Experimental Studies on Teacher-student Training of Deep Neural Network Acoustic Models

Q. Li, C. Zhang, F. L. Kreyssig and P. C. Woodland

UNIVERSITY OF CAMBRIDGE

Cambridge University Engineering Department

Abstract

- Teacher-student training investigated for DNN acoustic model compression [1].
- Teacher-student modelling allows faster and cheaper implementation of deep learning models without much loss of performance [2].
- Experiments show that soft-label trained student models outperform the hard-label trained counterpart [3].
- For a given teacher model, the student performs better as the student model complexity increases.
- For a given student model, better teacher models will result in improved student performance.
- An ensemble teacher trains student to reduce error rate further.

Teacher-student Training Overview

- Objective: train a smaller and shallower *student* model to mimic the output from a larger and deeper *teacher* model.
- Objective function: Kullback-Leibler divergence between posterior distribution of teacher $P_T(s|x)$ and student $P_S(s|x)$, *i.e.*

$$\sum_t \sum_{i=1}^{N} P_T(s_i|x_t) \log \left(\frac{P_T(s_i|x_t)}{P_S(s_i|x_t)} \right)$$

where s belongs to a set of tied triphone states, N is the total number of HMM states, and x_t is the input vector at time t.
- Equivalent to minimising the cross entropy between P_T and P_S

$$-\sum_t \sum_{i=1}^{N} P_T(s_i|x_t) \log P_S(s_i|x_t)$$

- In Figure 1, student and teacher model are concatenated.

