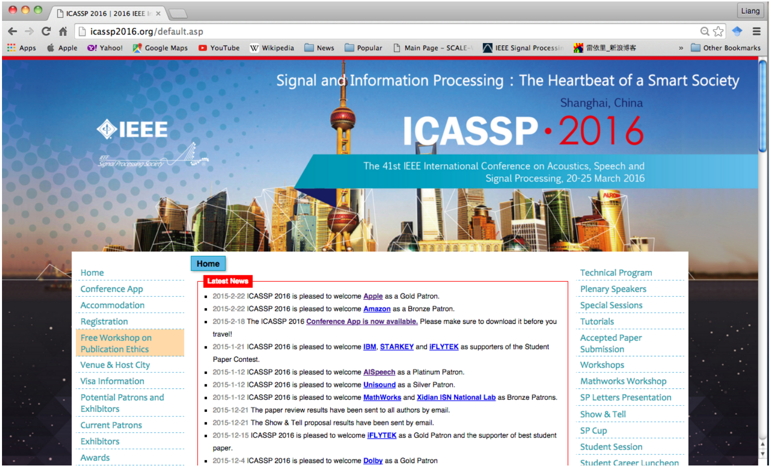
Deep learning for end-to-end speech recognition

Liang Lu

Centre for Speech Technology Research
The University of Edinburgh

11 March 2016



Signal and Information Processing：The Heartbeat of a Smart Society

Shanghai, China

ICASSP · 2016

The 41st IEEE International Conference on Acoustics, Speech and Signal Processing, 20-25 March 2016

- 2015-2-22 ICASSP 2016 is pleased to welcome Apple as a Gold Patron.
- 2015-2-22 ICASSP 2016 is pleased to welcome Amazon as a Bronze Patron.
- 2015-2-18 The ICASSP 2016 Conference App is now available. Please make sure to download it before you travel!
- 2015-1-21 ICASSP 2016 is pleased to welcome IBM, STARKEY and iFLYTEK as supporters of the Student Paper Contest.
- 2015-1-12 ICASSP 2016 is pleased to welcome AISpeech as a Platinum Patron.
- 2015-1-12 ICASSP 2016 is pleased to welcome Unisound as a Silver Patron.
- 2015-1-12 ICASSP 2016 is pleased to welcome MathWorks and Xidian ISN National Lab as Bronze Patrons.
- 2015-12-21 The paper review results have been sent to all authors by email.
- 2015-12-21 The Show & Tell proposal results have been sent by email.
- 2015-12-15 ICASSP 2016 is pleased to welcome iFLYTEK as a Gold Patron and the supporter of best student paper.
- 2015-12-4 ICASSP 2016 is pleased to welcome Dolby as a Gold Patron

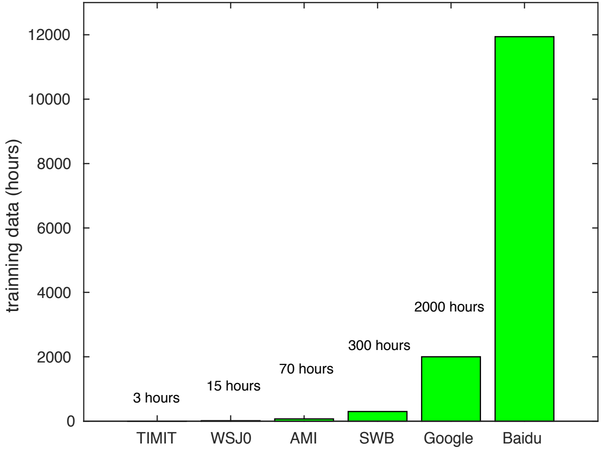
Golden age for speech technology?

Speech technology is around us

ICASSP 2016 Patrons
Mobvoi 出门问问
思必驰 AISPEECH
Microsoft
科大讯飞 iFLYTEK
HUAWEI
DOLBY
云知声 Unisound 知心知音 智慧生活
Google
facebook
iSN STATE KEY LABORATORY
IT FOR YOU
MathWorks
amazon alexa

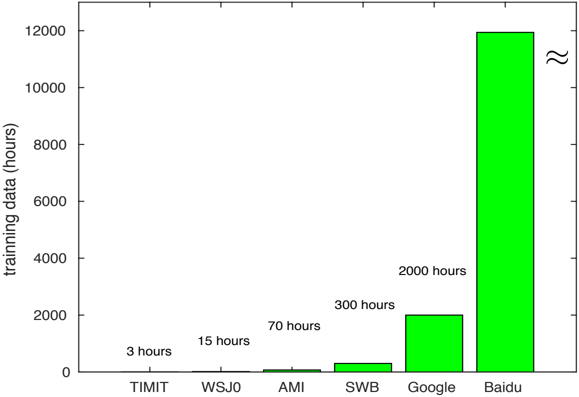
Golden age for speech technology?

