A study of RNN encoder-decoder for LVCSR

[1]Liang Lu, [1]Xingxing Zhang, [2]Kyunghyun Cho and [1]Steve Renals

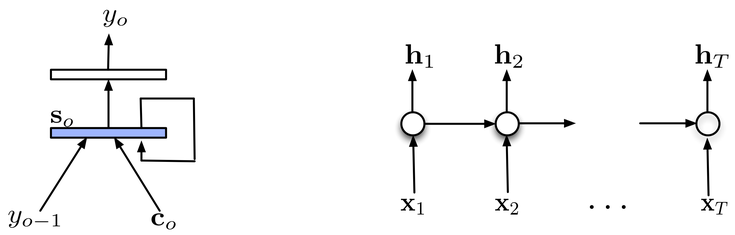
[1]The University of Edinburgh, [2]University of Montreal

Introduction

In this paper, we study using RNN encoder-decoder for end-to-end speech recognition which replaces HMMs.

- Introduction to RNN encoder-decoder
- Model architecture and training details
- Experiments on switchboard dataset

RNN encoder-decoder



1. The objective function - The RNN encoder-decoder is a neural network model that directly computes the conditional probability of the output sequence given the input sequence

$$P(y_1, \ldots, y_O | \mathbf{x}_1, \ldots, \mathbf{x}_T) = \prod_{o=1}^{O} P(y_o | y_1, \ldots, y_{o-1}, \mathbf{c}_o).$$

where $\mathbf{c}_o$ is a fixed length hidden representation of the input sequence obtained from the encoder, to which we refer as *context vector*.

2. The decoder - The posterior probability of y_o is computed as

$$P(y_o | y_1, \ldots, y_{o-1}, \mathbf{c}_o) = g(\mathbf{y}_{o-1}, \mathbf{s}_o, \mathbf{c}_o) \quad (1)$$

where $\mathbf{s}_o$ denotes the output of a recurrent hidden layer, and $g(\cdot)$ is a *softmax* function in the output layer.

3. The encoder - The encoder maps the input sequence into a fixed length of vector representation using, e.g. bi-directional RNN

$$\mathbf{c}_o = \sum_t \alpha_{ot} \mathbf{h}_t \quad (2)$$

where $\alpha_{ot} \in [0, 1]$ and $\sum_t \alpha_{ot} = 1$. The weights $\alpha_o t$ is obtained by the alignment model

$$\alpha_{ot} = \frac{\exp(e_{ot})}{\sum_{t'} \exp(e_{ot'})}, \quad e_{ot} = a(\mathbf{s}_{o-1}, \mathbf{h}_t) \quad (3)$$

where $a(\cdot)$ is a neural network (with one or two hidden layers).

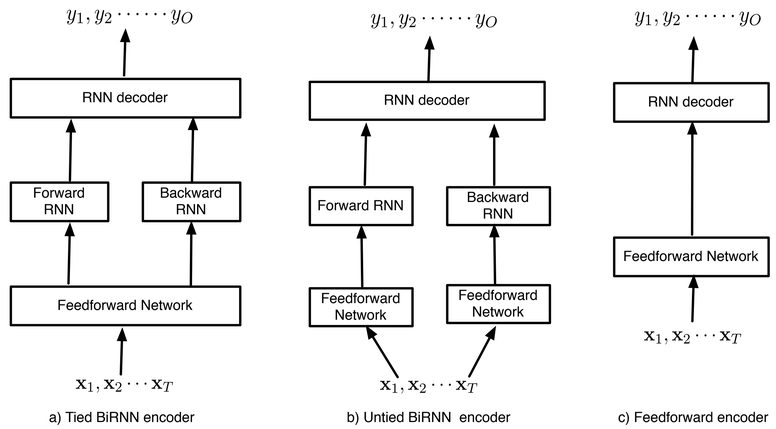
Model architecture and training

1. Model training - The model can be trained by maximising the average conditional log-likelihood over all the training set as

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \frac{1}{N} \sum_{n=1}^{N} \log P(y_1^n, \ldots, y_O^n | \mathbf{x}_1^n, \ldots, \mathbf{x}_T^n, \mathcal{M})$$

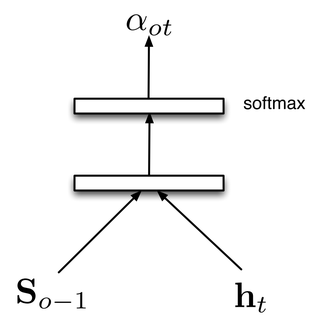
where $\mathcal{M}$ denotes the set of model parameters, and N is the number of training utterances. The model was trained using stochastic gradient decent (SGD) with the Adadelta algorithm to automatically estimate the learning rates.

2. Types of encoders



3. Downsampling - Modelling such long sequences is challenging to RNNs even with gated recurrent units, especially due to high computational requirement. We explore an approach of downsampling the input sequence at the low-level acoustic features and at the level of encoder representations. It can give 2 - 3 times speedup.

4. Deep alignment model



Experiments

1. System configuration

- we trained the model using 300hrs Switchboard data
- 1000 hidden units and ∼30k vocabulary size
- It took ∼ 15 epochs to converge and ∼ 4 - 5 days to train with downsampling

2. Results of using different encoders

Encoder	#layers	CHM	SWB	Avg
Tied BiRNN	1	70.5	55.0	62.8
Tied BiRNN	2	67.3	51.3	59.3
Tied BiRNN	3	68.1	54.0	61.1
Untied BiRNN	1	61.3	40.8	51.1
Untied BiRNN	2	60.5	41.2	50.9
Untied BiRNN	3	67.7	46.2	57.0
FeedForward	1	93.2	86.5	89.9

3. Results of downsampling

Step	Splicing	Space	CHM	SWB	Avg
1	±5	feature	62.7	47.6	55.2
2	±5	feature	61.3	40.8	51.1
3	±5	feature	59.9	**38.8**	**49.4**
4	±5	feature	60.2	41.7	51.0
2	±5	hidden	60.7	42.3	51.5
3	±5	hidden	**58.9**	41.7	50.3

4. Results of using deep alignment model

#layers	CHM	SWB	Avg
1	59.9	38.8	49.4
2	60.6	40.8	50.8

Conclusion - In this paper, we study the application of an RNN encoder-decoder model for large vocabulary end-to-end speech recognition. Without using any language model or pronunciation lexicon, we have obtained encouraging recognition accuracy on the Switchboard corpus. Our experiments show that using RNNs in the encoder is essential for the success of this model architecture, and using separate feedforward neural networks for feature extraction in the encoder can reduce the word error rate.

Acknowledgement The research was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology)