Knowledge Distillation for Small-footprint Highway Networks

Liang Lu[†], Michelle Guo[‡] and Steve Renals[*]

[†]Toyota Technological Institute at Chicago
[‡]Stanford University
[*]The University of Edinburgh

6 March 2017

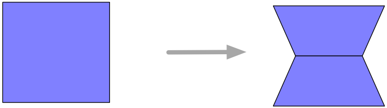
Why smaller model?

- Deep learning has made a tremendous impact
  - Large amount of data for training
  - Powerful computational devices
  - Connected to a server
- Embedded (*Client-side*) deep learning
  - Local inference
  - Small footprint
  - Energy efficient

Why smaller model for speech recognition?

- Speech recognition as an interface (requiring internet connection)
  - Google Home
  - Amazon Alexa
  - Microsoft Cortana
  - Apple Siri
  - ...
- Local speech service
  - Internet is unavailable
  - Privacy issues
  - Low latency

Background: smaller models

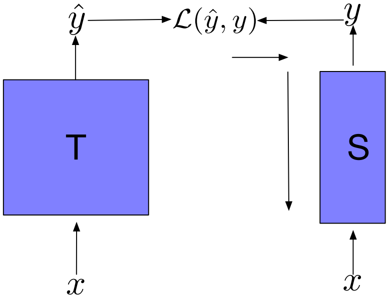


- Low-ranks matrices for DNNs
  - J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition." in Proc. INTERSPEECH, 2013
  - T.N.Sainath, B.Kingsbury, et al., "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in Proc. ICASSP. IEEE, 2013

Background: smaller models

- Structured linear layers
  - V. Sindhwani, T. N. Sainath, and S. Kumar, "Structured transforms for small-footprint deep learning", in Proc. NIPS, 2015.
  - M. Moczulski, M. Denil, J. Appleyard, and N. de Freitas, "ACDC: A Structured Efficient Linear Layer," ICLR 2016

Background: smaller models



- FitNet by teacher-student training
  - J.Li, R.Zhao, J.-T.Huang, and Y.Gong, "Learning small-size DNN with output-distribution-based criteria," in Proc. INTERSPEECH, 2014
  - R. Adriana, B. Nicolas, K. Samira Ebrahimi, C. Antoine, G. Carlo, and B. Yoshua, "FitNets: Hints for thin deep nets," in Proc. ICLR, 2015

Model

$$\boldsymbol{h}_l = \sigma(\boldsymbol{h}_{l-1}, \theta_l) \circ \underbrace{T(\boldsymbol{h}_{l-1}, \boldsymbol{W}_T)}_{\text{transform gate}} + \boldsymbol{h}_{l-1} \circ \underbrace{C(\boldsymbol{h}_{l-1}, \boldsymbol{W}_c)}_{\text{carry gate}} \tag{1}$$

- Shortcut connections with gates
- Similar to Residual networks
- $\boldsymbol{W}_T$ and $\boldsymbol{W}_C$ are layer independent

[1] R.K.Srivastava, K.Greff, and J.Schmidhuber, "Training very deep networks," in Proc. NIPS, 2015
[2]Y. Zhang, et al. "Highway Long Short-Term Memory RNNs for Distant Speech Recognition", in Proc. ICASSP 2015
[3] L. Lu and S. Renals, "Small-footprint deep neural networks with highway connections for speech recognition", in Proc. Interspeech 2016

Loss Functions

- Cross Entropy Loss

$$\mathcal{L}^{(CE)}(\theta) = -\sum_{j} \underbrace{\hat{y}_{jt}}_{\text{label}} \log \underbrace{y_{jt}}_{\text{prediction}}, \tag{2}$$

where j is the class index, and t is the time step.

- Teacher-Student Loss (KL-divergence)

$$\mathcal{L}^{(KL)}(\theta) = -\sum_{j} \underbrace{\tilde{y}_{jt}}_{\text{prediction-T}} \log \underbrace{y_{jt}}_{\text{prediction-S}}, \tag{3}$$

Loss Functions

- Sequence-level teacher-student loss

$$\mathcal{L}^{(sKL)}(\theta) \approx \sum_{\mathcal{W} \in \Phi} P^*(\mathcal{W}|\boldsymbol{X}) \log P(\mathcal{W}|\boldsymbol{X}, \theta) \tag{4}$$

where $P(\mathcal{W}|\boldsymbol{X}, \theta)$ is the posterior given by MMI or sMBR.

