Segmental Recurrent Neural Networks for End-to-end Speech Recognition

Liang Lu, Lingpeng Kong, Chris Dyer,
Noah Smith and Steve Renals

TTI-Chicago, UoE, CMU and UW

9 September 2016

Background

- A new wave of sequence modelling
  - I. Sutskever, et al., "Sequence-to-Sequence Learning with Neural Networks", NIPS 2014
  - D. Bahdanau, et al., "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015
  - A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks, ICML 2014

Background

- Maybe time to review sequence modelling for speech
- Why speech recognition is special?
  - monotonic alignment
  - long input sequence
  - output sequence is much shorter (word/phonme)

Speech Recognition

- monotonic alignment
  - challenges for attention models
- long input sequence
  - expensive for globally (sequence-level) normalised model
- output sequence is much shorter (word/phonme)
  - length mismatch – alignment model or not?

Speech Recognition

- Hidden Markov Model
  - monotonic alignment $\surd$
  - long input sequence $\rightarrow$ locally (frame-level) normalised
  - length mismatch $\rightarrow$ hidden states
- Connectionist Temporal Classification
  - monotonic alignment $\surd$
  - long input sequence $\rightarrow$ locally normalised
  - length mismatch $\rightarrow$ blank state

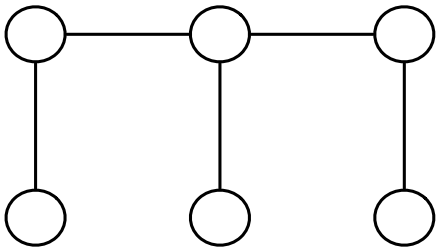
Speech Recognition

- Locally normalised models:
  - conditional independence assumption
  - label bias problem
  - better results given by sequence training: local $\rightarrow$ global normalisation
- Question:
  Why not sticking to the globally normalised models from scratch?

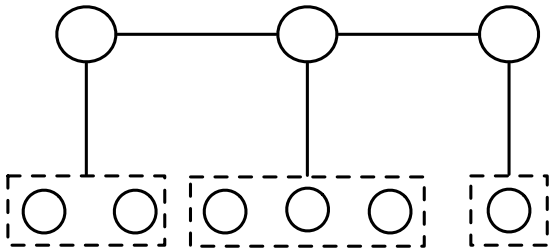
[1] D. Andor, et al, "Globally Normalized Transition-Based Neural Networks", ACL, 2016.

[2] D. Povey, et al, "Purely sequence-trained neural networks for ASR based on lattice-free MMI" Interspeech, 2016

(Segmental) Conditional Random Field



CRF



segmental CRF

(Segmental) Conditional Random Field

- CRF [Lafferty et al. 2001]

$$P(\mathbf{y}_{1:L} \mid \mathbf{x}_{1:T}) = \frac{1}{Z(\mathbf{x}_{1:T})} \prod_j \exp\left(\mathbf{w}^\top \Phi(y_j, \mathbf{x}_{1:T})\right) \tag{1}$$

  where $L = T$.

- Segmental (semi-Markov) CRF [Sarawagi and Cohen 2004]

$$P(\mathbf{y}_{1:L}, \mathbf{E}, \mid \mathbf{x}_{1:T}) = \frac{1}{Z(\mathbf{x}_{1:T})} \prod_j \exp\left(\mathbf{w}^\top \Phi(y_j, \mathbf{e}_j, \mathbf{x}_{1:T})\right) \tag{2}$$

  where $\mathbf{e}_j = \langle s_j, n_j \rangle$ denotes the beginning (s_j) and end (n_j) time tag of y_j; $\mathbf{E} = \{\mathbf{e}_{1:L}\}$ is the latent segment label.

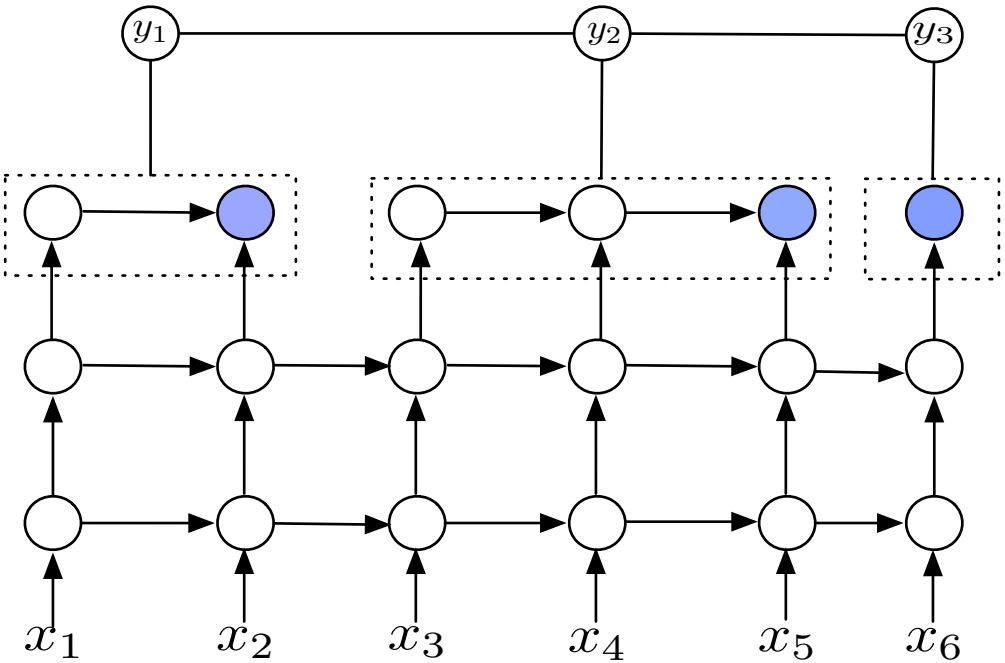
(Segmental) Conditional Random Field

$$\frac{1}{Z(\mathbf{x}_{1:T})} \prod_j \exp\left(\mathbf{w}^\top \Phi(y_j, \mathbf{x}_{1:T})\right)$$

- Learnable parameter $\mathbf{w}$
- Engineering the feature function $\Phi(\cdot)$
- Designing $\Phi(\cdot)$ is much harder for speech than NLP

Segmental Recurrent Neural Network

- Using (recurrent) neural networks to learn the feature function $\Phi(\cdot)$.



Segmental Recurrent Neural Network

- Training criteria
  - Conditional maximum likelihood

$$\begin{aligned}\mathcal{L}(\theta) &= \log P(\mathbf{y}_{1:L} \mid \mathbf{x}_{1:T}) \\ &= \log \sum_{\mathbf{E}} P(\mathbf{y}_{1:L}, \mathbf{E} \mid \mathbf{x}_{1:T}) \end{aligned} \tag{3}$$

  - Hinge loss – similar to structured SVM

.

`Not studied yet!`

Segmental Recurrent Neural Network

- Viterbi decoding
  - Partially Viterbi decoding

$$\mathbf{y}^*_{1:L} = \arg\max_{\mathbf{y}_{1:L}} \log \sum_{\mathbf{E}} P(\mathbf{y}_{1:L}, \mathbf{E} \mid \mathbf{x}_{1:T}) \tag{4}$$

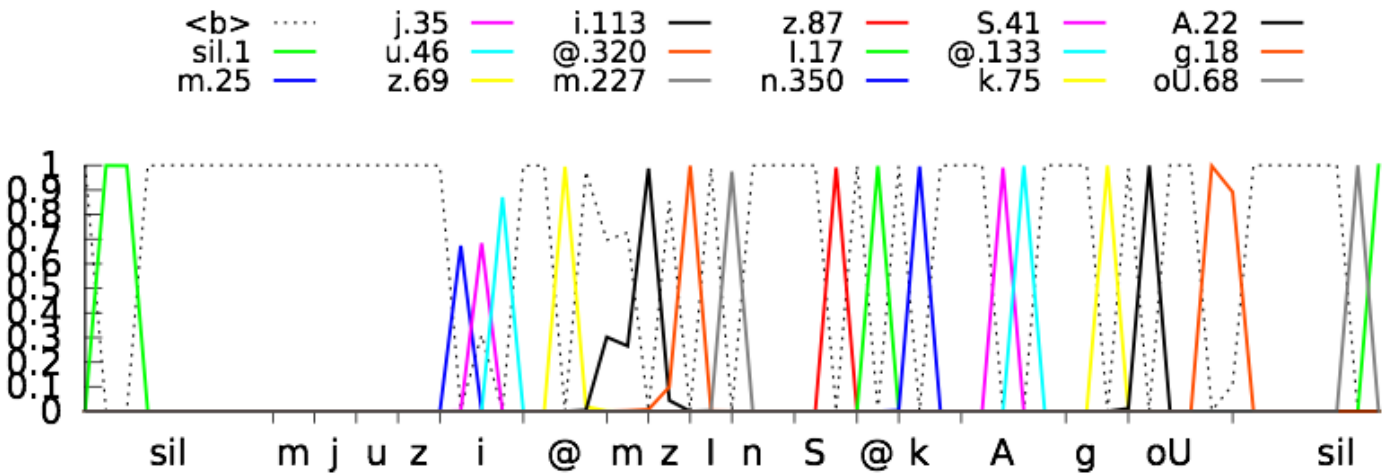
  - Fully Viterbi decoding

$$y^*_{1:L}, \mathbf{E}^* = \arg\max_{\mathbf{y}_{1:L}, \mathbf{E}} \log P(\mathbf{y}_{1:L}, \mathbf{E} \mid \mathbf{x}_{1:T}) \tag{5}$$

Related works

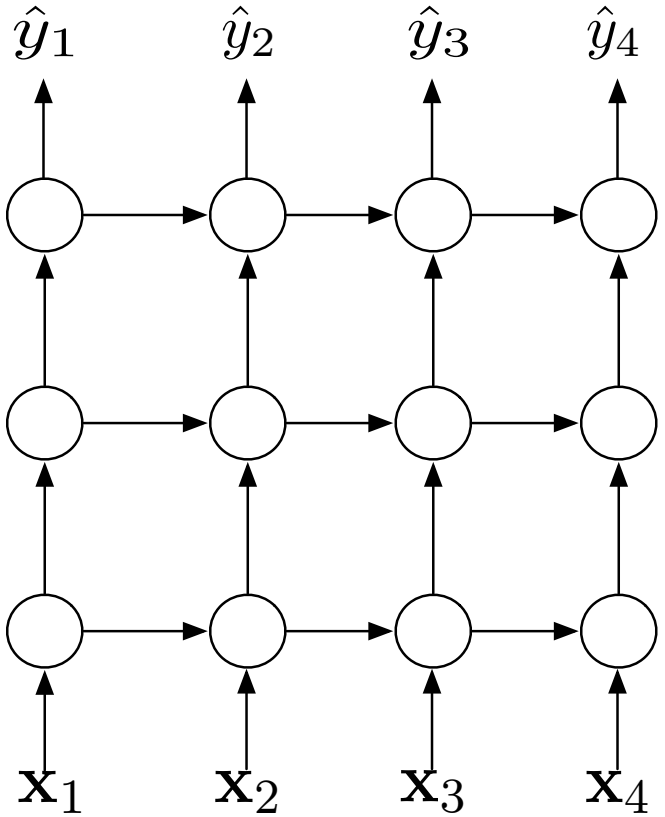
- (Segmental) CRFs for speech
- Neural CRFs
- Structured SVMs
- Two good review papers
  - M. Gales, S. Watanabe and E. Fosler-Lussier, "Structured Discriminative Models for Speech Recognition", IEEE Signal Processing Magazine, 2012
  - E. Fosler-Lussier et al. "Conditional random fields in speech, audio, and language processing, Proceedings of the IEEE, 2013

