UNIVERSITY OF CAMBRIDGE

Improved TDNNs using Deep Kernels and Frequency Dependent Grid-RNNs

Florian Kreyssig, Chao Zhang and Phil Woodland

April 18, 2017

Cambridge University Engineering Department

Overview

- Introduction
- Models
  - Baseline TDNN
  - Deep Kernels
  - Frequency-Dependent Grid-RNN
  - Frequency-Dependent CNN (for comparison)
- Experimental Setup (MGB3 English)
- Experimental Results
- Conclusions

Introduction

Neural Network Depth

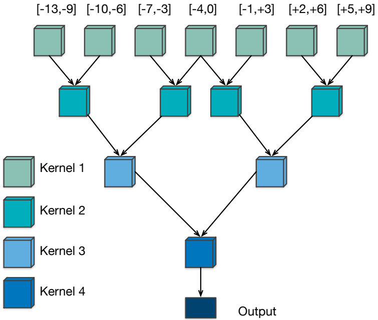
- deepening Neural Networks often yields improved performance
- structure of the TDNN restricts its depth
- we deepen the TDNN by exchanging each Kernel of the temporal convolution through a deeper structure

Frequency Dependent Grid-RNNs

- recently 2D-LSTM designs were shown to improve acoustic modelling
- we propose an efficient 2D-RNN design with frequency dependent parameters that as the front-end to a TDNN

Time-Delay Neural Networks (TDNNs) [1]

- consists of FC layers repeated at different time-steps
- parameters are shared across time
- incorporates that the same feature can occur at any time-step
- similar to 1-D (temporal) CNNs
- modern versions use shifts of more than one frame
- version from [2] is used in this work

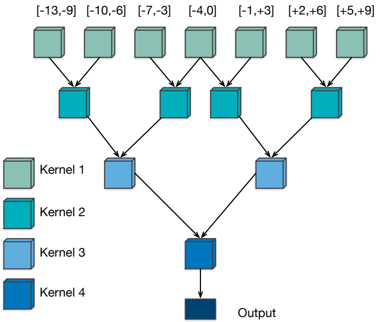


[1] A. Waibel et. al. “Phoneme recognition using time-delay neural networks,” 1989

[2] V. Peddinti et. al.“A time delay neural network architecture for efficient modeling of long temporal contexts,” 2015

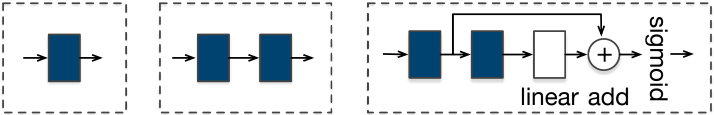
Time-Delay Neural Networks (TDNNs)

- current TDNNs are rel. shallow since deep TDNNs need larger input contexts
- TDNN design does not incorporate the structure of the frequency domain
- this work deepens the TDNN, by deepening each convolutional kernel
- spectro-temporal variations will be modelled using a 2D-RNN as a front-end to the TDNN
- both alterations can be combined



Deep Kernels

- replace each convolution kernel in a TDNN with a Deep Kernel
- parameters are still shared across time-domain
- Double Kernel consists of two FC layers
- Resnet Kernel consists of FC layer followed by two further FC layers bypassed with a residual connection
- *linear* activation function is needed since output range of $\sigma(\cdot)$ is positive



Standard Kernel   Double Kernel   ResNet Kernel

Figure: Darker blocks are FC layers with $\sigma(\cdot)$ activation function. The white block denotes an FC layer with linear activation function.

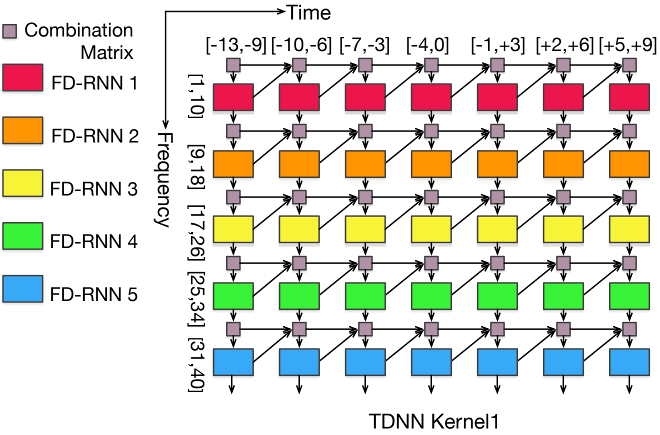
Frequency-Dependent Grid-RNNs

- 2D-LSTM architectures have shown promising results [3,4]
- LSTMs are unfolded along both the time- and frequency axis
- allows units to influence each other within the same layer
- unfolding for one time-step at a time is expensive
- we exploit TDNN structure and unfold for 7 time-steps (time-bins)
- features at low and high end of the frequency scale are different
- translational weight sharing along frequency axis is sub-optimal

[3] J. Li, et. al., “Exploring multidimensional LSTMs for large vocabulary ASR,” 2016

[4] T. Sainath and B. Li, “Modeling time-frequency patterns with LSTM vs. convolutional architectures for LVCSR tasks,” 2016

Frequency-Dependent Grid-RNNs

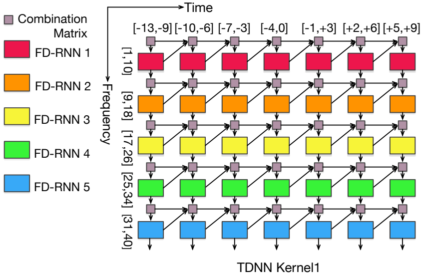


Frequency-Dependent Grid-RNNs

$$\mathbf{h}^{l}_{t,k} = W^{l}_{F_1}\mathbf{h}^{F}_{t,k} + W^{l}_{F_2}\mathbf{h}^{F}_{t,k-1} + V^{l}_{l}\mathbf{h}^{l}_{t-1,k} + \mathbf{b}^{l}$$ Combination Matrix

