# Small-footprint Deep Neural Networks with Highway Connections for Speech Recognition

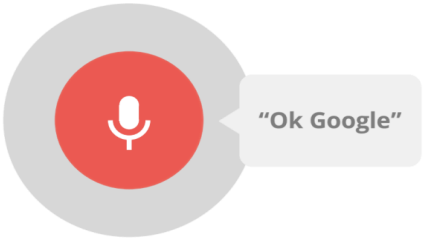
Liang Lu, Steve Renals

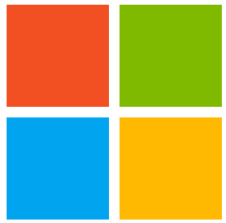
[1]The University of Edinburgh
[2]Toyota Technological Institute at Chicago

9 September 2016

# Background

- Deep learning has made a significant difference



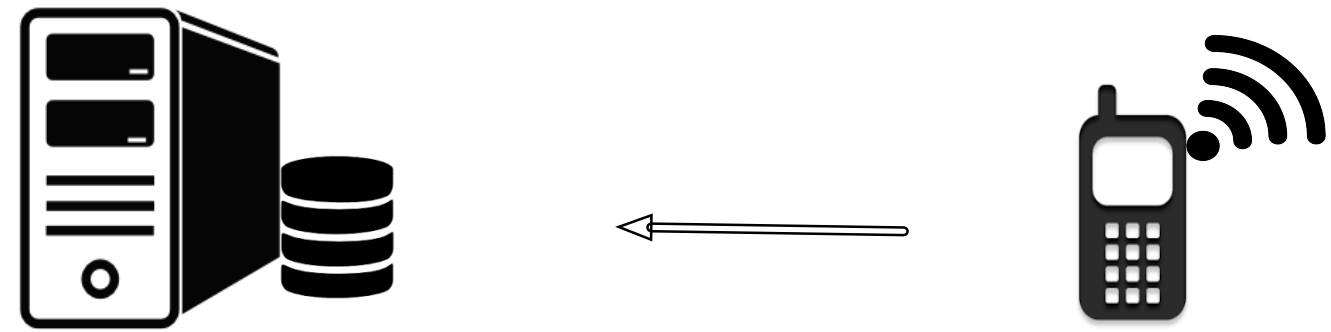






... and many more!

# Background

- The current business model



Big model on the server

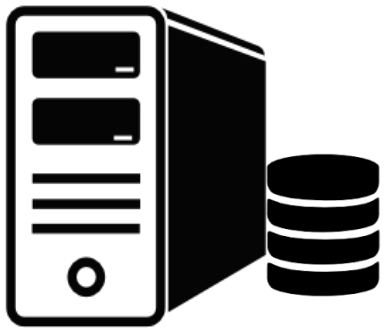
# Background

- However, the problems ...

Big model on the server

# Background

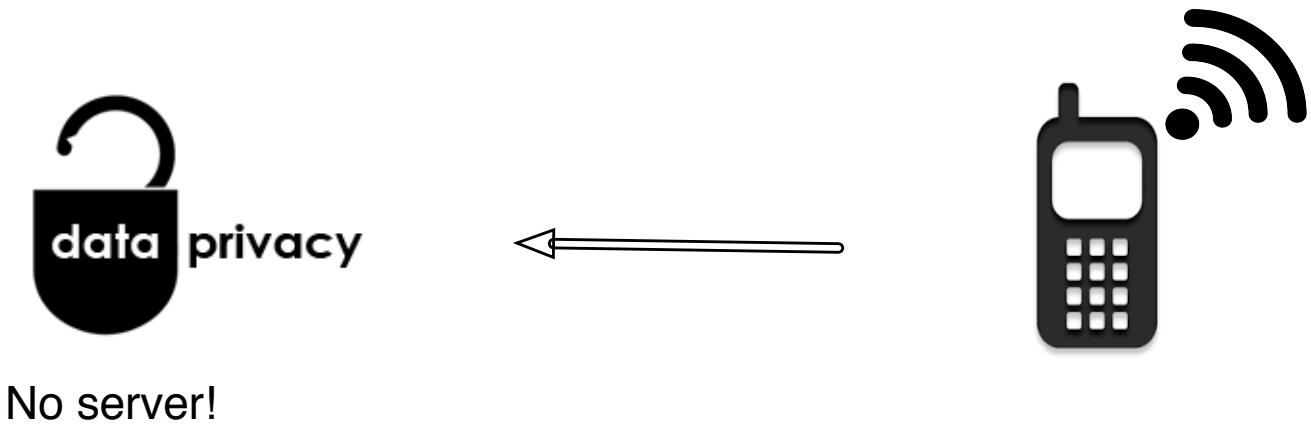
- However, the problems ...



