The Use of Deep Neural Network for Speech Recognition

Chao Liu

2013/05/13

Review: Speech Recognition

Deep Neural Network(DNN)

Using DNN for ASR

Experiments

Review: Speech Recognition

Overview

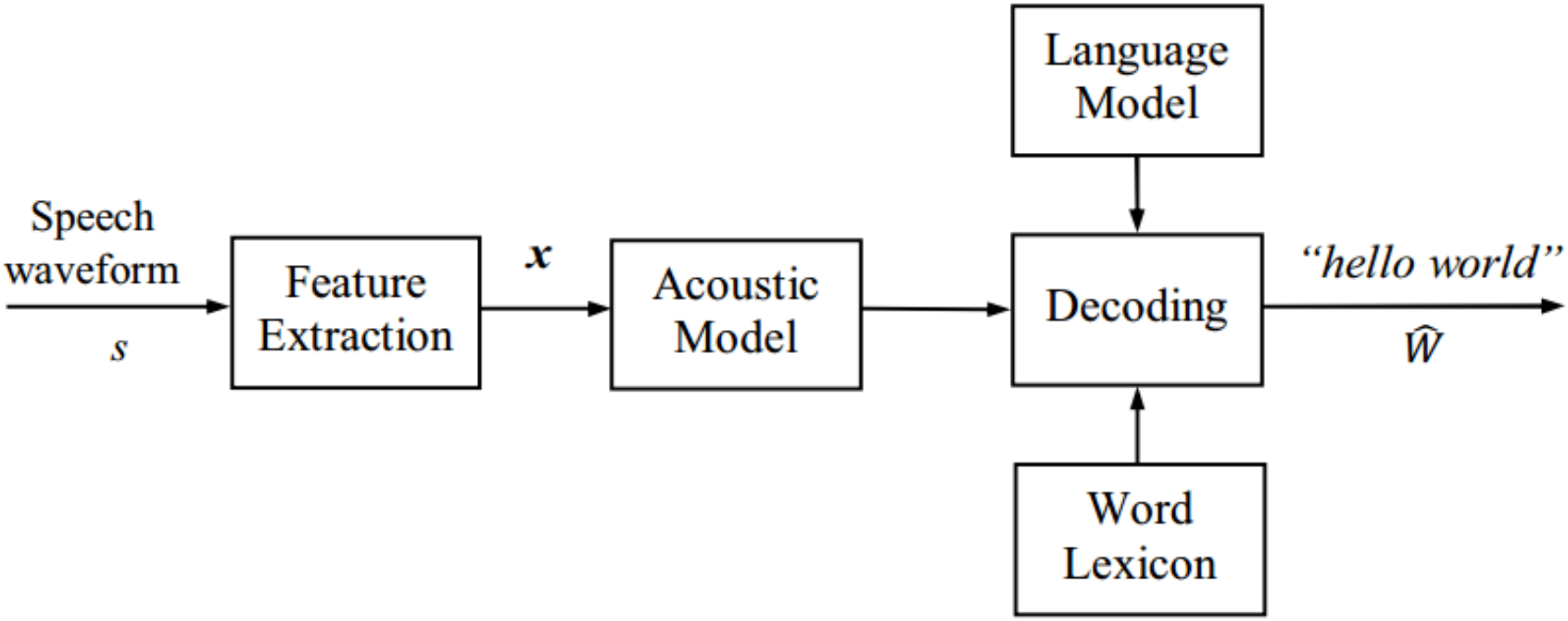


Figure : A typical speech recognition system

- Get Best hypothesized word sequence

$$\hat{W} = \underset{W}{argmax} P(X|W)P(W)$$

HMM-GMM framework

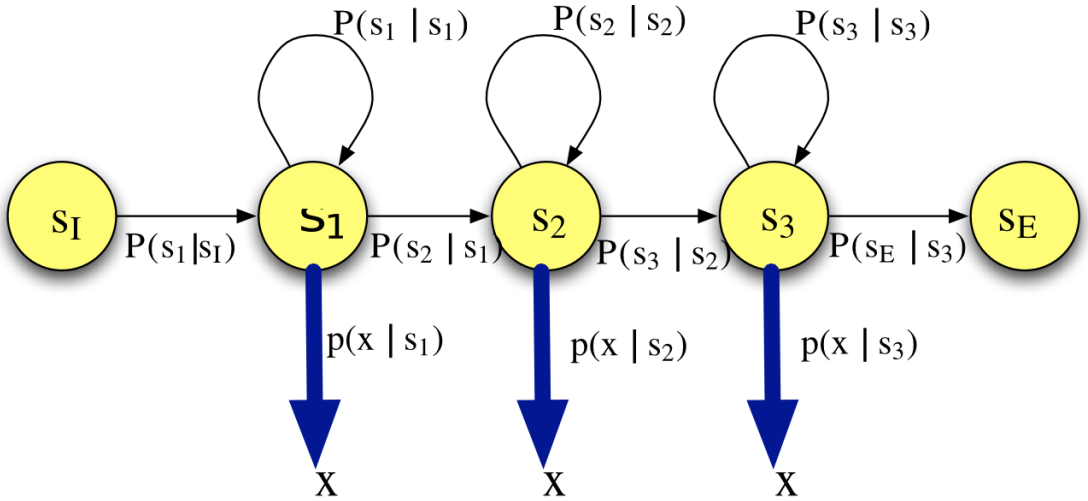


Figure : Hidden Markov Model(HMM)

- Gaussian mixture model(GMM)

$$P(x|s) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i)$$

- Maximum-likelihood estimation(MLE) using EM

Deep Neural Network(DNN)

Overview

- Deep architecture
  - Multi Layer Perceptron(MLP) with many hidden layers
  - Highly non-linear and varying functions
- Hard to train
  - Back Propagation often trapped in poor local minima
  - Pretraining by stacking Restricted Boltzmann Machine(RBM)
    - Unsupervised learning

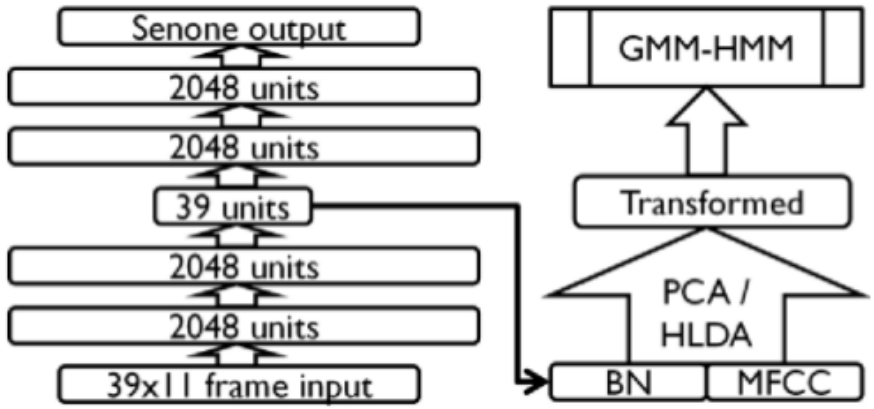
Generative and discriminative models

- The goal is classification
  - more appropriate to use discriminative models
- Cons for discriminative models
  - Can not describe properties
  - Difficult to adapt

Using DNN for ASR

DNN for feature extraction

- Bottleneck feature