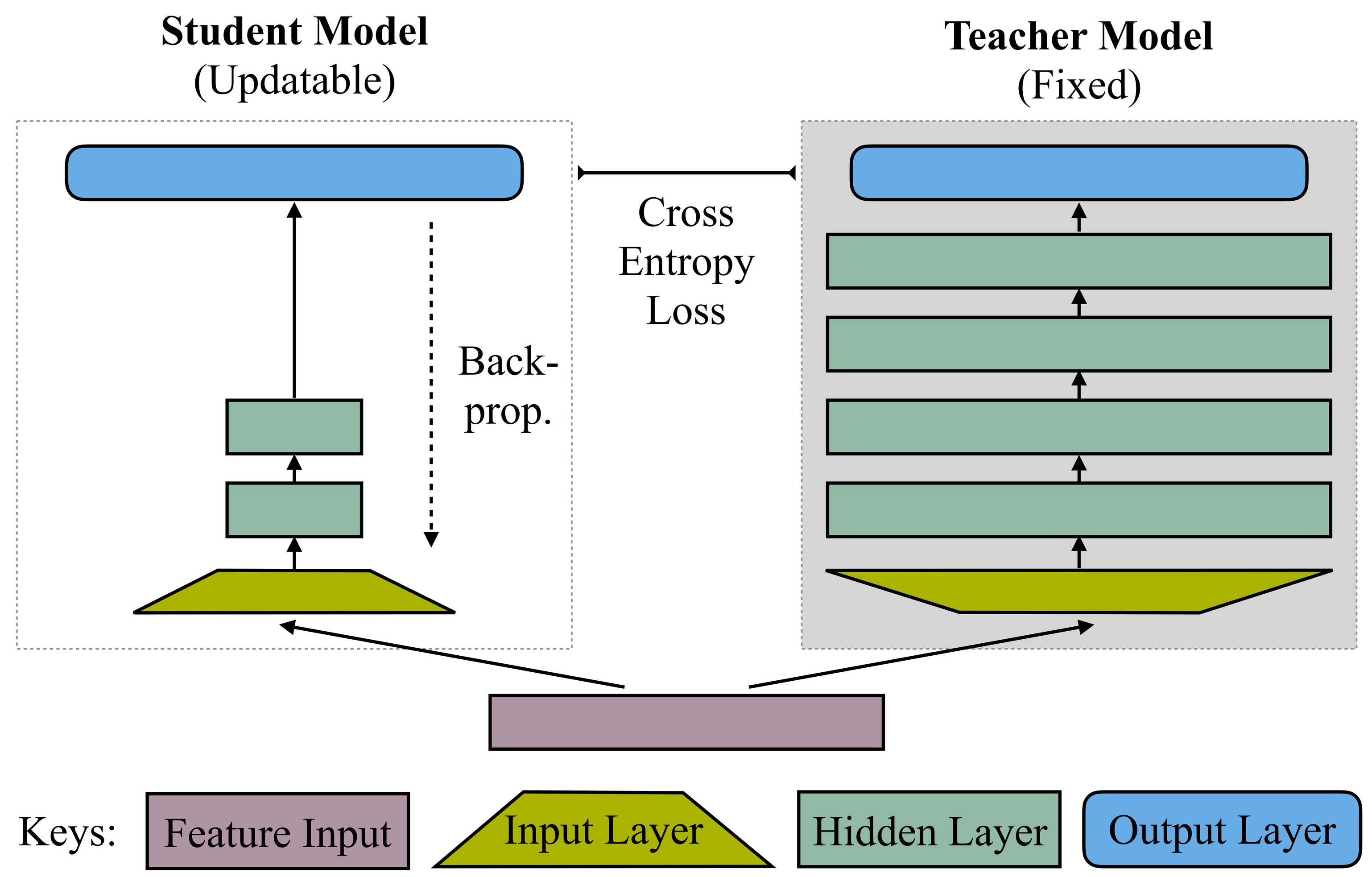

Figure 1: Teacher-student model training.

- By holding all teacher parameters fixed, only student parameters are updated.
- Loss is computed as cross entropy after Softmax output from each model, *i.e.* the target for the student is output distribution from teacher model, instead of one-hot hard labels.
- Advantages of teacher-student training:
  - Fast to train, since the student network is generally small.
  - Untranscribed data could be used for training.
  - Simple and fast in decoding.
  - Cheap to deploy on devices with limited computing resources.

Experimental Setup

- Phomeme recognition experiments are conducted on TIMIT corpus.
  - Training set: 3696 utterances (3504 training, 192 cross validation) from 462 speakers, 3.14 hours.
  - Full test set: 1344 utterances from 168 speakers, 0.81 hours.
  - 13 dimensional MFCC features with Δ and $\Delta\Delta$.
  - Standard dictionary and bigram language model.
- All systems are trained and decoded using HTK 3.5.

Experimental Results

- Fully-connected 7-layer teacher models (Figure 2):

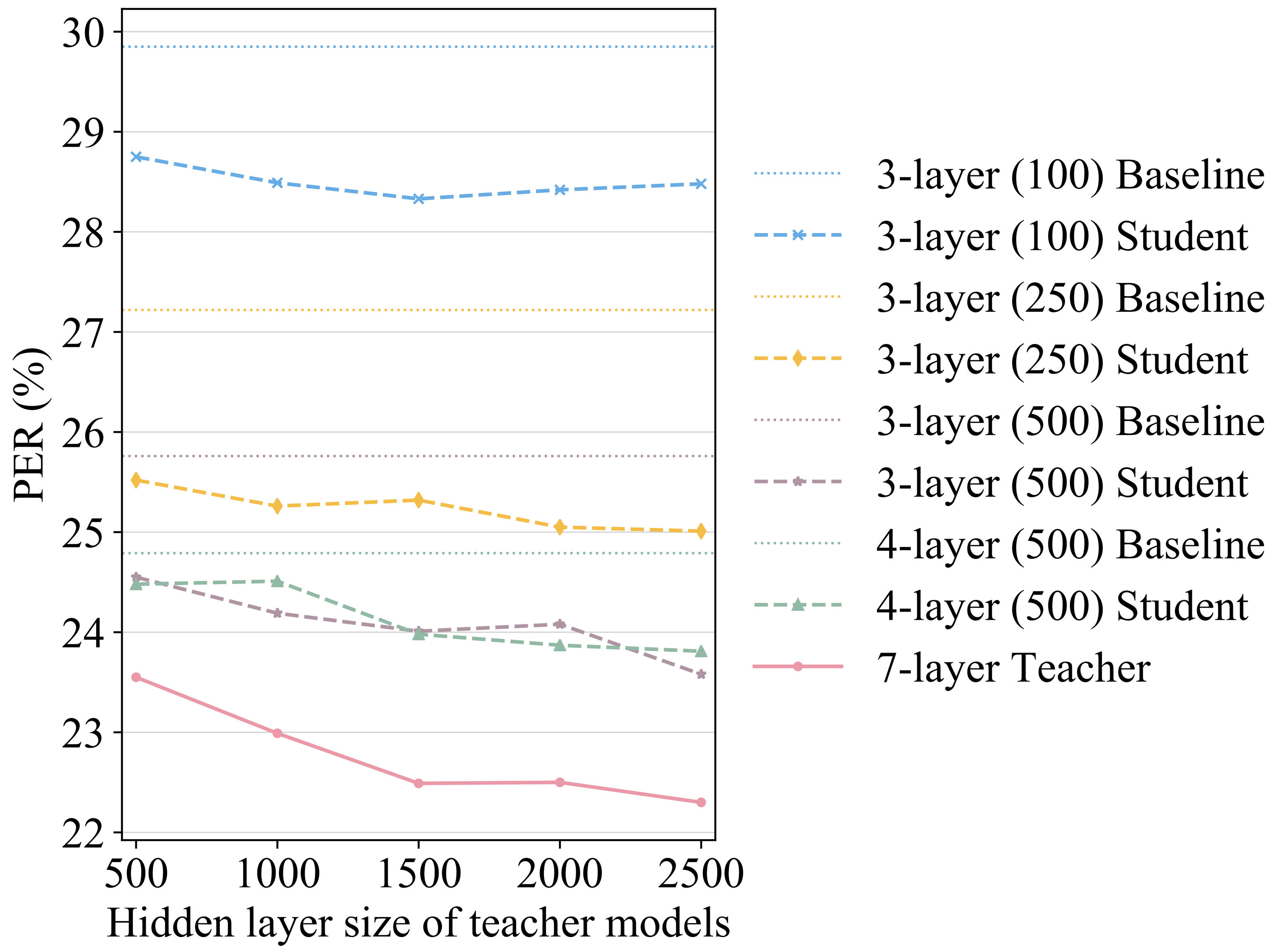

Figure 2: Phone error rate (PER) of shallow student models trained from 7-layer teacher models with various layer sizes.

  - All student models perform better than hard-label trained baselines.
  - Student model performance is restricted by both model complexity and teacher performance.
  - For a very simple model, the gain is limited due to its weak modelling capability.
  - For a more complex student model, as the gap between baseline and teacher performance narrows, the gain diminishes.
  - 3-layer (250) student model outperforms 3-layer (500) baseline, with ~ 50% parameters.
  - 3-layer (500) student model outperforms 4-layer (500) baseline, with ~ 70% parameters.
- RNN model [4] and ensemble model [5]:
  - RNN architecture: 1 recurrent layer followed by a hidden layer and an output layer.
  - Ensemble architecture: linear ensemble between the above RNN and the fully-connected 7-layer model with layer size of 500, *i.e.* the arithmetic average of two Softmax outputs.

Teacher Arch.	T-S PER (%)	Ref. PER (%)
7-layer (500)	24.55	23.55
RNN	23.84	20.59
Ensemble	**23.73**	**20.34**

Table 1: RNN and ensemble teacher models with a 3-layer (500) fully-connected student model. (student baseline PER 25.76%)

Conclusions

- Capability of simple models is not fully exploited by hard-label training: improved by teacher-student training.
- Soft labels easier to match as richer and smoother knowledge available.
- Teacher-student training less prone to incorrect hard labels, which may contribute to the student gain.
- Teacher-student training is generally applicable for model compression with little restrictions on model architecture.

References

[1] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in Proc. *KDD*, 2006.

[2] J. Ba and R. Caurana, "Do deep nets really need to be deep?" in Proc. *NIPS*, 2014.

[3] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size dnn with output-distribution-based criteria," in Proc. *Interspeech*, 2014.

[4] A. Graves, A. rahman Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. *ICASSP*, 2013.

[5] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in Proc. *Interspeech*, 2014.

UK Speech 2017, Sept. 11-12 | Cambridge, U.K.

Email: {ql264,cz277,flk24,pcw}@eng.cam.ac.uk