Big model on the server

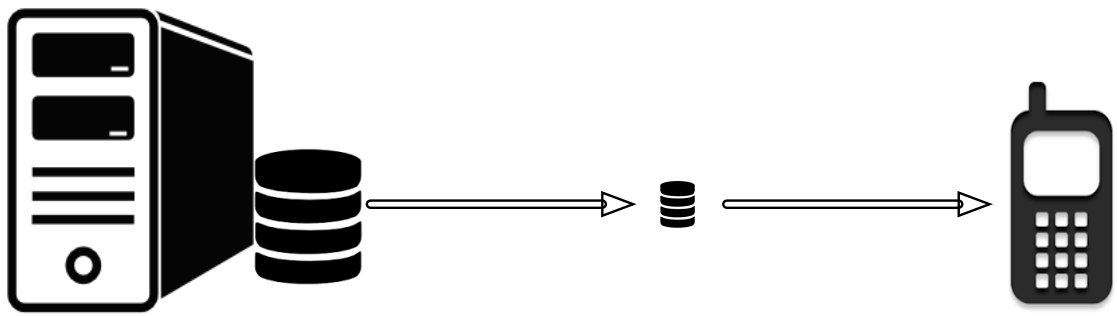
# Background

- However, the problems ...



# Background

- Small-footprint models
- Running speech recognition locally



Big model on the server

Small model in your phone

# Background: smaller models

- LSTMs and CNNs vs. DNNs
- 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

- LSTMs and CNNs vs. DNNs
- Low-ranks matrices for DNNs
- 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

# Background: smaller models

- LSTMs and CNNs vs. DNNs
- Low-ranks matrices for DNNs
- FitNet by teacher-student training
- 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

# This paper

- FitNet – Thinner and deeper networks
- No teacher-student training
- Highway connections

# This paper

- Based on papers addressing *How deep is deep?*
  - R.K.Srivastava, K.Greff, and J.Schmidhuber, "Training very deep networks," in Proc. NIPS, 2015
  - K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition in Proc. CVPR, 2016
- Image recognition now employs 100+ convolutional layers.
- Our experience
  - Reducing the error rate is difficult
  - Reducing the mode size is much easier

# Model

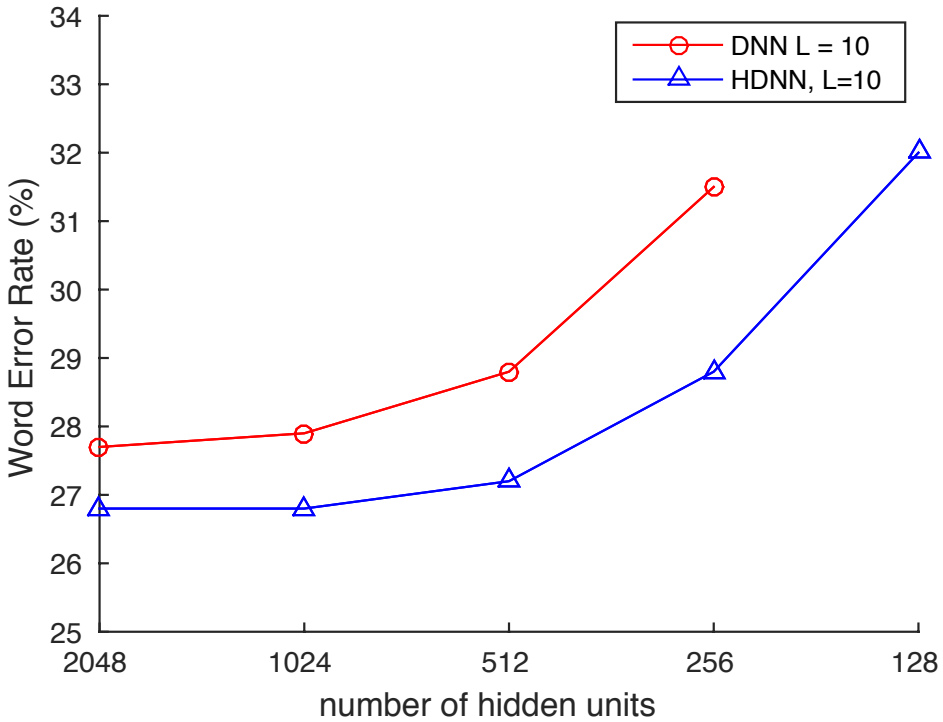
$$\mathbf{h}_l = \sigma(\mathbf{h}_{l-1}, \theta_l) \circ \underbrace{T(\mathbf{h}_{l-1}, \mathbf{W}_T)}_{\text{transform gate}} + \mathbf{h}_{l-1} \circ \underbrace{C(\mathbf{h}_{l-1}, \mathbf{W}_c)}_{\text{carry gate}} \quad (1)$$