Comparison to CTC



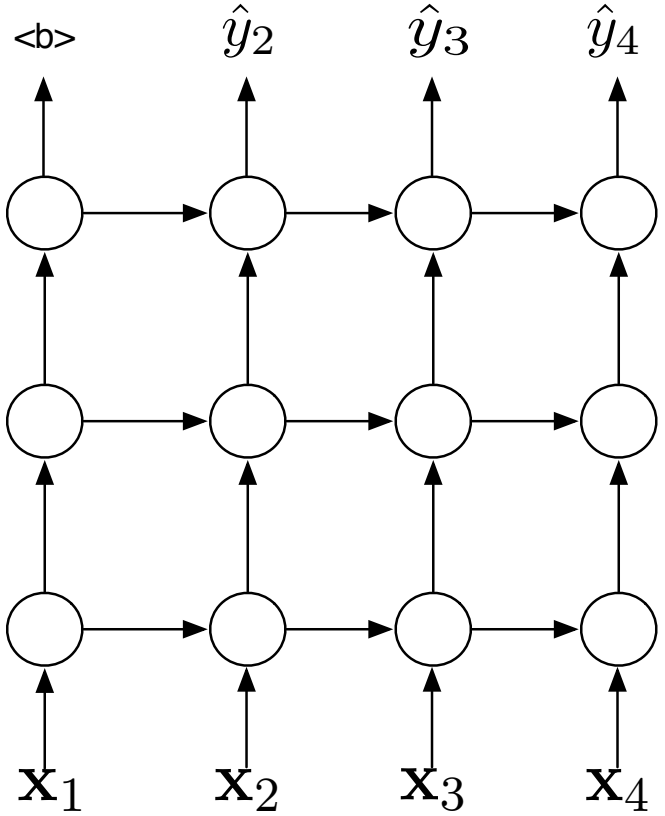
[1] A. Senior, et al, "Acoustic Modelling with CD-CTC-sMBR LSTM RNNs", ASRU 2015.

Comparison to CTC



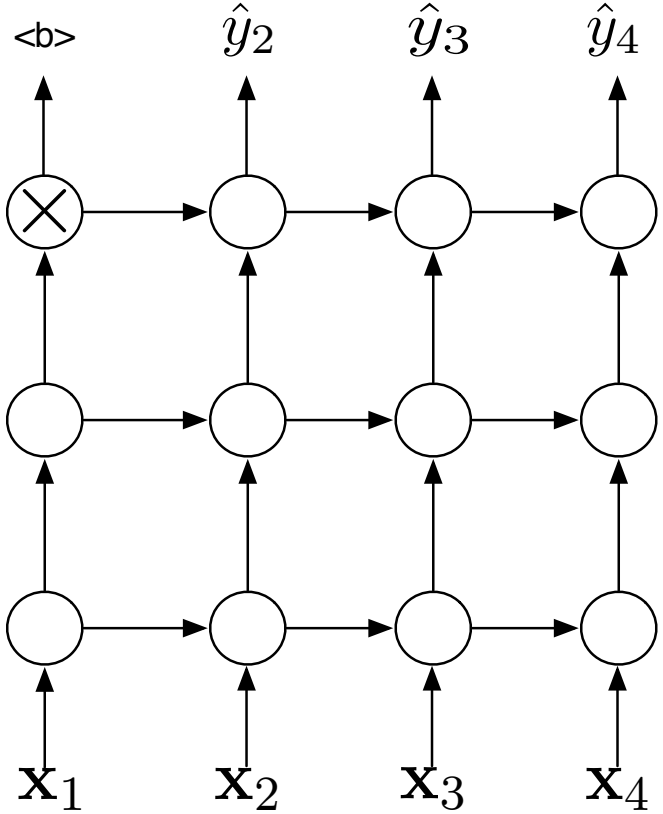
$$p(\mathbf{y} \mid \mathbf{x}) = p(\hat{y}_1|x_1) \times p(\hat{y}_2|x_2) \times p(\hat{y}_3|x_3) \times p(\hat{y}_4|x_4)$$

