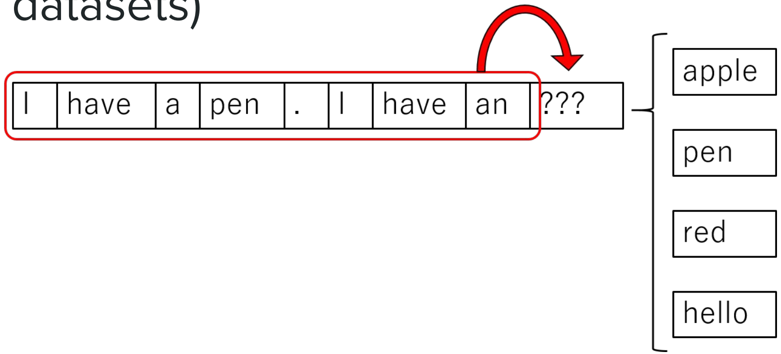
A Review on Neural Language Modeling

Anuar Maratkhan
School of Science and Technology
Nazarbayev University

Introduction

- Language Model (LM) is an algorithm for predicting next words in a text document
- It is central task of Natural Language Processing (NLP) and used in speech recognition, machine translation, optical character recognition like tasks
- LM performance is measured with *perplexity* (lower is better)
- Neural network based language models outperform standard backoff models (which need larger datasets)



Recurrent Neural Network (RNN)

- RNN is an enhancement of feed-forward neural network
- RNN passes previous hidden layer result (context) to the current input layer
- RNN have dynamic (unlimited) size of context in contrast to the fixed-size in feed-forward neural network
- Backpropagation is main technique for optimizing RNN and other neural networks

Simple RNN

$$x(t) = w(t) + s(t-1) \tag{1}$$

$$s_j(t) = f\left(\sum_i x_i(t) u_j i\right) \tag{2}$$

$$y_k(t) = g\left(\sum_j s_j(t) v_k j\right) \tag{3}$$

where x(t), s(t), y(t), w(t) - input, hidden layer, output, word at time t respectively,
s(t-1) - previous hidden layer, u and v - weights (coefficients),
f(z) - sigmoid activation function, g(z) - softmax function

Simple RNN

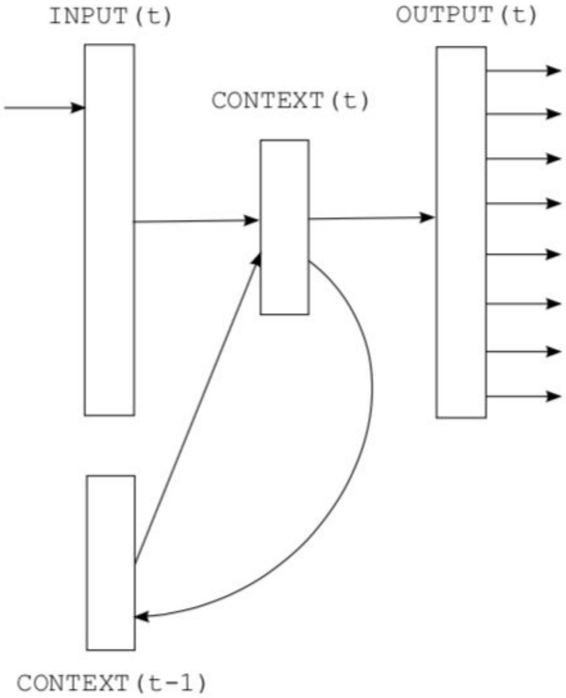


Figure 1. Simple recurrent neural network from (T. Mikolov et al. 2010)

Recurrent Neural Network based Language Model

- T. Mikolov et al. (2010) were first to use recurrent neural networks for language modeling tasks
- RNN based language models outperform feed-forward neural network based language models in terms of perplexity
- Backpropagation through time (BPTT) enhances the performance of RNN based language models (T. Mikolov et al. 2011)
- W.Zaremba et al. (2014) tried Long-Short Term Memory (LSTM) recurrent neural network for language modeling that showed better results

