NOTE for part of ICASSP 16

LSTM

mlsp-l2.2: simplifying long short-term memory acoustic models for fast training and decoding

1. deriving input gates from forget gates, as they show a negative correlation
2. removing recurrent inputs from output gates
3. frame skipping

sp-l1.1: exploring multidimensional lstms for large vocabulary asr
:: LSTM can scan the frames along the time or/and frequency axis

sp-p4.9: exploiting lstm structure in deep neural networks for speech recognition
:: the expansion of LSTM along time axis is introduced to DNN while along the layer

sp-p11.7: highway long short-term memory rnns for distant speech recognition
:: a gate connects cells of layers directly with dropout

sp-p14.6: recurrent support vector machines for speech recognition
:: replacing the softmax layer in RNN with Support Vector Machines, frame-level max-margin

hlt-l1.2: learning compact recurrent neural networks
:: low-rank factorizations and share low-rank across layers

hlt-l1.4: on the compression of recurrent neural networks with an application to lvcsr acoustic modeling for embedded speech recognition
:: factorizing recurrent and inter-layer matrices, sharing a recurrent projection matrix

----

CTC

mlsp-p6.5: an empirical exploration of ctc acoustic models

1. initialize the bias vector of the LSTMs forget gates to larger values, avoiding decay gradients
2. more data, convolution layer in front, bi-directional lstm, front-end(VTLNs, etc.)

➔ better performance

sp-p4.1: flat start training of cd-ctc-smbr lstm rnn acoustic models

1. BLSTM with flat start CTC training to align phonemes and get CD phones
2. Another LSTM trained with the alignments
3. sMBR

Attention/End-to-end

sp-l1.2: end-to-end attention-based large vocabulary speech recognition
:: encoder-decoder with attention

sp-l1.5: listen, attend and spell: a neural network for large vocabulary conversational speech recognition
:: encoder-decoder with attention, encoder is a pyramidal BLSTM

sp-l5.1: on training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition

1. learning rate schedule, tricky
2. decoder with long memory by introducing another recurrent layer for implicit language modelling

----

CNN

sp-l1.4: very deep multilingual convolutional neural networks for lvcsr

1. very deep cnn for lvcsr and multilingual
2. multi-scale feature inputs, e.g., 3 input channels with different contexts like RGB

sp-p11.2: noise robust speech recognition using recent developments in neural networks for computer vision

1. 3 channel inputs(with delta plus double-delta), different CNN architecture
2. A convolution to simulate that dynamic feature
3. Parametric Rectifier

sp-p14.9: filterbank learning using convolutional restricted boltzmann machine for speech recognition

:: Convolutional Restricted Boltzmann Machine

Sparsity

mlsp-p7.1: ranking the parameters of deep neural networks using the fisher information

1. non-parametric Fisher Information to rank the parameters
2. removing redundant unimportant parameters and quantizing the remaining

----

Optimization

mlsp-p7.3: batch normalized recurrent neural networks

:: batch normalization only to the input-to-hidden transition, makes convergence faster, not improves the generalization performance, similar to the way dropout applied to RNN

mlsp-p7.7: learning deep neural network using max-margin minimum classification error

1. measure the misclassification error
2. ReLU-like function as the loss function of above
3. combine cross-entropy and that method

:: Max-margin means the finite range of the loss

sp-p2.4: character-level incremental speech recognition with recurrent neural networks

:: tree-based online beam search for CTC end-to-end

sp-p4.10: self-stabilized deep neural network

:: each weight matrix multiplied by a trainable parameter as stabilizer

hlt-l1.3: towards implicit complexity control using variable-depth deep neural networks for automatic speech recognition

:: for an already trained DNN, criterion to choose output of which hidden layer to be the final output

----

Noise labels

mlsp-p7.8: training deep neural-networks based on unreliable labels

1. an additional noise layer(confusion matrix) converting the unreliable labels to right ones(latent variables)
2. EM to optimize

Adaptation

sp-l3.1: combining i-vector representation and structured neural networks for rapid adaptation
:: i-vectors are used to predict multi-basis transform weight

sp-l3.2: low-rank plus diagonal adaptation for deep neural networks
:: a layer is decomposed to a low-rank matrix and a diagonal one with cross-layer link

sp-l3.5: investigations on speaker adaptation of lstm rnn models for speech recognition

1. KL regularization is important for adaptation
2. adapting top hidden layers is clearly more effective for LSTM-RNN
3. adapting cell internal matrix is effective, adapting the projection weight matrix and hidden activations to cell internal memory are the most effective

sp-p1.8: speaker adaptation of rnn-blstm for speech recognition based on speaker code
:: speaker code (d-vector similarly) to cell or gates
This paper is similar to upstairs.

