Multi-Mode LSTM Network for Energy-Efficient Speech Recognition

Junseo Jo, Seokha Hwang, Sunggu Lee, and Youngjoo Lee
Department of Electrical Engineering, POSTECH, Pohang, Korea,
youngjoo.lee@postech.ac.kr

Abstract— **We newly introduce a novel processing scenario of long short-term memory (LSTM) network for the energy-efficient speech recognition. Compared to the conventional single-mode processing based on the fixed computing scheme, the proposed LSTM processing contains multiple operating cells providing attractive tradeoff between the recognition accuracy and the energy consumption. For the case study, the state-of-the-art LSTM network is modified to have two types of processing cells, strong and weak cells, which are dedicated to the accuracy-aware and energy-aware LSTM sequences, respectively. By allocating as many weak cells with low energy as possible, experimental results show that the proposed work saves the energy consumption for speech recognition by 75% compared to the original network.**

Keywords; LSTM; Low-power architecture; Speech recognition;

I. INTRODUCTION

Recently, the LSTM based recurrent neural network (RNN), has received great interests due to its excellent performance in speech recognition systems by preserving temporal memories for long periods using forget, input and output gates [1]–[3]. Applying the bidirectional LSTM networks associated with the mel-frequency centrum coefficients (MFCC), DeepSpeech in [4] is one of the most accurate recognition systems, achieving a character-level error rate of less than 10%. After calculating the MFCC values, as depicted in Fig. 1, three fully-connected (FC) layers are firstly applied to find different features as many as possible, and then two LSTM networks having the different directions followed by additional FC layers are operated to extract the received character. Based on the word-level libraries, finally, the connectionist temporal classification (CTC) process eliminates redundant characters, making the output sentence.

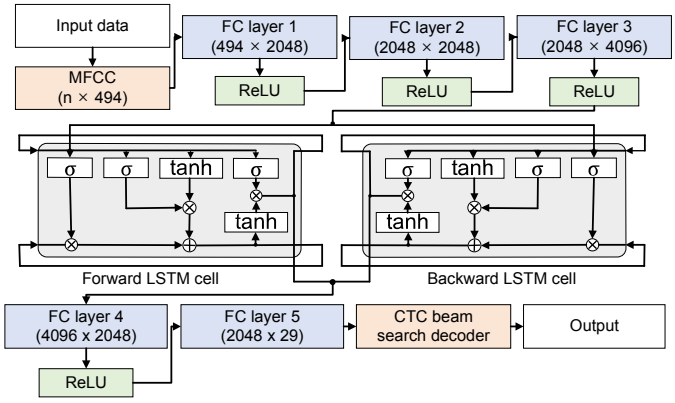

Fig. 1. DeepSpeech architecture

This work was supported by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-TB1703-07.

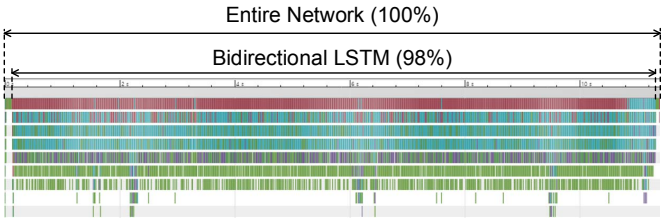

Fig. 2. Timeline of DeepSpeech processing

Although DeepSpeech network shows an attractive accuracy, however, it requires a huge amount of computational complexity caused by two *n*-length LSTM processing sequences, where *n* is determined by the number of windows issued to the MFCC calculator. Note that the value of *n* varies according to the length of the captured speech, and typically ranges from 500 to 1000 [5]. Therefore, the bidirectional LSTM in DeepSpeech requires more than 1000 same LSTM operations, each of which consists of numerous matrix multiplications as well as several look-up tables (LUTs) for nonlinear functions. As shown in Fig. 2, therefore, the bidirectional LSTM covers more than 98% of total processing costs, leading to the energy-starving recognition system. Therefore, it is urgent to develop the low-power LSTM processing scenario with the acceptable accuracy.

II. MULTI-MODE LSTM PROCESSING

In the bidirectional LSTM sequences of DeepSpeech, as shown in Fig. 3(a), total *n* cells are repeatedly processed in both directions, consuming a huge amount of energy. Even though a single LSTM cell can be simplified with any technique [6], in general, all the LSTM sequences in the improved architecture are still based on the identical cells. Hence, in this work, the previous architecture having the same simplified LSTM cell is referred to as the single-mode LSTM architecture.

To further reduce the energy consumption of bidirectional LSTM, on the other hand, we newly introduce the multi-mode LSTM processing. We first define several LSTM cells having the different processing accuracies on their operations, i.e., the multiple modes. It is natural that we consume more processing energy for the stronger cells associated with the more precise computations. Therefore, we may place weaker LSTM cells to non-critical positions for removing the excessive computations. This energy-accuracy tradeoffs can be achieved by adjusting various factors including quantization level, filter pruning, data compression, and even approximate computation.

