Our approach

Investigating 2nd order optimization techniques

- Faster and more stable training for deep neural networks (DNNs)
- Applied to recurrent neural network language model (RNNLM)
- Efficient GPU based training parallelization

- Convergence: 19 epochs using 5604s for SGD; 9 epochs using 4709s for L-BFGS
- 0.7% abs. WER reductions obtained by L-BFGS before interpolation with 4-gram

OBSERVED 2ND ORDER IMPROVEMENTS:

- Results on Switchboard
- Convergence: 16 epochs using 1453s for SGD; 7 epochs using 675s for L-BFGS
- 0.8% abs. WER reductions obtained by L-BFGS before interpolation with 4-gram
- Observed on both tasks, the combination between SGD and L-BFGS is complementary since consistent improvements are obtained

Results on Babel Cantonese

- Convergence: 16 epochs using 1453s for SGD; 7 epochs using 675s for L-BFGS
- 0.8% abs. WER reductions obtained by L-BFGS before interpolation with 4-gram
- Observed on both tasks, the combination between SGD and L-BFGS is complementary since consistent improvements are obtained

Convergence: 16 epochs using 1453s for SGD; 7 epochs using 675s for L-BFGS
- 0.8% abs. WER reductions obtained by L-BFGS before interpolation with 4-gram
- Observed on both tasks, the combination between SGD and L-BFGS is complementary since consistent improvements are obtained

Conclusion

- L-BFGS optimization for RNNLM training & Future work
- Successfully applied to RNNLM training
- Consistent improvements over SGD on multiple speech recognition tasks
- Future research on L-BFGS training of advanced forms of NNS

This research was supported by MSRA grant no. 6904412 and Chinese University of Hong Kong (CUHK) grant no. 4055065