Driven by data



Golden age for speech technology?

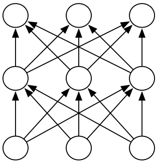
Driven by data



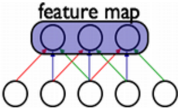
$\approx 5 \times 10^9 (5 \text{ billion})$

Golden age for speech technology?

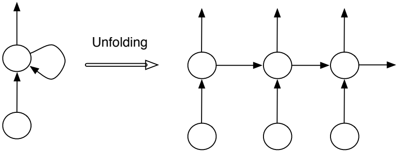
Driven by deep learning



Feed-forward neural network



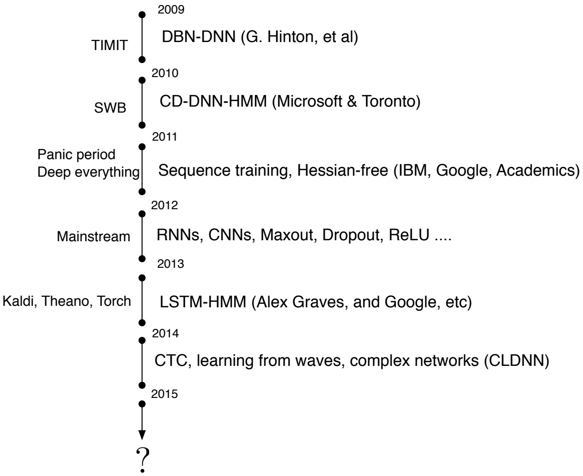
Convolutional neural network



Recurrent neural network

Golden age for speech technology?

Driven by deep learning



2009

TIMIT — DBN-DNN (G. Hinton, et al)

2010

SWB — CD-DNN-HMM (Microsoft & Toronto)

2011

Panic period
Deep everything — Sequence training, Hessian-free (IBM, Google, Academics)

2012

Mainstream — RNNs, CNNs, Maxout, Dropout, ReLU ....

2013

Kaldi, Theano, Torch — LSTM-HMM (Alex Graves, and Google, etc)

2014

CTC, learning from waves, complex networks (CLDNN)

2015

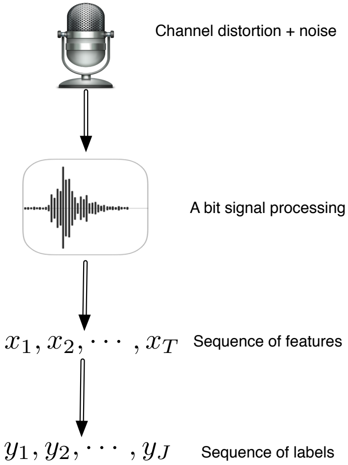
?

But, what is next?

- Open challenges in speech recognition
  - Efficient **adaptation** to speakers, environment, etc
  - **Distant** speech recognition, from close-talk microphone to distant microphone(s)
  - **Small footprint** models, reduce the model size for mobile devices
  - **Open-vocabulary** speech recognition
  - **Low-resource** languages
  - ...
- In this talk, I would like to revisit the fundamental architecture for speech recognition

Speech recognition problem

- Speech recognition is a typical 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
  - ...



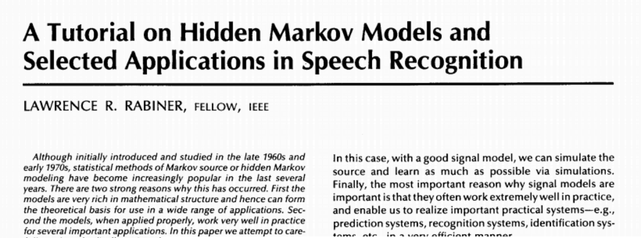
Channel distortion + noise

A bit signal processing

$x_1, x_2, \cdots, x_T$ Sequence of features

$y_1, y_2, \cdots, y_J$ Sequence of labels

Hidden Markov Models



A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition

LAWRENCE R. RABINER, FELLOW, IEEE

Although initially introduced and studied in the late 1960s and early 1970s, statistical methods of Markov source or hidden Markov modeling have become increasingly popular in the last several years. There are two strong reasons why this has occurred. First the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications. Second the models, when applied properly, work very well in practice for several important applications. In this paper we attempt to carefully and methodically review th