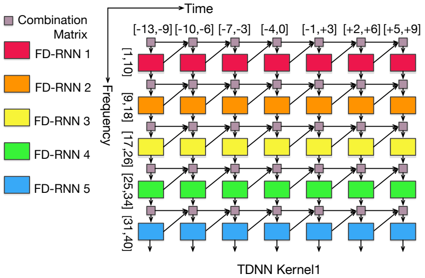
$$\mathbf{h}^{F}_{t,k} = \sigma\left(W^{F}_{(k)}\mathbf{x}_{t,k} + V^{F}_{(k)}\mathbf{h}^{l}_{t,k-1} + \mathbf{b}^{F}_{(k)}\right)$$ FD-RNN k

- $\mathbf{x}_{t,k}$ is the input at time step k and frequency step k
- linear activation in Combination Matrix for better information flow



Frequency-Dependent Grid-RNNs

- architecture separates information flow and feature extraction
- one or both axes can be reversed to yield bi-directional FD-RNN
- 5 frequency bins and 7 time bins for easy combination with TDNN
- frequency bins have separate weights (note colours)
- 'FD-RNN 5' (blue) is followed by the TDNN



Frequency-Dependent CNN (for comparison)

- the 7 time bins of the TDNN have width 5
- split frequency axis into 7 overlapping frequency bins
- each time-frequency bin is convolved with a set of 5x5 filters
- separate set of filters for each frequency bin
- output is 6x1 for each filter within a time-frequency bin
- reduced to 3x1 via maxpooling
- output of the convolutions within a time bin are passed to the TDNN

Experimental Setup

Data

- 55h and 275h from English Multi-Genre Broadcast (MGB) Challenge 3
- A trigram word level LM with a 63k word dictionary
- **dev17b** test set contains 5.5h data with reference segmentation

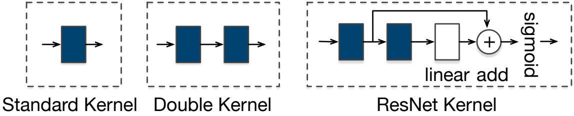
Systems

- All experiments were conducted by extending HTK 3.5
- 40-dim log-Mel filter bank features were used, with Δ for LSTMP
- number of parameters was kept constant by adjusting layer-sizes
- trained using cross-entropy criterion
- initialized using discriminative pre-training
- evaluation used confusion network decoding

Results 55h: Comparing the three Kernels

- Deep Kernels yield significant improvement

ID	System	WER	WERR
ST_1^{55h}	TDNN	32.7	–
DT_1^{55h}	Double-TDNN	31.5	3.7%
RT_1^{55h}	ResNet-TDNN	30.5	6.7%



Results 55h: Comparing with appending FC-layers

- appending FC-Layers also yields improvement
- can be combined with ResNet-Kernel
- gains from ResNet-Kernels

ID	System	WER	WERR
ST_1^{55h}	TDNN	32.7	–
RT_1^{55h}	ResNet-TDNN	30.5	6.7%
ST_2^{55h}	TDNN + 1 FC	31.9	2.4%
ST_3^{55h}	TDNN + 2 FC	30.9	5.5%
ST_4^{55h}	TDNN + 3 FC	30.5	6.7%
RT_2^{55h}	ResNet-TDNN + 3 FC	29.8	8.9%

Experimental Results 55h: Combination with Grid-RNN

- frequency-Dependent parameters are important
- bi-directional model further improves results
- bi-directional FD-Grid-RNN outperforms frequency dependent CNN

ID	System	WER	WERR
ST_1^{55h}	TDNN	32.7	–
RT_1^{55h}	ResNet-TDNN	30.5	6.7%
RC_1^{55h}	FD-CNN-ResNet-TDNN	29.9	8.6%
RG_1^{55h}	Grid-RNN-ResNet-TDNN	30.1	8.0%
RG_2^{55h}	FD-Grid-RNN-ResNet-TDNN	29.6	9.5%
RG_3^{55h}	BD-FD-Grid-RNN-ResNet-TDNN	29.0	11.3%
L_1^{55h}	2L-LSTMP	30.6	6.4%

Experimental Results 275h

- alterations also give large improvements for the larger dataset
- ResNet-Kernel is more effective than appending FC layers on 275h dataset in comparison to 55h dataset

ID	System	WER	WERR
ST_1^{275h}	TDNN	26.7	–
ST_4^{275h}	TDNN + 3 FC	25.7	3.7%
RT_1^{275h}	ResNet-TDNN	25.0	6.4%
RT_2^{275h}	ResNet-TDNN + 3 FC	24.7	7.5%
RG_3^{275h}	BD-FD-Grid-RNN-ResNet-TDNN	24.3	9.0%
L_1^{275h}	2L-LSTMP	25.6	4.1%

Conclusions

- replacing convolutional kernels in a TDNN with deeper structures improves acoustic modelling (6.4% WERR)
- the best deep kernel consists of three FC layers with a ResNet connection from the output of the first the output of the third
- 2D-RNNs can be used as front-end to TDNN to effectively model spectro-temporal variations
- 2D-RNN design need not rely on LSTMs
- parameters of the 2D-RNNs should be frequency dependent
- the alterations are complimentary (9.0% WERR)

Thanks for listening!