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

# Experiments

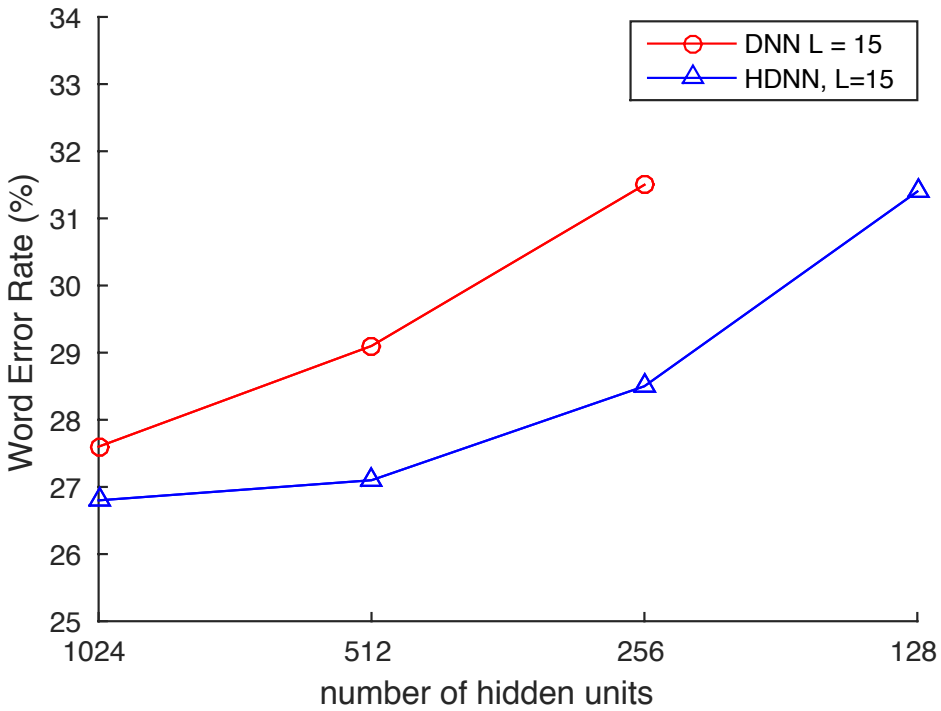
- AMI meeting speech transcription with 80h training data
- Using the standard Kaldi recipe
  - fMLLR acoustic features
  - 3-gram language models
- CNTK was used to build HDNN models
- The same decision tree was used