of statistical modeling and show

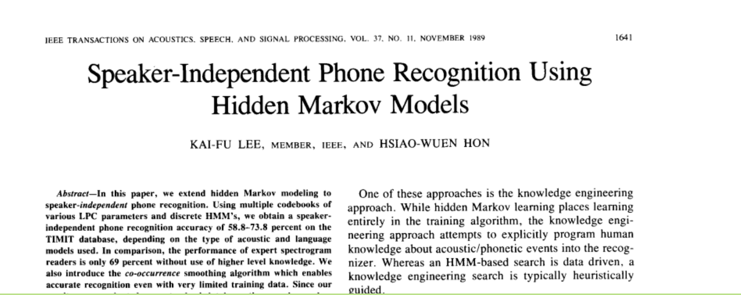
selected problems in machine re

I. Introduction

Real-world processes gene

puts which can be characteri

In this case, with a good signal model, we can simulate the source and learn as much as possible via simulations. Finally, the most important reason why signal models are important is that they often work extremely well in practice, and enable us to realize important practical systems—e.g., prediction systems, recognition systems, identification systems, etc., in a very efficient manner.



IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 11, NOVEMBER 1989 1641

Speaker-Independent Phone Recognition Using Hidden Markov Models

KAI-FU LEE, MEMBER, IEEE, AND HSIAO-WUEN HON

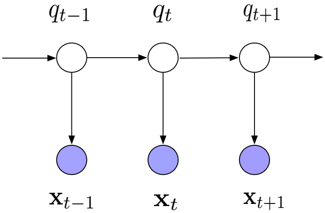
***Abstract*—In this paper, we extend hidden Markov modeling to speaker-*independent* phone recognition. Using multiple codebooks of various LPC parameters and discrete HMM's, we obtain a speaker-independent phone recognition accuracy of 58.8–73.8 percent on the TIMIT database, depending on the type of acoustic and language models used. In comparison, the performance of expert spectrogram readers is only 69 percent without use of higher level knowledge. We also introduce the *co-occurrence* smoothing algorithm which enables accurate recognition even with very limited training data. Since our**

One of these approaches is the knowledge engineering approach. While hidden Markov learning places learning entirely in the training algorithm, the knowledge engineering approach attempts to explicitly program human knowledge about acoustic/phonetic events into the recognizer. Whereas an HMM-based search is data driven, a knowledge engineering search is typically heuristically guided.

Hidden Markov Models

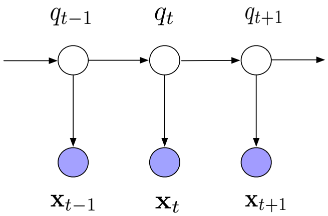
- Why the hidden Markov model works for speech recognition?
- It converts the sequence-level classification problem into a frame-level problem

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



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 → alignment → triphone → alignment → neural net → alignment → 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
- Segmental recurrent neural networks

End-to-end speech recognition

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

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

[1] A. Graves, et al, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks", ICML 2006

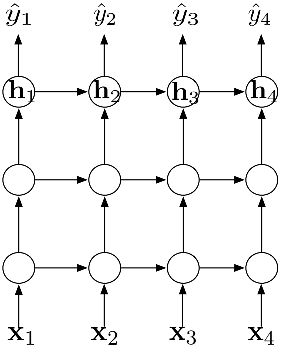
[2] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks", ICML 2014

[3] A. Hannun, et al, "Deep Speech: Scaling up end-to-end speech recognition", arXiv 2014

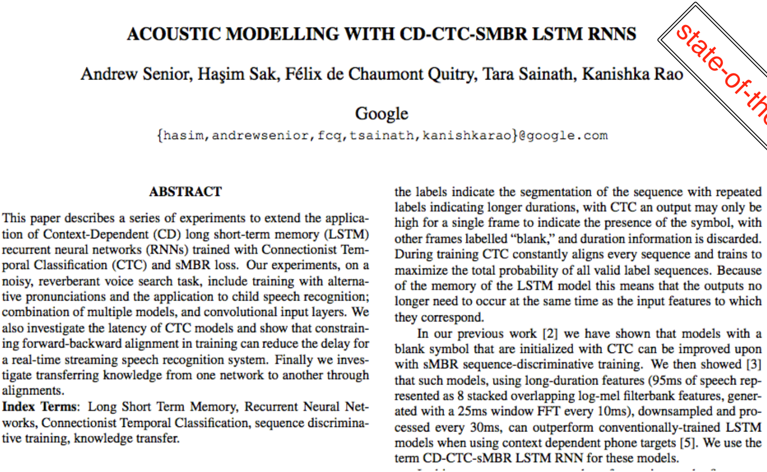
[4] H. Sak, et al, "Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition", INTERSPEECH 2015

End-to-end speech recognition

- Maximum Entropy Markov Model (MEMM)
- Still reply on the **independence** assumption



End-to-end speech recognition



ACOUSTIC MODELLING WITH CD-CTC-SMBR LSTM RNNS

Andrew Senior, Haşim Sak, Félix de Chaumont Quitry, Tara Sainath, Kanishka Rao

Google

{hasim,andrewsenior,fcq,tsainath,kanishkarao}@google.com

ABSTRACT

This paper describes a series of experiments to extend the application of Context-Dependent (CD) long short-term memory (LSTM) recurrent neural networks (RNNs) trained with Connectionist Temporal Classification (CTC) and sMBR loss. Our experiments, on a noisy, reverberant voice search task, include training with alternative pronunciations and the application to child speech recognition; combination of multiple models, and convolutional input layers. We also investigate the latency of CTC models and show that constraining forward-backward alignment in training can reduce the delay for a real-time streaming speech recognition system. Finally we investigate transferring knowledge from one network to another through alignments.

Index Terms: Long Short Term Memory, Recurrent Neural Networks, Connectionist Temporal Classification, sequence discriminative training, knowledge transfer.

the labels indicate the segmentation of the sequence with repeated labels indicating longer durations, with CTC an output may only be high for a single frame to indicate the presence of the symbol, with other frames labelled “blank,” and duration information is discarded. During training CTC constantly aligns every sequence and trains to maximize the total probability of all valid label sequences. Because of the memory of the LSTM model this means that the outputs no longer need to occur at the same time as the input features to which they correspond.

In our previous work [2] we have shown that models with a blank symbol that are initialized with CTC can be improved upon with sMBR sequence-discriminative training. We then showed [3] that such models, using long-duration features (95ms of speech represented as 8 stacked overlapping log-mel filterbank features, generated with a 25ms window FFT every 10ms), downsampled and processed every 30ms, can outperform conventionally-trained LSTM models when using context dependent phone targets [5]. We use the term CD-CTC-sMBR LSTM RNN for these models.

state-of-the-art

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

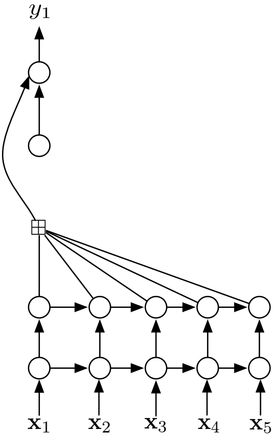
[2] J. Chorowski, et al, "Attention-Based Models for Speech Recognition", NIPS 2015

[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

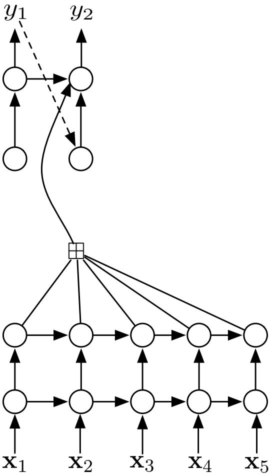
End-to-end speech recognition

- Attention-based RNN encoder-decoder



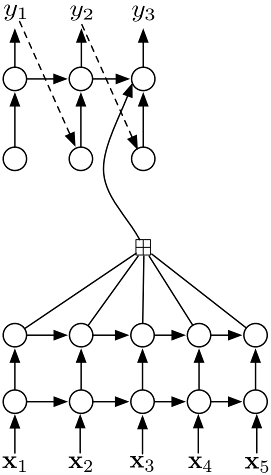
End-to-end speech recognition

- Attention-based RNN encoder-decoder



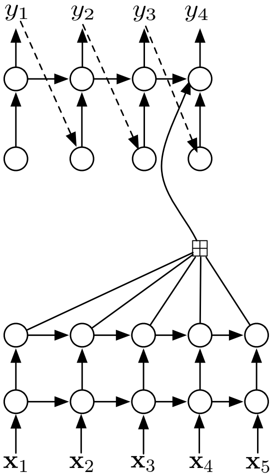
End-to-end speech recognition

- Attention-based RNN encoder-decoder



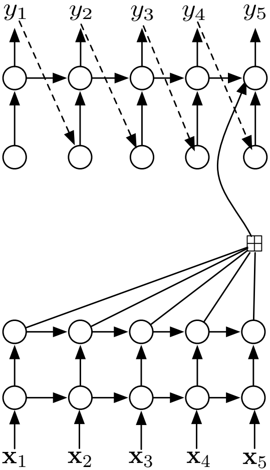
End-to-end speech recognition

- Attention-based RNN encoder-decoder



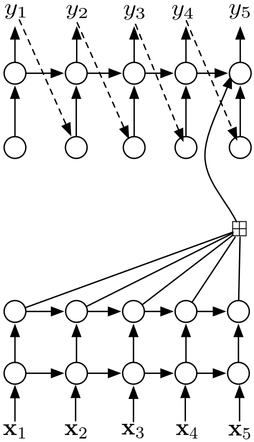
End-to-end speech recognition

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

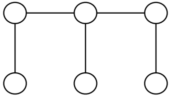
Encoder $\mathbf{h}_{1:T} = \text{RNN}(\mathbf{x}_{1:T})$

End-to-end speech recognition

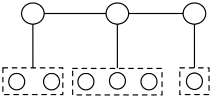
- Attention-based RNN encoder-decoder
  - A flexible sequence-to-sequence transducer
  - "Revolutionising" machine translation
  - Popularising the attention-based scheme
  - But it may not be a natural model for speech recognition, why?

End-to-end speech recognition

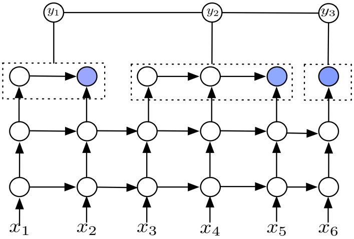
- Segmental recurrent neural network – segmental CRF + RNN



CRF



segmental CRF



segmental RNN

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

[3] Many many more on (segmental) CRFs

Segmental recurrent neural network

- CRF [Lafferty et al. 2001]

$$P(\mathbf{y} \mid \mathbf{X}) = \frac{1}{Z(\mathbf{X})} \prod_j \exp\left(\mathbf{w}^\top \Phi(y_j, \mathbf{X})\right) \quad (5)$$

where the length of **y** and **X** should be equal.

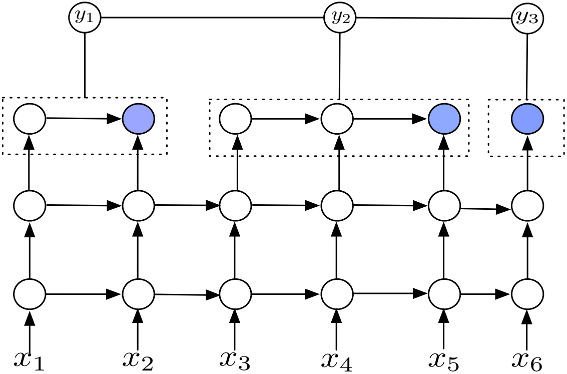
- Segmental (semi-Markov) CRF [Sarawagi and Cohen 2004]

$$P(\mathbf{y}, \mathbf{E}, \mid \mathbf{X}) = \frac{1}{Z(\mathbf{X})} \prod_j \exp\left(\mathbf{w}^\top \Phi(y_j, \mathbf{e}_j, \mathbf{X})\right) \quad (6)$$

where $\mathbf{e}_j = \langle s_j, n_j \rangle$ denotes the beginning (s_j) and end (n_j) time tag of y_j; $\mathbf{E} = \{\mathbf{e}_1, \cdots, \mathbf{e}_J\}$ is the latent segment label.

Segmental recurrent neural network

- Segmental recurrent neural network – using neural networks to learn the feature function $\Phi(\cdot)$.



