teacher	WER (%)			
		Hub5'00		RT03S	
		SWB	CH	FSH	SWB
FPCNN	–	15.1	26.9	20.3	29.6
BCNN	–	21.1	34.7	26.8	37.1
BCNN	FPVDCNN	18.8	31.5	23.9	33.8
BCNN-FPFC	–	16.3	28.8	21.4	31.2
BCNN-FPFC	FPVDCNN	15.2	27.4	20.3	29.8

WER: word error rate; FPCNN: full precision CNN; BCNN: binary CNN; BCNN-FPFC: BCNN with full precision values on fully-connected layers; FPVDCNN: full precision very deep CNN

Conclusions

- The binary neural network acoustic model can be 3-4 times faster than its floating-point version during the inference with the help of bit operations.
- Experiments on the Switchboard speech recognition task demonstrated that the binary neural network based acoustic model has acceptable performance (increase of the word error rate is no more than 15%) when it is trained using knowledge distillation.