On Training the RNN Encoder-Decoder for Large Vocabulary End-to-end Speech Recognition

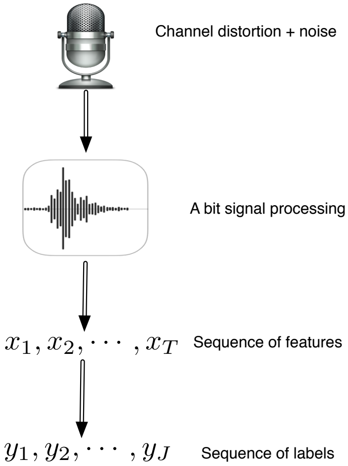
Liang Lu, Xingxing Zhang, Steve Renals

Centre for Speech Technology Research
The University of Edinburgh

23 March 2016

Speech recognition problem – review

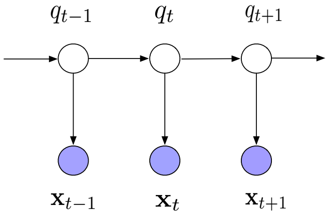
- A sequence to sequence transduction problem
- Given $\mathbf{y} = \{y_1, \cdots, y_J\}, y \in \mathcal{Y}$ and $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$, compute $P(\mathbf{y} \mid \mathbf{X})$
- However, it is difficult
  - $T \gg J$ and T can be large (> 1000)
  - Large size of vocabulary $|\mathcal{Y}| \approx 60K$
  - Uncertainty and variability in features
  - Context-dependency problem
  - ...



Hidden Markov Models

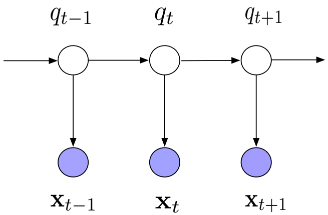
- Hidden Markov Models — convert the sequence-level classification problem into a frame-level problem

$$
\begin{aligned}
P(\mathbf{y} \mid \mathbf{X}) &\propto p(\mathbf{X} \mid \mathbf{y}) \\
&\approx p(\mathbf{X}_{1:T} | Q_{1:T}) P(\mathbf{y}) \\
&\approx P(\mathbf{y}) \prod_t p(\mathbf{x}_t | q_t) p(q_t | q_{t-1})
\end{aligned}
$$



Hidden Markov Models

- Problems of HMMs:
  - Loss function: minimise the word error $\mathcal{L}(\mathbf{y}, \tilde{\mathbf{y}})$ versus maximise the likelihood $p(\mathbf{X}_{1:T} | Q_{1:T})$
  - Conditional independence assumption
  - Weak sequence model – first order Markov rule
  - System complexity: monophone $\rightarrow$ alignment $\rightarrow$ triphone $\rightarrow$ alignment $\rightarrow$ neural net $\rightarrow$ alignment $\rightarrow$ neural net



End-to-end speech recognition

- Can we train a model that directly computes $P(\mathbf{y} \mid \mathbf{X})$?
- CTC – Connectionist Temporal Classification
- Attention-based recurrent neural network (RNN) encoder-decoder

End-to-end speech recognition

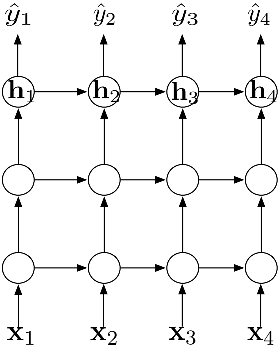
- CTC – Connectionist Temporal Classification
  - Method: $\{y_1, \cdots, y_J\} \rightarrow \{\hat{y}_1, \cdots, \hat{y}_T\} \rightarrow \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$
  - Replicate the labels (a b c $\rightarrow$ a a b b b $\oslash$ c) with *blank* symbol $\oslash$
  - Approximate the conditional probability

$$P(\hat{\mathbf{y}} \mid \mathbf{X}) = \prod_{t=1}^{T} P(\hat{y}_t \mid \mathbf{x}_t) \tag{1}$$