  - learning high level discriminative feature for HMM-GMM system
  - concatenated with conventional short time features helps

DNN for acoustic modeling

- HMM-DNN hybrid system
  - Replace GMM with DNN
  - Get posterior probability from network directly
    - Calculate conditional probaility using Bayes rule

$$P(x|s) = \frac{P(s|x)P(x)}{P(s)}$$

Experiments

Data and setup

- Data
  - Training: tencent speech data, about 400 hours
  - Testing: tencent speech data, about 19 hours
- Language model
  - 3-gram LM trained with ?G text
  - 1e-5 / 1e-9 pruned biglm WFST decoding
- Computer
  - 5 x 4 cpu cores used for conventional training and testing, 1 gpu used for DNN training
- DNN
  - both approaches share a 300-1200-1200-1200-40-1200-3858 BN network

Result

WER%	fMMI HMM-GMM	BN HMM-GMM	HMM-DNN
1900	7.35	6.57	7.27
2044	24.03	21.77	20.24
online1	34.33	31.44	30.53
online2	26.80	24.10	23.89
map	27.69	23.79	22.46
notepad	21.75	15.81	12.74
general	38.90	33.61	31.55
speedup	26.81	22.82	22.00

Conclusion

- DNN is useful for speech recognition
- HMM-DNN outperforms BN HMM-GMM
- Open questions
  - Choices of feature input
  - Choices of network structure
  - Parallel mechanism for training (DNN training takes much more time)
  - Sparse network structure
  - ...

Thank you

- **Q&A**