Recurrent Neural Network based Language Model

Model	Param	Validation	Test
Mikolov & Zweig (2012) – RNN-LDA + KN-5 + cache	9M	-	92.0
Zaremba et al. (2014) – LSTM	20M	86.2	82.7
Gal & Ghahramani (2016) – Variational LSTM (MC)	20M	-	78.6
Kim et al. (2016) – CharCNN	19M	-	78.9
Merity et al. (2016) – Pointer Sentinel-LSTM	21M	75.7	70.9
Grave et al. (2016) – LSTM + continuous cache pointer	-	-	72.1
Inan et al. (2016) – Tied Variational LSTM + augmenter loss	24M	75.7	73.2
Zilly et al. (2016) – Variational RHN	23M	67.9	65.4
Zoph & Le (2016) – NAS Cell	25M	-	64.0
Melis et al. (2017) – 2-layer skip connection LSTM	24M	60.9	58.3
Merity et al. (2017) – AWD-LSTM w/o finetune	24M	60.7	58.8
Merity et al. (2017) – AWD-LSTM	24M	60.0	57.3
Salakhutdinov et al. [5] – AWD-LSTM-MoS w/o finetune	22M	58.08	55.97
Salakhutdinov et al. [5] – AWD-LSTM-MoS	22M	**56.54**	**54.4**
Merity et al. (2017) – AWD-LSTM + continous cache pointer	24M	53.9	52.8
Krause et al. (2017) – AWD-LSTM + dynamic evaluation	24M	51.6	51.1
Salakhutdinov et al. [5] – AWD-LSTM-MoS + dynamic evaluation	22M	**48.33**	**47.69**

Table 1. Single model perplexity on validation and test sets on Penn Treebank (Z. Yang et al. 2017)

Subword-level Language Model

- Vocabulary in language modeling represents a number of unique words in the document
- Word-level language models are unable to deal with new words, so called Out-of-Vocabulary (OOV) words
- Character-level neural language models are not efficient relative to word-level neural language models (T. Mikolov et al. 2011)
- Subword-level language models combine word-level and character-level language models (example below)

```
new company dreamworks interactive
new company dre+ am+ wo+ rks:  in+ te+ ra+ cti+ ve:
```

Reinforcement Learning based Language Modeling

- B. Zaph and Q. V. Le (2017) present neural architecture search for language modeling using reinforcement learning (Figure 2)
- RNN controller tries different neural network architecture for LM, and then gets rewarded according to the performance of that neural network architecture