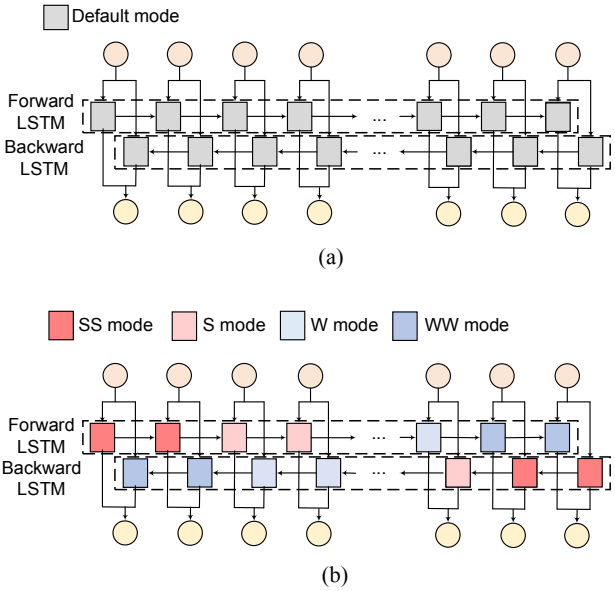

Fig. 3. The conceptual diagram of (a) the conventional single-mode LSTM processing and (b) the proposed multi-mode operation.

After designing the different LSTM cells, we then arrange the new energy-efficient processing scenario based on the important observation that that weak LSTM cells suffers from the erroneous results, and the errors tend to be accumulated at the subsequent cells. Therefore, it is necessary to process the earlier LSTM cells with stronger modes. On the other hand, the later LSTM steps can tolerate more aggressive approximations, allowing the usage of weak modes. Starting from the first LSTM cell of each direction, therefore, the proposed energy-efficient DeepSpeech network gradually changes its processing mode from the strongest to the weakest, minimizing the overall energy consumption without degrading the recognition accuracy.

III. Case Study: Dual-Mode Quantization

To show the effects of the multi-mode LSTM processing, as a case study, we divide the *n*-length bidirectional LSTM network into two regions, i.e., the accuracy-ware and the energy-aware regions. Then two LSTM cells, denoted as strong and weak cells, are defined to have the different quantization levels by using 19 and 7 bits, respectively. According to the numerous simulations, the ratio of two regions is carefully determined to 1:9, which means that the accuracy-aware region only covers the first 10% of LSTM steps to minimize the energy consumption under the 1% accuracy drop as summarized in Table I. Note that the straightforward quantization based on the single-mode 19-bit operations results the similar level of recognition accuracy, but it obviously has the limitation on the amount of energy reduction. For the quantitative comparison, we also design the multi-mode LSTM operator in a 65nm CMOS process. By allocating the dual-mode operations dynamically, as shown in Fig. 4, the proposed scheme reduces the energy consumption by 75% and 52% compared to the baseline network using 32-bit floating-point numbers and the straightforward single-mode architecture with the 19-bit quantized values, respectively.

TABLE I. Performance of LSTM Networks

LSTM architecture	Word error rate (%)
Baseline (32-bit floating-point)	8.4733
Single-mode (19-bit fixed-point)	9.3589
Dual-mode (19-bit / 7-bit fixed-point)	9.4409

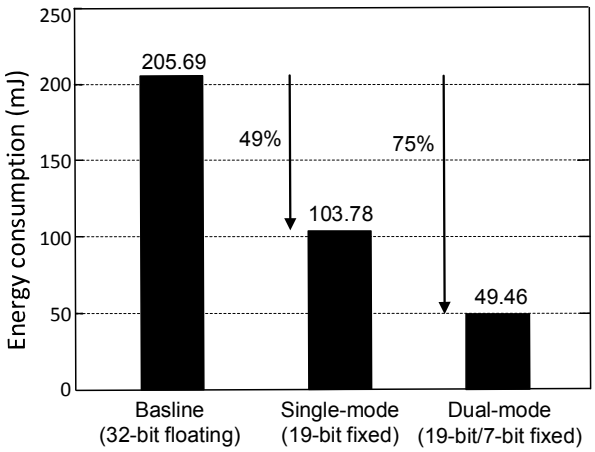

Fig. 4. The energy reduction from the proposed dual-mode operation.

IV. Conclusion

In this paper, we have introduced the multi-mode LSTM network for the energy-efficient speech recognition. Allocating the different processing modes properly, the proposed scheme provides more attractive operating point reducing the processing energy of speech recognition. Compared to the straightforward single-mode approach, the case study based on the dual-mode quantization shows that our architecture saves more than 50% of recognition energy while maintaining the recognition accuracy.

References

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," arXiv prepring arXiv:1402.11281, 2014.

[3] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 4580–4584.

[4] A. Hannun *et al.*, "Deepspeech: Scaling up end-to-end speech recognition," arXiv preprint arXiv:1412.5567, 2014.

[5] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf, Acoust., Speech, Signal Process. (ICASSP),* 2013, pp. 6645-6649.

[6] S. Han *et al.*,"ESE: Efficient speech recognition engine with sparse LSTM on FPGA," in *Proc. ACM/SIGDA Int. Symp. Field-Programmable Gate Arrays*, 2017, pp. 75–84.