Segmental recurrent neural network

- Training criteria
  - Conditional maximum likelihood

$$\begin{aligned}\mathcal{L}(\theta) &= \log P(\mathbf{y} \mid \mathbf{X}) \\ &= \log \sum_{\mathbf{E}} P(\mathbf{y}, \mathbf{E} \mid \mathbf{X}) \qquad (7)\end{aligned}$$

  - Max-margin – maximising the distance between the ground truth and negative labels

$$\mathcal{L}(\theta) = \sum_{\hat{\mathbf{y}} \in \Omega} \underbrace{\mathcal{D}_\theta(\mathbf{y}, \tilde{\mathbf{y}})}_{\text{model distance}} \qquad (8)$$

H. Tang, et al, "A comparison of training approaches for discriminative segmental models", INTERSPEECH 2014

Segmental recurrent neural network

- Viterbi decoding
  - Partially Viterbi decoding

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \log \sum_{\mathbf{E}} P(\mathbf{y}, \mathbf{E} \mid \mathbf{X}) \tag{9}$$

  - Fully Viterbi decoding

$$\mathbf{y}^*, \mathbf{E}^* = \arg\max_{\mathbf{y},\mathbf{E}} \log P(\mathbf{y}, \mathbf{E} \mid \mathbf{X}) \tag{10}$$

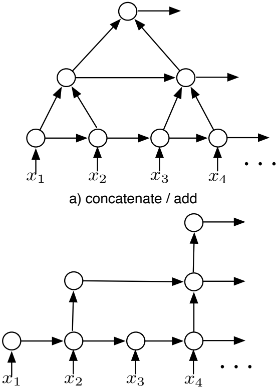
More details: L. Lu, L. Kong, et al, "Segmental Recurrent Neural Networks for End-to-end Speech Recognition", arXiv 2016.

Experiment 1

- 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 1

- Speed up training



a) concatenate / add

b) skip

Experiment 1

Table: Results of hierarchical subsampling networks.

System	LSTM layers	hidden	PER(%)
skip	3	128	21.2
conc	3	128	21.3
add	3	128	23.2
skip	3	250	20.1
conc	3	250	20.5
add	3	250	21.5

Experiment 1

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
×	6	250	21.9

Experiment 1

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 1

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 baseline [Chorowski 2015]	-	×	17.6
Segmental RNN	×	×	18.9
Segmental RNN	×	√	17.3

Experiment 2

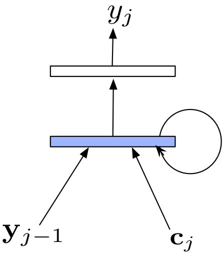
- Switchboard dataset ($\sim$ 300 hours $\approx$ 100 million frames)
- Attention-based RNN systems (EncDec)
- No language model in baseline systems

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

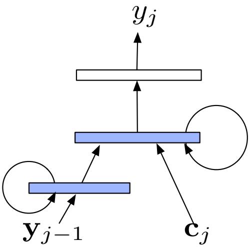
System	Output	Avg
HMM-DNN sMBR [Vesely 2013]	-	18.4
CTC no LM [Maas 2015]	char	**47.1**
+7-gram	char	35.9
+RNNLM (3 hidden layers)	char	30.8
Deep Speech [Hannun 2014]	char	25.9
EncDec no LM	word	36.4
EncDec no LM	char	**37.8**

Experiment 2

- Long memory decoder



a) Baseline decoder



b) LongMem decoder

Experiment 2

Table: Results of language model rescoring and using long memory decoder.

System	Output	Avg
EncDec no LM	word	37.6
+ LongMem	word	36.4
+ 3-gram rescoring	word	**36.0**
EncDec no LM	char	42.8
+ LongMem	char	41.3
+ 5-gram rescoring	char	40.5

L. Lu, et al, "On Training the Recurrent Neural Network Encoder-Decoder for Larger Vocabulary End-to-End Speech Recognition", ICASSP 2016.

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

Summary

- End-to-end speech recognition is a new and exiting research area
- Three new models have been discussed
    - Connectionist Temporal Classification (CTC)
    - Attention-based recurrent neural network
    - Segmental recurrent neural network

Acknowledgement

- Joint work with
  - Xingxing Zhang (Ph.D student at Edinburgh)
  - Lingpeng Kong (Ph.D student at CMU)
- Funed by the NST project

EPSRC
Engineering and Physical Sciences
Research Council

UNIVERSITY OF
CAMBRIDGE

The
University
Of
Sheffield.

Thank you ! Questions?