Comparison to CTC



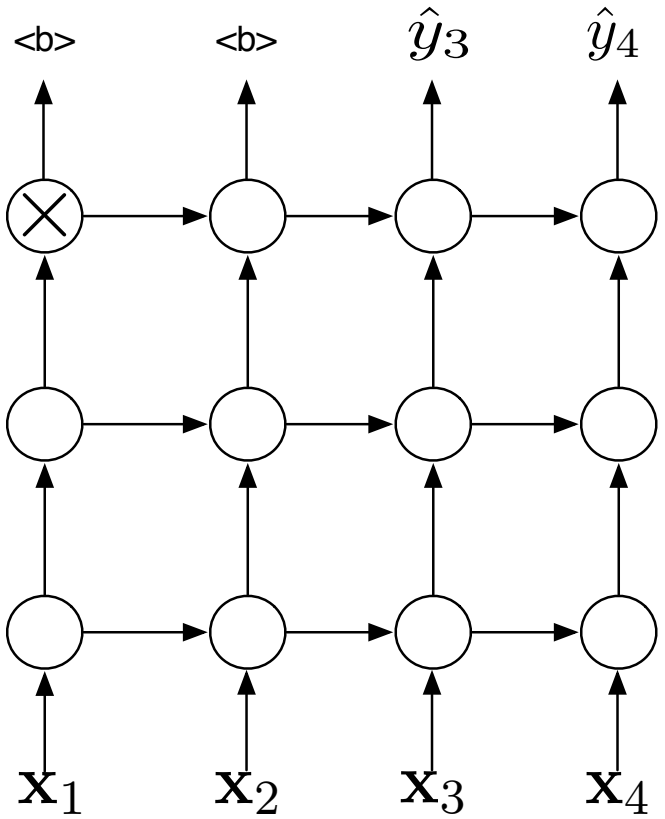
$$p(\mathbf{y} \mid \mathbf{x}) = \underbrace{p(\langle b \rangle | x_1)}_{=1} \times p(\hat{y}_2 | x_2) \times p(\hat{y}_3 | x_3) \times p(\hat{y}_4 | x_4)$$

Comparison to CTC



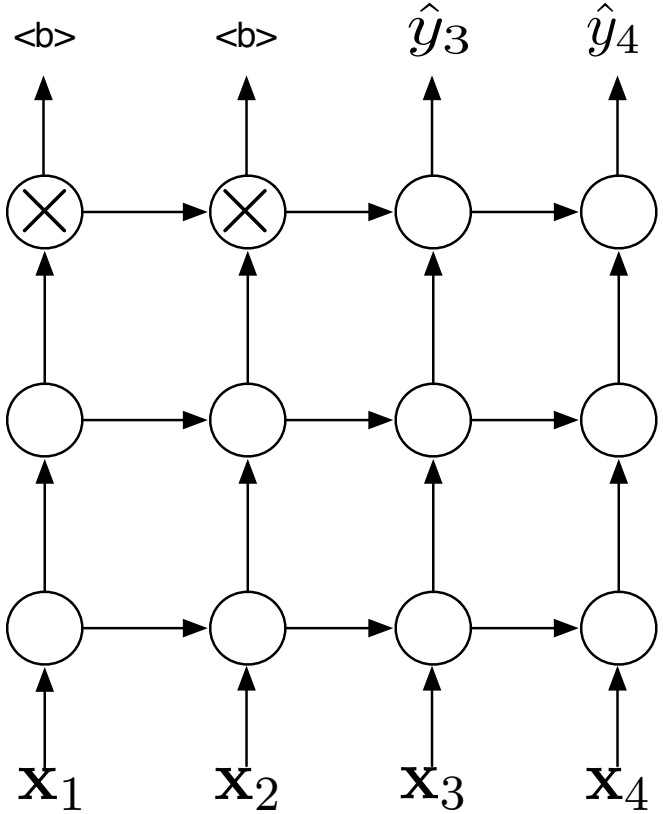
$$p(\mathbf{y} \mid \mathbf{x}) = \cancel{p(\langle b \rangle | x_1)} \times p(\hat{y}_2|x_2) \times p(\hat{y}_3|x_3) \times p(\hat{y}_4|x_4)$$