# Experiments – Depth and Width



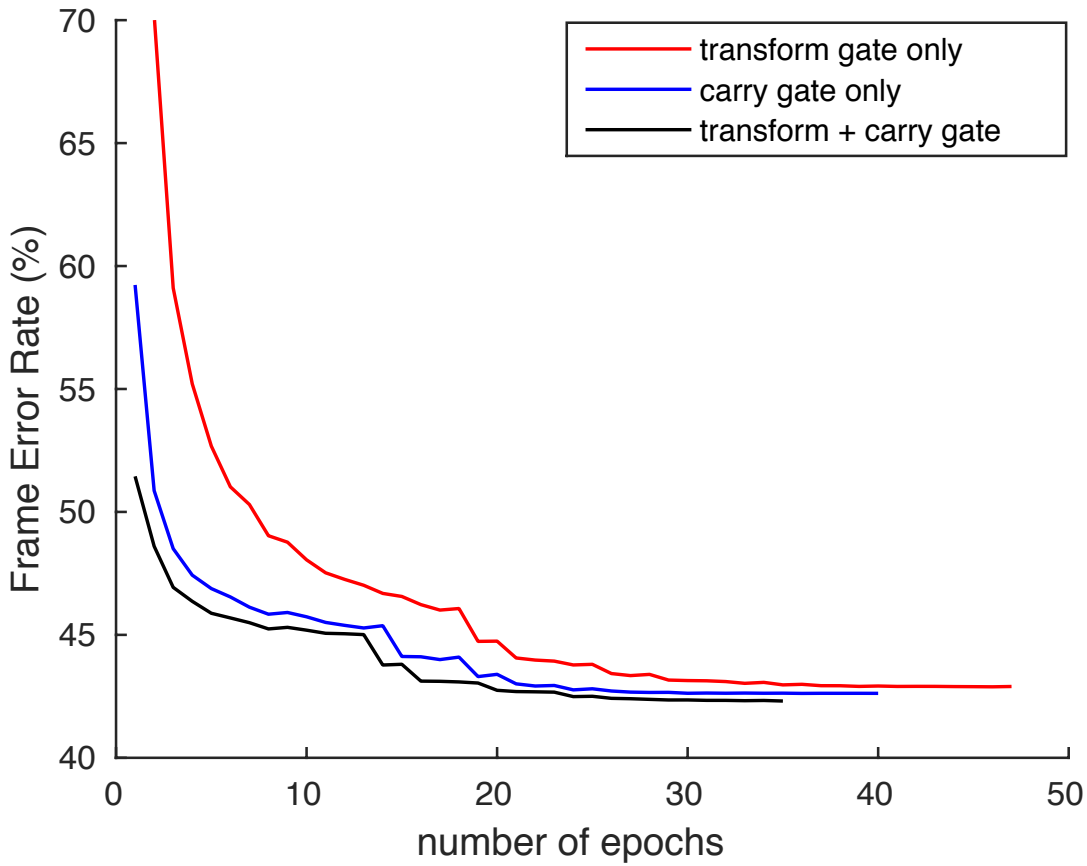

- DNNs were trained using Kaldi with RBM pretraining

# Experiments – Depth and Width




- DNNs were trained using Kaldi with RBM pretraining

# Experiments – Convergence Rate



# Experiments – Gates

Table: With and without the transform and/or carry gate.

| System | #Layer | Dim | Transform | Carry | WER |
|---|---|---|---|---|---|
| DNN* | 10 | 512 | × | × | 28.8 |
| HDNN | 10 | 512 | √ | √ | **27.2** |
| HDNN | 10 | 512 | √ | × | 27.6 |
| HDNN | 10 | 512 | × | √ | 27.5 |

- Works best with both gates

# Experiments – Constraint Gates

Table: Results of using constrained carry gate, where $C(\cdot) = \mathbf{1} - T(\cdot)$.

| System | #Layer | Dim | Constrained | WER |
|---|---|---|---|---|
| HDNN | 10 | 1024 | × | 26.8 |
| HDNN | 10 | 1024 | √ | 28.0 |
| HDNN | 10 | 512 | × | 27.2 |
| HDNN | 10 | 512 | √ | 27.4 |

- Constraint carry gate results in accuracy loss.

# Experiments – Sequence Training

Table: $\theta_h$ denotes all the model parameters of the hidden layers, $\theta_g = (\mathbf{W}_T, \mathbf{W}_c)$, and $\theta_c$ is the parameters in the softmax layer.

| Model | sMBR Update $\theta_h$ | $\theta_g$ | $\theta_c$ | WER |
|---|---|---|---|---|
| HDNN-$H_{512}L_{10}$ | × | × | × | 27.2 |
| | √ | √ | √ | 24.9 |
| | × | √ | √ | **25.2** |
| | × | √ | × | 25.8 |

[1] L. Lu, et al, "Sequence Training and Adaptation of Highway Deep Neural Networks", arXiv 2016.

# Experiments – Adaptation

Table: Results of unsupervised speaker adaptation.

| Model | Seed | Update | WER (eval) SI | SD |
|---|---|---|---|---|
| HDNN-$H_{512}L_{10}$ | sMBR | $\theta_g$ | 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, et al, "Sequence Training and Adaptation of Highway Deep Neural Networks", arXiv 2016.

# Conclusion

- Small-footprint models using highway networks
  - 2M HDNN model $\approx$ 30M DNN model after sequence training
- More adaptable and controllable
  - The tied gates largely controls the whole network
- Teacher-student training can further improve the accuracy
  - L. Lu, et al, "Knowledge Distillation for Small-footprint Highway Networks", arXiv 2016. (improves the model with $< 0.8M$ parameters)


THE UNIVERSITY of EDINBURGH
informatics

Thank you ! Questions?