Neural Segmental CRFs for Sequence Modelling

Liang Lu

Based on the work with Lingpeng Kong @ CMU

28 June 2016

Carry on from Mark[1]

- Sequence-to-sequence modelling
  - speech synthesis:
    word sequence $\rightarrow$ waveform
  - speech recognition:
    waveform $\rightarrow$ word sequence
  - machine translation:
    word sequence $\rightarrow$ word sequence

[1] Assuming that you were in the talk with attention and long-term memory.

Next Question

- Is speech recognition more special?
  - monotonic alignment
  - long input sequence
  - output sequence is much shorter (word/phonme)

Speech Recognition

- monotonic alignment
  - encoder-decoder model does not naturally apply
  - $\mathbf{x}_{1:T} \rightarrow \mathbf{c} \rightarrow \mathbf{y}_{1:L}$
- long input sequence
  - expensive for global normalised model
- output sequence is much shorter (word/phonme)
  - length mismatch

Speech Recognition

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

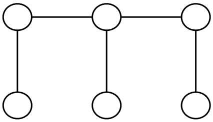
Speech Recognition

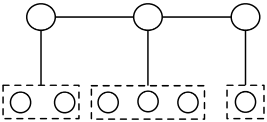
- Locally normalised models:
  - conditional independence assumption
  - label bias problem
  - better results given by sequence training: local $\rightarrow$ global normalisation
- Question:
  Why not sticking to the globally normalised models from scratch?

[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





(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.

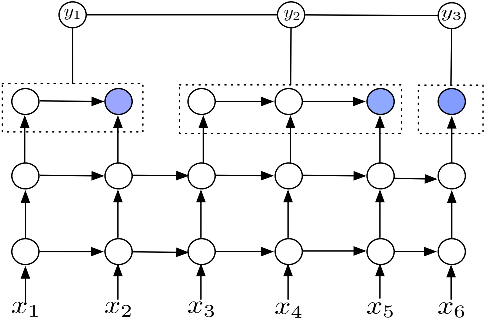
(Segmental) Conditional Random Field

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

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

Neural Segmental CRF

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



[1] Y. Liu, et al, "Exploring Segment Representations for Neural Segmentation Models", arXiv 2016.

Neural conditional random fields

- 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

.

`something complicated!`

Neural conditional random fields

- 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

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

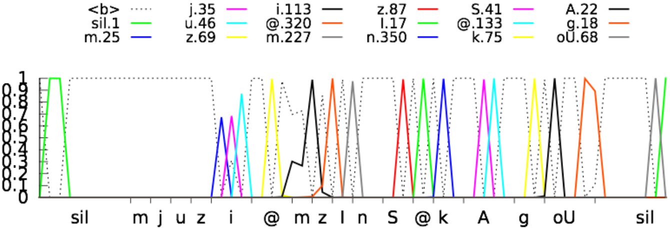
---

[1] L. Lu, et al, "Segmental Recurrent Neural Networks for End-to-end Speech Recognition", Interspeech 2016.

Related works

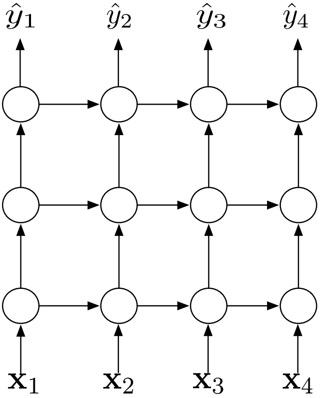
- (Segmental) CRFs for speech
- Neural CRFs
- Structured SVMs

Comparison to CTC

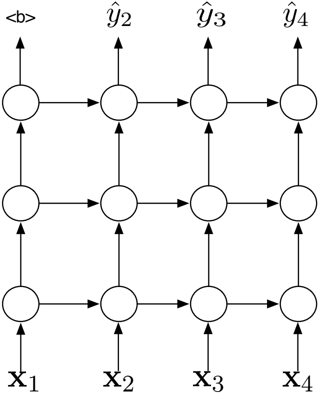


[1] A. Senior, et al, "Acoustic Modelling with CD-CTC-sMBR LSTM RNNs", ASRU 2015.