Comparison to CTC



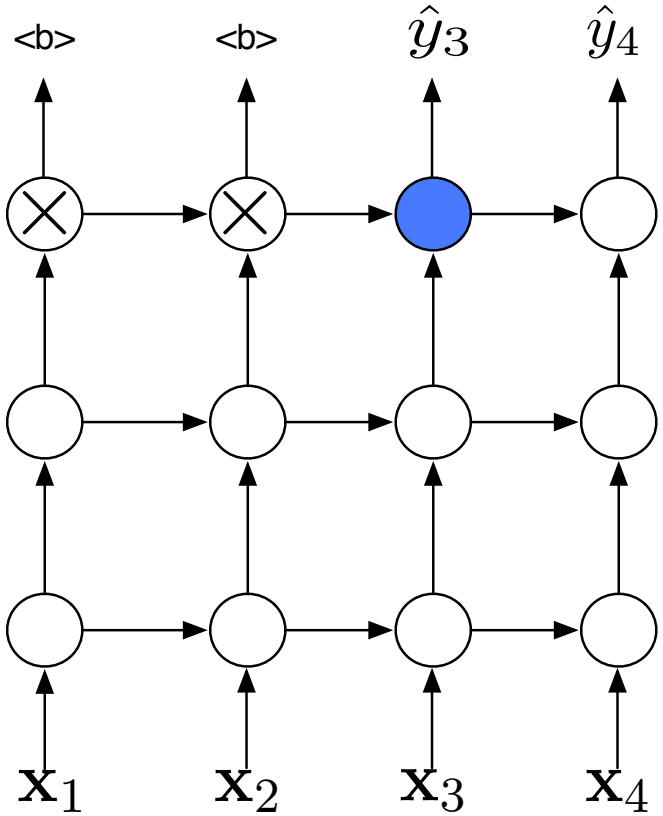
$$p(\mathbf{y} \mid \mathbf{x}) = \underbrace{p(\langle b \rangle | x_1)}_{=1} \times \underbrace{p(\langle b \rangle | x_2)}_{=1} \times p(\hat{y}_3 | x_3) \times p(\hat{y}_4 | x_4)$$