sp-l3.6: joint acoustic factor learning for robust deep neural network based automatic speech recognition
sp-l5.2: discriminatively trained joint speaker and environment representations for adaptation of deep neural network acoustic models
:: enhance the input feature with a bottleneck feature learned separately by a DNN with speaker or/and phone or/and noise as targets(multi-task or combine them)

sp-p1.1: context adaptive deep neural networks for fast acoustic model adaptation in noisy conditions
:: a layer with separate parallel weight matrices to model different context, context class weights are computed by a small nnet, trained jointly, input of the small nnet here is i-vector

sp-p1.3: speaker-aware training of lstm-rnns for acoustic modelling
:: bottleneck speaker vector as auxiliary feature, extracted separately by a nnet with speaker-id or/and monophone as targets

sp-p1.4: non-negative intermediate-layer dnn adaptation for a 10-kb speaker adaptation profile
:: inserting a compact linear layer on top of SVD layer, with non-negative constraints such as, a positive threshold, setting small-positive weights in the non-negative model to zero

sp-p1.4: non-negative intermediate-layer dnn adaptation for a 10-kb speaker adaptation profile
:: bottleneck features extracted by a nnet to enhance the input frames

sp-p1.6: speaker cluster-based speaker adaptive training for deep neural network acoustic modeling

1. train a common base dnn
2. cluster the training set based on i-vector distance, adapt the base nnet respectively
3. when decoding, find the specific model with the nearest cluster based on i-vector

sp-p1.7: dnn speaker adaptation using parameterised sigmoid and relu hidden activation functions

:: parameterised sigmoid or rectifier for adaptation

sp-p1.9: efficient non-linear feature adaptation using maxout networks

:: when adapting, convert the final layer from linear to maxout, and a new linear layer above

sp-p1.10: sequence summarizing neural network for speaker adaptation

:: a network(summary history) to produce extra input feature is trained together with the main nnet

sp-p14.1: comparison of unsupervised sequence adaptations for deep neural networks

:: three unsupervised sequence adaptation techniques: maximum a posteriori (MAP), entropy minimization, and Bayes risk minimization,

----

SVD / low-dimensional structrue

sp-l5.6: linearly augmented deep neural network

:: when doing SVD, a linear weight is imported to transform the input to output, making training very deep networks without pre-training

sp-p10.4: exploiting low-dimensional structures to enhance dnn based acoustic modeling in speech recognition

1. train a common dnn
2. dictionary learning for dnn posteriors
3. reconstruct the dnn posteriors based on the dictionary and sparse representation

:: improvement based on the fact that the true information is embedded in a low-dimensional subspace, sparse reconstruction separates out the high dimensional erroneous estimates, **interesting**!

Low-resource/multilingual

sp-p2.1: supervised and unsupervised active learning for automatic speech recognition of low-resource languages

:: supervised: select a larger set of annotated data for training based on a first-pass dnn
unsupervised: select data based on diversity reward

sp-p4.2: multilingual data selection for training stacked bottleneck features

:: select more similar data to the target low-resource language from rich-resource data(by Language Identification system) to train a system, then adapt it to the target language
DNN is just used for bottleneck feature extraction, still gmm-hmm for acoustic

sp-p4.4: a study of rank-constrained multilingual dnns for low-resource asr

1. train the lower several layer with all low-resource languages
2. adapt it to a target one

sp-p4.6: multilingual region-dependent transforms
:: introduce Region Dependent Transform to Stacked bottleneck (SBN) feature scheme as paper "sp-p4.2: multilingual data selection for training stacked bottleneck features"

Joint training / multi-task
sp-p4.3: prediction-adaptation-correction recurrent neural networks for low-resource language speech recognition
1. an additional nnet with state or phone as targets (therefore call it prediction DNN) jointly trained with the main nnet with recurrences with each other
2. introduce Bottleneck features from Stacked bottleneck (SBN) features as paper "sp-p4.2: multilingual data selection for training stacked bottleneck features"

sp-p4.5: sequence training of multi-task acoustic models using meta-state labels
:: Combine CD states inventories (multiple outputs) to "meta-states", design a decoder subsequently

sp-p11.10: integrated adaptation with multi-factor joint-learning for far-field speech recognition
:: acoustic model and far-field model trained together, not interact enough
----
Multistream
sp-p10.2: novel neural network based fusion for multistream asr
:: use only one fusion DNN with dropout for the input

DAE
sp-p11.9: two-stage noise aware training using asymmetric deep denoising autoencoder
1. input of DAE is noised feature, outputs are two sets of labels (clean feature and noise)
2. the clean output is used for ASR

Language identification
sp-p13.1: a hierarchical framework for language identification
:: Tree Structure to cluster languages, using cosine similarity score

sp-p13.2: local fisher discriminant analysis for spoken language identification
:: local Fisher discriminant to extract the discriminative features from i-vectors

sp-p13.3: language recognition using deep neural networks with very limited training data
:: dnn trained with labeled data to estimate unlabeled data, all data together trains another one
----
Biological
sp-p14.2: synaptic depression in deep neural networks for speech processing
:: synaptic depression(actually weight decay along time) into DNN

Other topics
Music, emotion, speaker
Noise, Echo, Feedback, Reverberation, far-field
Matrix Factorization
topic models
beamforming, multichannel