[1] A. Graves, et al, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks", ICML 2006
[2] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks", ICML 2014
[3] A. Hannun, et al, "Deep Speech: Scaling up end-to-end speech recognition", arXiv 2014
[4] H. Sak, et al, "Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition", INTERSPEECH 2015

End-to-end speech recognition

- Still reply on the **independence** assumption
- RNN may help to mitigate the problem



End-to-end speech recognition

- Attention-based RNN encoder-decoder

$$P(\mathbf{y} \mid \mathbf{X}) \approx \prod_j P(y_j \mid y_1, \cdots, y_{j-1}, \mathbf{c}_j) \tag{2}$$

$$\mathbf{h}_{1:T} = \mathrm{RNN}(\mathbf{x}_{1:T}) \tag{3}$$

$$\mathbf{c}_j = \mathrm{Attend}(\mathbf{h}_{1:T}) \tag{4}$$

[1] D. Bahdanau, et al, "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015

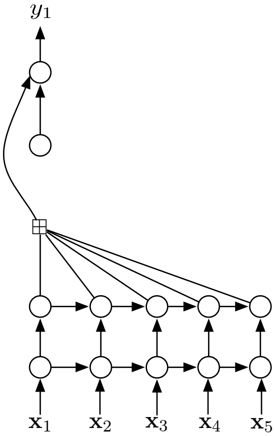
[2] J. Chorowski, et al, "Attention-Based Models for Speech Recognition", NIPS 2015

[3] L. Lu et al, "A Study of the Recurrent Neural Network Encoder-Decoder for Large Vocabulary Speech Recognition", INTERSPEECH 2015