Comparison to CTC

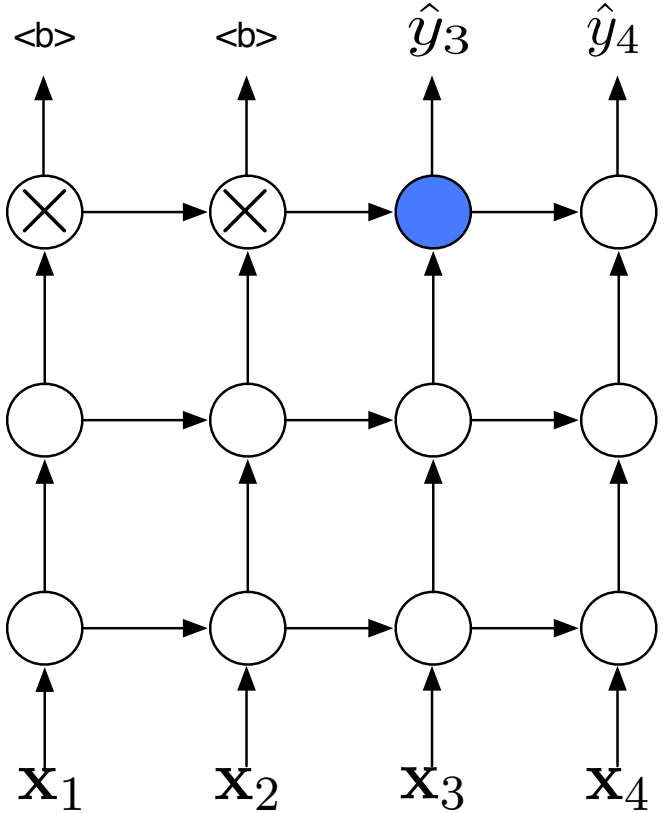


$$p(\mathbf{y} \mid \mathbf{x}) = \cancel{p(\langle b \rangle | x_1)} \times \cancel{p(\langle b \rangle | x_2)} \times p(\hat{y}_3 | x_3) \times p(\hat{y}_4 | x_4)$$

Comparison to CTC



Comparison to CTC



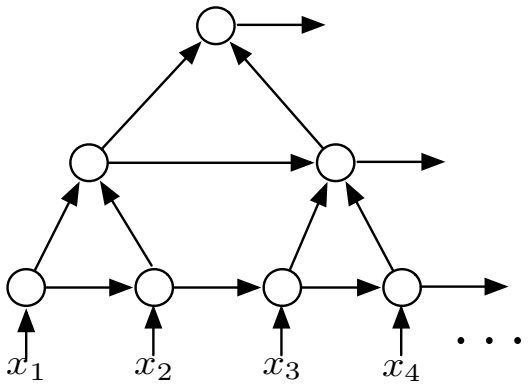
- CTC loss may do some kind of segmental modelling

Experiment

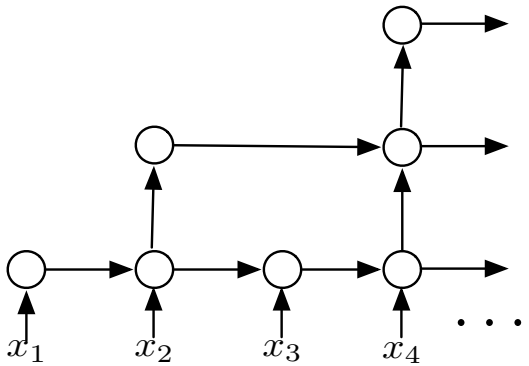
- TIMIT dataset
  - 3696 training utterances ($\sim$ 3 hours)
  - core test set (192 testing utterances)
  - trained on 48 phonemes, and mapped to 39 for scoring
  - log filterbank features (FBANK)
  - using LSTM as an implementation of RNN

Experiment

- Limit the lengths of segments
- Recurrent subsampling networks – over 10x speedup



a) concatenate / add



b) skip

Experiment

- Large model with dropout works the best

Table: Results of dropout.

Dropout	layers	hidden	PER
0.2	3	128	21.2
	3	250	20.1
	6	250	19.3
0.1	3	128	21.3
	3	250	20.9
	6	250	20.4
×	6	250	21.9

Experiment

Table: Results of three types of acoustic features.

Features	Deltas	$d(\mathbf{x}_t)$	PER
24-dim FBANK	$\surd$	72	19.3
40-dim FBANK	$\surd$	120	18.9
Kaldi	$\times$	40	17.3

Kaldi features – 39 dimensional MFCCs spliced by a context window of 7, followed by LDA and MLLT transform and with feature-space speaker-dependent MLLR

Experiment

Table: Comparison to related works. LM = language model, SD = speaker dependent feature

System	LM	SD	PER
HMM-DNN	$\surd$	$\surd$	18.5
CTC [Graves 2013]	$\times$	$\times$	18.4
RNN transducer [Graves 2013]	–	$\times$	17.7
Attention-based RNN [Chorowski 2015]	–	$\times$	17.6
Segmental RNN	$\times$	$\times$	18.9
Segmental RNN	$\times$	$\surd$	17.3

Conclusion

- Segmental CRFs with recurrent neural networks
- Potential for end-to-end training
- Computational cost is the main bottleneck
- Need to evaluate on large vocabulary tasks

THE UNIVERSITY of EDINBURGH
informatics

Thank you ! Questions?