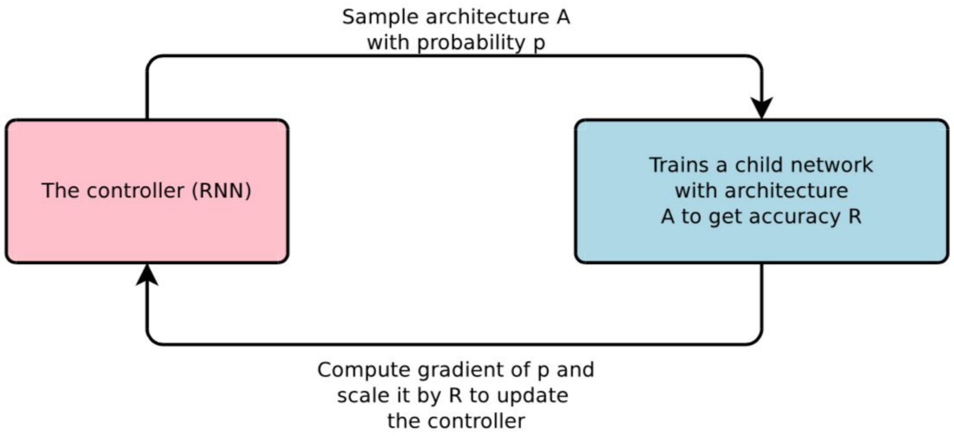
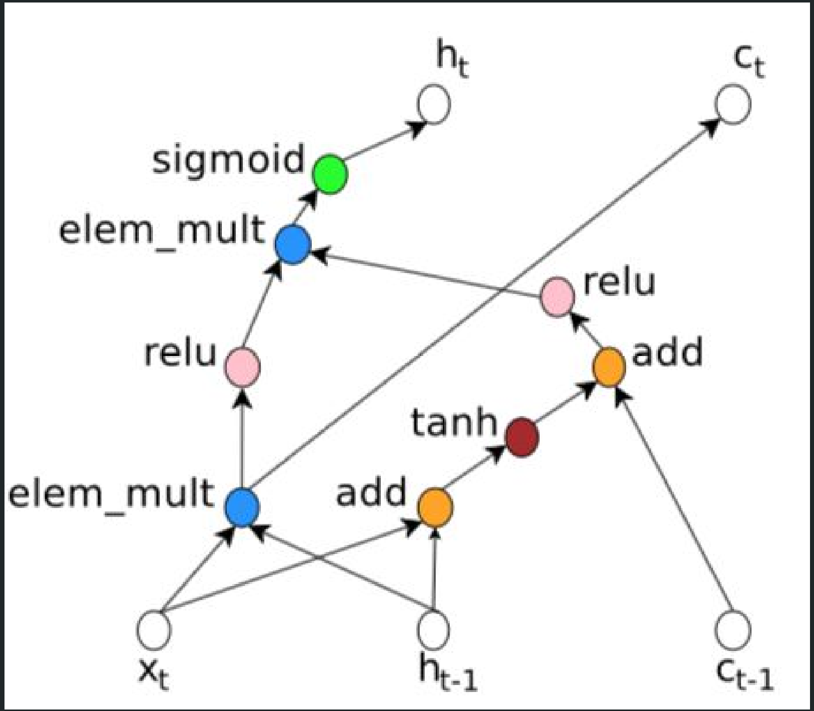


Figure 2. An overview of neural architecture search (B. Zaph and Q. V. Le, 2017)

Reinforcement Learning based Language Modeling

RNN controller tries neural network architecture from search space (figure 3) and gets rewarded according to the performance of neural network

Figure 3. Neural architecture search space (B. Zaph and Q. V. Le, 2017)



Conclusion and Future Suggestions

- Neural network based language models outperforms standard statistical (backoff) language models in terms of perplexity and dataset size
- RNN based language models trained with BPTT are simple and intelligent models for language modeling tasks
- Subword-level neural language model combines advantages of word-level and character-level neural language models
- Reinforcement learning can be used to train RNN controller to find the best neural architecture for language modeling

References

[1] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur. “Recurrent Neural network based language model”. 2010

[2] Q. V. Le, M. A. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Deau, A. Y. Ng. “Building high-level feature using large scale unsupervised learning”. 2012

[3] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, S. Khudanpur. “Extensions of recurrent neural network language model”. 2011

[4] W. Zaremba, I. Sutskever, and O. Vinyals. “Recurrent neural network regularization”. 2014

References

[5] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen. “Breaking the softmax bottleneck: A high-rank RNN language model”. 2017

[6] T. Mikolov, I. Sutskever, A. Deoras, H.-S. Le, S. Kombrink, and J. Cernocky. “Subword language modeling with neural networks”. 2011

[7] B. Zoph and Q. V. Le. “Neural architecture search with reinforcement learning”. 2017

[8] Y. LeCun, Y. Bengio, and G. Hinton “Deep Learning”. 2015

[9] A. Karpathy, J. Johnson, and L. Fei-Fei. “Visualizing and understanding recurrent networks”. 2016

Thank you for your attention!