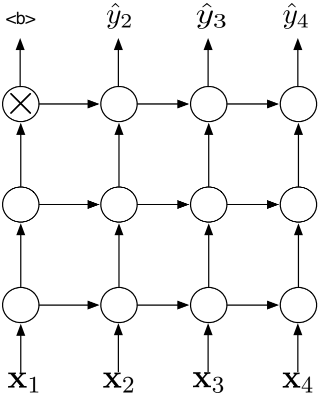
Comparison to CTC



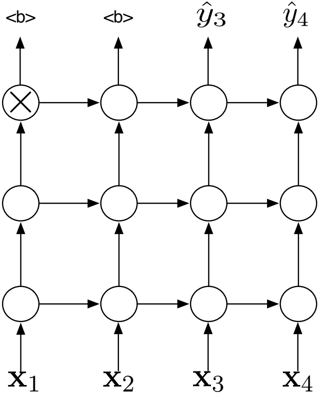
Comparison to CTC



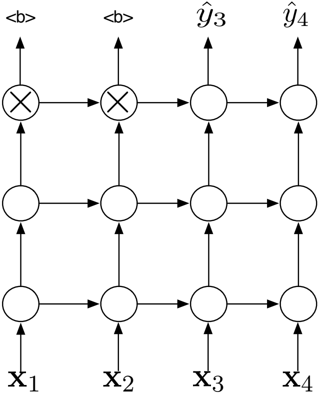
Comparison to CTC



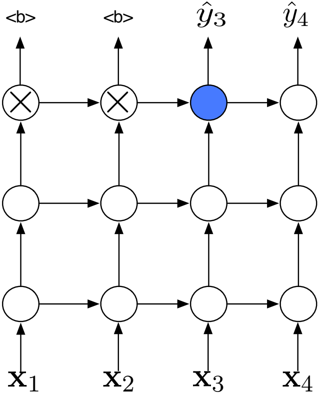
Comparison to CTC



Comparison to CTC



Comparison to CTC

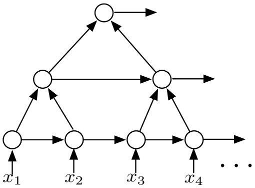


Experiment

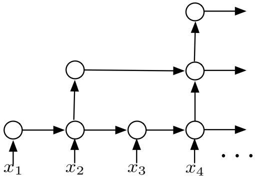
- TIMIT dataset
  - 3696 training utterances ($\sim$ 3 hours)
  - core test set (192 testing utterances)
  - trained on 48 phonemes, and mapped to 39 for scoring
  - log filterbank features (FBANK)
  - using LSTM as an implementation of RNN

Experiment

- Speed up training



a) concatenate / add



b) skip

Experiment

Table: Results of tuning the hyperparameters.

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
$\times$	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.

System	LM	SD	PER
HMM-DNN	√	√	18.5
first-pass SCRF [Zweig 2012]	√	×	33.1
Boundary-factored SCRF [He 2012]	×	×	26.5
Deep Segmental NN [Abdel 2013]	√	×	21.9
Discriminative segmental cascade [Tang 2015]	√	×	21.7
+ 2nd pass with various features	√	×	19.9
CTC [Graves 2013]	×	×	18.4
RNN transducer [Graves 2013]	–	×	17.7
Attention-based RNN [Chorowski 2015]	–	×	17.6
Segmental RNN	×	×	18.9
Segmental RNN	×	√	17.3

Conclusion

- Neural Segmental CRFs are flexible and powerful sequence models
  - handwriting recognition
  - joint word segmentation and POS tagging
- However, speed matters for large vocabulary speech recognition
  - WFST-based decoder
  - context-dependent vs. context-independent phones

[1] L. Kong, et al, "Segmental Recurrent Neural Networks", ICLR 2016.

Thank you ! Questions?