- Teacher-student training to sequence training

$$\widehat{\mathcal{L}(\theta)} = \mathcal{L}^{(sMBR)}(\theta) + p\mathcal{L}^{(KL)}(\theta). \tag{5}$$

where p is the regularization weight.

[1] J. Wong and M. Gales, "Sequence Student-Teacher Training of Deep Neural Networks," in Proc. Interspeech, 2016

[2]Y. Kim and A. Rush, "Sequence-Level Knowledge Distillation", in arXiv 2016

Experiments

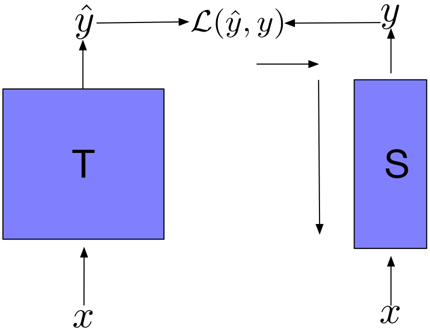
- AMI corpus 80h training data (28 million frames)
- Using the standard Kaldi recipe
  - fMLLR acoustic features
  - 3-gram language models
- CNTK was used to build HDNN models
- The same decision tree was used

Smaller model by highway networks

Model	Size	eval CE	eval sMBR
DNN-$H_{2048}L_6$	$30M$	26.8	**24.6**
DNN-$H_{512}L_{10}$	$4.6M$	28.0	25.6
DNN-$H_{256}L_{10}$	$1.7M$	30.4	27.5
DNN-$H_{128}L_{10}$	$0.71M$	34.1	30.8
HDNN-$H_{512}L_{10}$	$5.1M$	26.5	**24.1**
HDNN-$H_{256}L_{10}$	$1.8M$	27.9	**25.0**
HDNN-$H_{128}L_{10}$	$0.74M$	–	28.7

[1] L. Lu and S. Renals, "Small-footprint Deep Neural Networks with Highway Connections for Speech Recognition," in Proc. Interspeech, 2016

Teacher-Student Training



Teacher-Student Training

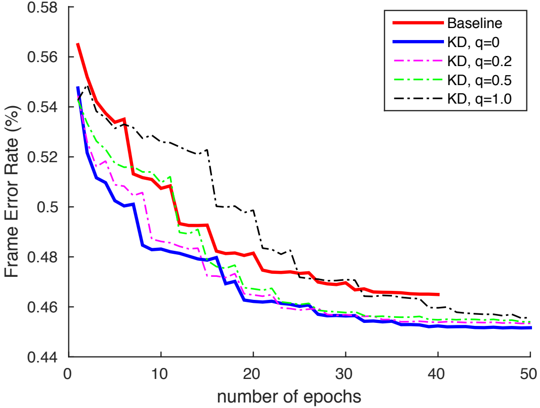
Table: Results of teacher-student training.

Model	q	T	WER eval	WER dev
DNN-$H_{128}L_{10}$	–	–	34.1	31.5
HDNN-$H_{128}L_{10}$ baseline	–	–	32.0	29.9
HDNN-$H_{128}L_{10}$	0	1	31.3	29.3
HDNN-$H_{128}L_{10}$	0.2	1	31.4	29.5
HDNN-$H_{128}L_{10}$	0.5	1	31.3	29.4
HDNN-$H_{128}L_{10}$	0	2	32.3	29.9
HDNN-$H_{128}L_{10}$	0	3	33.0	30.6

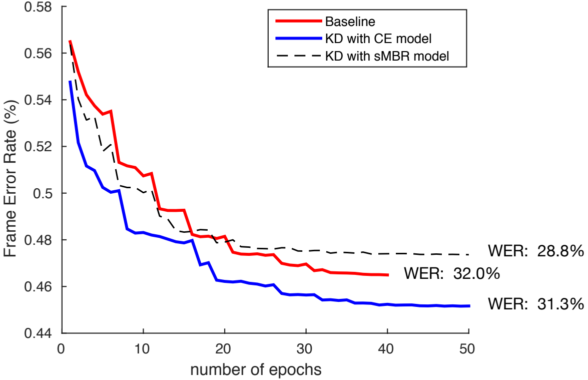
[1] G. Hinton et al.,"Distilling the knowledge in a neural network," in Proc. NIPS workshop, 2015 $T: y_{jt} = \text{softmax}(z_{jt}/T)$; $q: \mathcal{L}(\theta) = \mathcal{L}^{(KL)} + q\mathcal{L}^{(CE)}(\theta)$

