Automatic Speech Recognition Using Deep Learning

Kenichi Iso, Takehiro Sekine
Yahoo JAPAN corporation, JPN
tasekine@yahoo-corp.jp

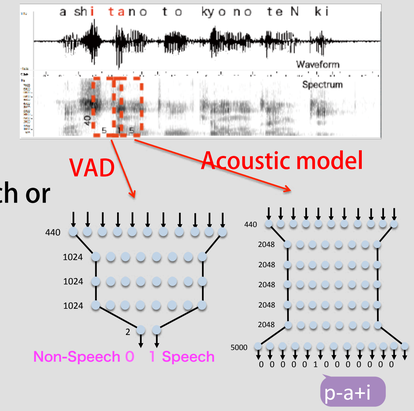
Abstract

Deep Neural Networks (DNNs) have become a popular foundation for state-of-the-art ASR systems. We have collected and transcribed more than 2000 hours of speech data and used it to train DNNs for acoustic models and VAD (voice activity detector) models. Implementation techniques for efficient speech DNN training on GPUs are explained and several evaluation results and the training times are shown using different amounts of training data and sizes of DNN. The trained DNNs are deployed in our ASR services in Japan.

Using DNNs in ASR

Input Sound → VAD → Acoustic model → Language Model → Recognized Text

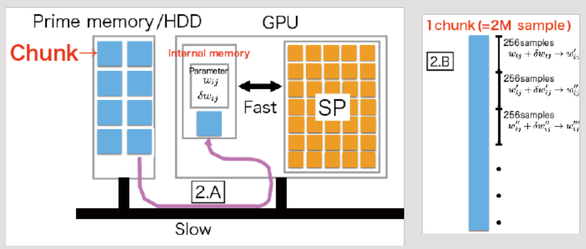
- Voice Activity Detection
  * DNN model has 3 hidden layers with 1024 units each.
  * Output layer has 2 units representing speech and non-speech classes.
  * Classify each 10 msec audio chunk into speech or non-speech.
- Acoustic Model
  * DNN models have 4-7 hidden layers with 1024-4096 units each.
  * Output layer has 5000 units representing tied triphone HMM states.



GPU Training Strategy

- Theano (Python Library)
  * Easy to implement backprop by Matrix notation and differential operators
  * Easy to handle CPU/GPU memory
  * Automatic compilation into C/C++ code at execution

- Training Data
  * 1sample=10msec audio chunk =440dimension=1.7kB(float32)
  * Ex.1000hours=360M sample=590GB



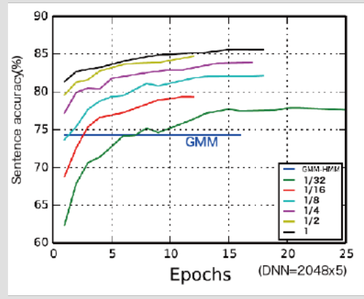
- Training Program(mini-batch Stochastic Gradient Descent)
  1. Shuffle all samples randomly and split them into chunks(ex.2M samples=3.4GB) within GPU memory size.
  2. Train chunk by chunk
     A) Copy a chunk to GPU memory
     B) Execute SGD on every 256 samples (execute forward and backward procedure on 256x440 dimensional matrix, and update DNN parameters)
     C) Repeat for all chunks
  3. After training all chunks(=1 epoch), evaluate accuracies for validation data, and decide whether to repeat training.

Main Result

- Accuracy
  * Accuracies with DNN are higher than those of GMM even using only 1/32 of training data used to train GMM.
  * Improved error rates with DNN: reduced by 30-40% compared to GMM.
  * No effectiveness observed using pretraining with RBM. [Fig.A]

- Training time [Fig.B]
  * Training time for acoustic model(2048x5) is about 1month for 15 epochs with 1000 hours speech data and 1GPU(M2090) machine (2weeks with Titan X).

Comparison of different amounts of training data



Dependency of DNN structures

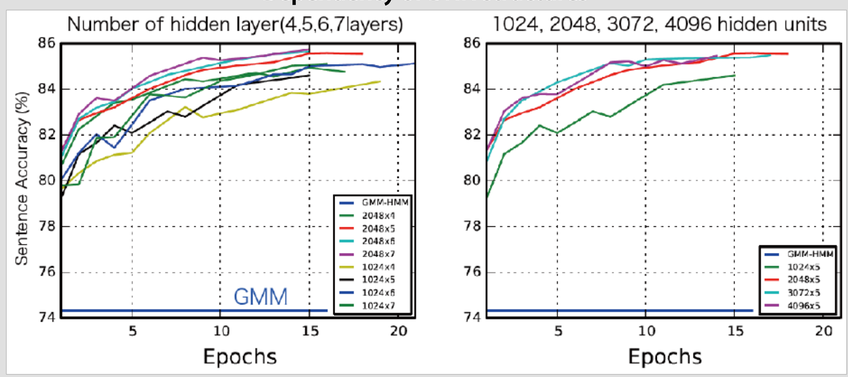
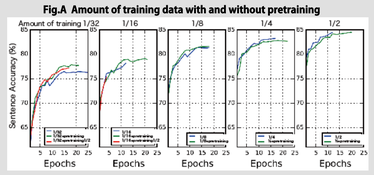
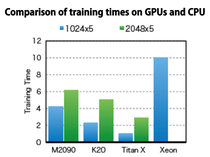


Fig.A Amount of training data with and without pretraining



Comparison of training times on GPUs and CPU



GPU training time per epoch

Fig.B

Conclusion and future work

- We have succeeded in deploying acoustic model and VAD DNNs in our ASR system.
  -- DNN trained using 2000 hours of speech data reduced the error rate by 30-40% compared to GMM.
  -- No effectiveness was observed using RBM pretraining
  -- GPU can drastically speed up DNN training compared to CPU
  -- Time for training increases..
     * approximately in proportion to the number of DNN hidden layers
     * with the number of hidden units in DNN hidden layer to the power of 1.5
- We are working to realize an end-to-end ASR system using the Deep Learning Methodology and parallel computing framework.