[4] W. Chan, et al, "Listen, Attend and Spell", arXiv 2015

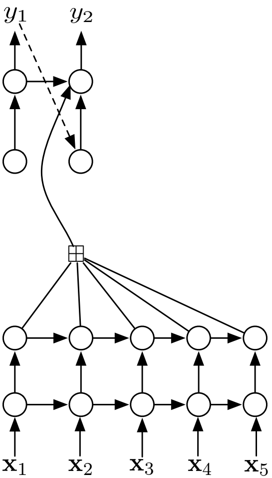
End-to-end speech recognition

- Attention-based RNN encoder-decoder



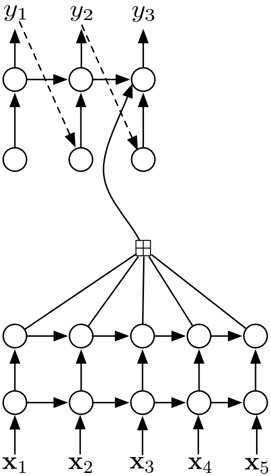
End-to-end speech recognition

- Attention-based RNN encoder-decoder



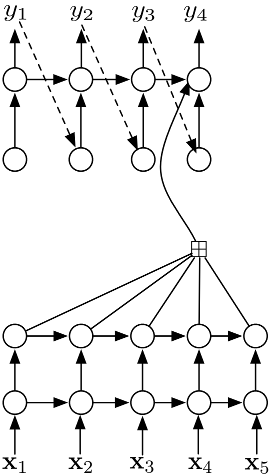
End-to-end speech recognition

- Attention-based RNN encoder-decoder



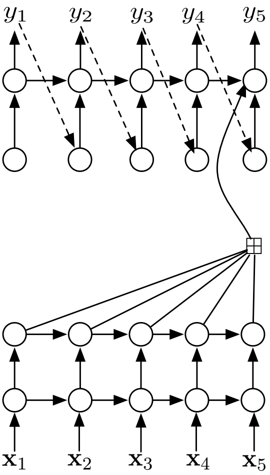
End-to-end speech recognition

- Attention-based RNN encoder-decoder



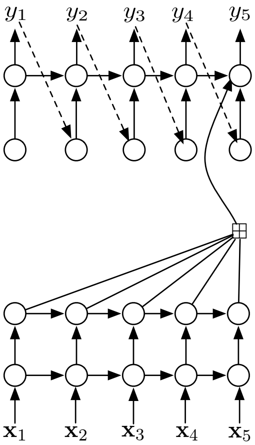
End-to-end speech recognition

- Attention-based RNN encoder-decoder



End-to-end speech recognition

- Attention-based RNN encoder-decoder



Decoder $P(y_j \mid y_1, \cdots, y_{j-1}, \mathbf{c}_j)$

Attention $\mathbf{c}_j = \mathrm{Attend}(\mathbf{h}_{1:T})$

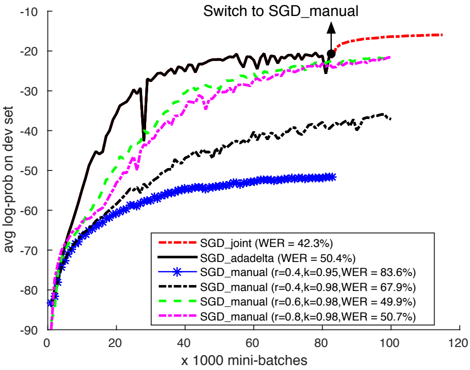
Encoder $\mathbf{h}_{1:T} = \mathrm{RNN}(\mathbf{x}_{1:T})$

End-to-end speech recognition

- In this paper, we look at three aspects of this model
  - SGD optimisation
  - Implicit language modelling
  - Word vs. Character output labels
- Dataset – Switchboard (300 hours $\approx$ 100 million frames)

Experiment

- SGD optimisation
  - It takes around 2 week to run 15 epochs in our baseline configuration
  - Tuning SGD learning rate is expensive
  - Adaptive SGD learning rate – AdaGrad, AdaDelta, Adam, ...



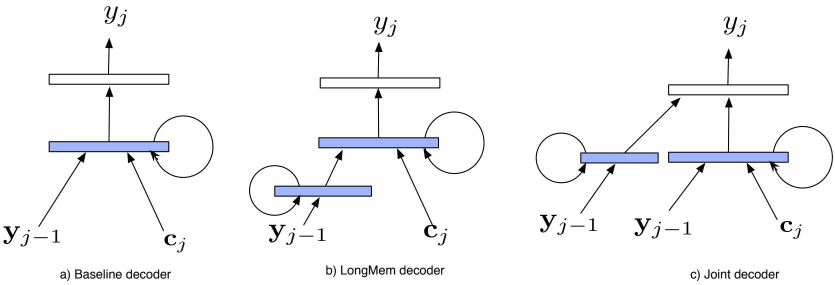
Experiments

Table: Scheduling the SGD learning rates.

SGD learning rate	Feature	SWB
SGD_adadelta	MFCC	38.8
+ manual SGD	MFCC	36.2
SGD_adadelta	FBANK	34.7
+ manual SGD	FBANK	26.8

Experiment

- Implicit RNN language modelling



Experiment

Table: Implicit RNN language modelling.

System	Output	Avg
EncDec no LM	word	26.8
+ LongMem	word	26.3
+ 3-gram rescoring	word	**25.8**
EncDec no LM	char	32.8
+ LongMem	char	30.9
+ 5-gram rescoring	char	30.5

Experiment

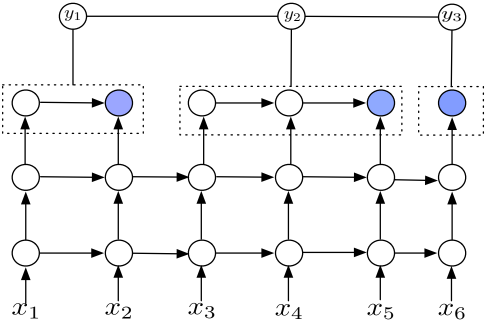
- Comparison to related works
- Results on Eval2000

Table: Attention-Based RNN vs. CTC and HMM-DNN hybrid systems.

System	Output	SWB
HMM-DNN sMBR [Vesely 2013]	-	12.6
CTC no LM [Maas 2015]	char	**38.0**
+7-gram	char	27.8
+RNNLM (3 hidden layers)	char	21.4
Deep Speech [Hannun 2014]	char	20.0
CTC+WFST decoder [Miao 2016]	phone	**14.5**
EncDec no LM	word	26.3
EncDec no LM	char	**27.3**

A new model without attention

- Segmental RNN – Segmental CRF with encoder RNN



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

Thank you ! Questions?