[2] K. Markov and T. Matsui, "Robust speech recognition using generalized distillation framework", in Proc. Interspeech 2016

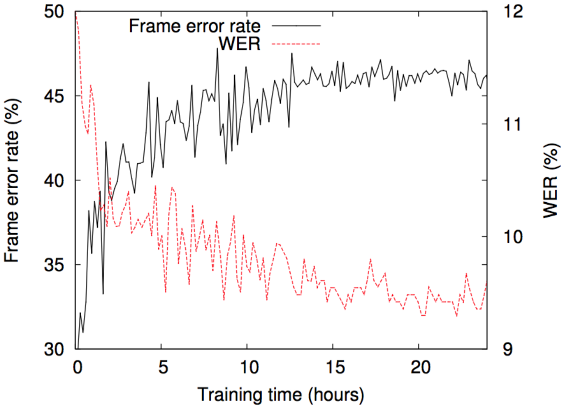
Teacher-Student Training



Teacher-Student Training

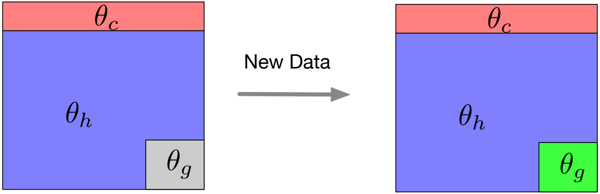


Teacher-Student Training



[1]G. Heigold, et al., "Asynchronous stochastic optimization for sequence training of deep neural networks," in Proc. ICASSP, 2014

Experiments – Adaptation



- θ_c: parameters in the softmax layer
- θ_h: parameters in the hidden layers
- θ_g: parameters in the gate functions

Experiments – Adaptation

Table: Results of unsupervised speaker adaptation.

Model	Seed	Update	WER (`eval`) SI	SD
HDNN-$H_{512}L_{10}$	sMBR	θ_g	24.9	**24.1**
HDNN-$H_{256}L_{10}$			26.0	**25.0**
HDNN-$H_{512}L_{10}$		$\{\theta_h, \theta_g, \theta_c\}$	24.9	24.5
HDNN-$H_{256}L_{10}$			26.0	25.4

[1] L. Lu, "Sequence Training and Adaptation of Highway Deep Neural Networks," in Proc. SLT 2016

Teacher-Student Training

Table: Results of unsupervised speaker adaptation.

Model	Loss	Update	eval SI	eval SD
HDNN-$H_{128}L_{10}$-KL	KL	$\{\theta_h, \theta_g, \theta_c\}$	28.4	27.5
HDNN-$H_{128}L_{10}$-KL	KL	θ_g	28.4	27.8
HDNN-$H_{128}L_{10}$-KL	CE	$\{\theta_h, \theta_g, \theta_c\}$	28.4	27.7
HDNN-$H_{128}L_{10}$-KL	CE	θ_g	28.4	27.1

Where we are?

Model	Size	eval CE	eval sMBR
DNN-$H_{2048}L_6$	$30M$	26.8	**24.6**
DNN-$H_{512}L_{10}$	$4.6M$	28.0	25.6
DNN-$H_{256}L_{10}$	$1.7M$	30.4	27.5
DNN-$H_{128}L_{10}$	$0.71M$	34.1	30.8
HDNN-$H_{512}L_{10}$	$5.1M$	26.5	**24.1**
HDNN-$H_{256}L_{10}$	$1.8M$	27.9	**25.0**
HDNN-$H_{128}L_{10}$	$0.74M$	–	$28.7 \rightarrow 27.1$

Conclusion

Teacher-student training + Highway networks

$\Downarrow$

Compact & Adaptable model

Thank you ! Questions?