LSTM, GRU, Highway and a Bit of Attention: An Empirical Overview for Language Modeling in Speech Recognition

Kazuki Irie, Zoltán Tüske, Tamer Alkhouli, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition
RWTH Aachen University

Introduction

Motivation

- Language models based on LSTM-RNN achieve state of the art performance [Sundermeyer et al. Interspeech 2012]
- Innovation of LSTM [Hochreiter and Schmidhuber 1997]: gating mechanism organized around the memory cell
- Trend in designing ANNs with intentionally organized information flow
  - Networks with **multiplicative gates** (Highway, Gated recurrent unit)
  - **Attention** mechanism provides both increase in performance and visualization of networks' decisions

Questions addressed in this work:

- How do different gated architectures compare for language modeling in terms of PPL and WER?
- Can we find a simple application of the attention mechanism for language modeling?

Experimental Setups

Task: Quaero English broadcast news and conversation speech recognition

Language modeling

- Vocabulary: **150 k**
- Training text:
  - **3.1 B** for baseline **4-gram count model** with Kneser-Ney smoothing
  - **50 M** subset for all **neural language models**
    Further fine-tuning on a 2 M most in-domain subset
- 1000 word classes are trained by the exchange algorithm and used to factorize the output layer of all neural LMs
- Dev 40 k, Eval 36 k
- All models are implemented within `rwthlm`

Acoustic modeling

- A hybrid 12-layer rectified linear unit based feedforward network
- Multilingually initialized on 4 languages
- MPE sequence-level discriminative training

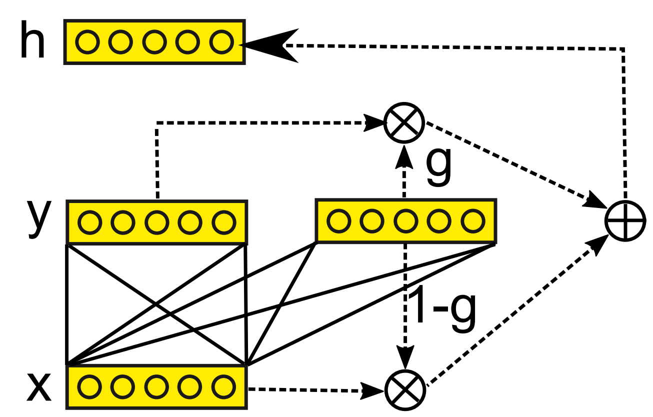
Neural networks with multiplicative gates

Highway connections in feedforward networks (FFNN)

[Srivastava et al. NIPS 2015, ICML 2015]
Input x, Output h:

$$
\begin{aligned}
y &= \sigma(W_y x + b_y) \\
g &= \sigma(W_g x + b_g) \\
h &= g \odot y + (1 - g) \odot x
\end{aligned}
$$

W_y and W_g are weight matrices, b_y and b_g are biases



- Extending the FFNN with a gated linear connection across layers.
- Allows unobstructed information flow through the network
- Interpolation between **transformed** and **untransformed** features
- Originally designed to train very deep networks: more than 900 layers
- Improvements even with shallow configurations
  - for language modeling (from 1 layer): [Kim et al, AAAI 2016]
  - for acoustic modeling (from 3 layers): [Zhang et al, ICASSP 2016]

Highway connections in feedforward networks

Perplexity results

- 20-gram feedforward models with 600 nodes per layer
- Perplexities on the development text

	Number of Layers			
Topology	2	3	4	5
Baseline FFNN	126.4	124.9	124.6	126.7
Sigmoid-Highway	126.5	120.4	119.8	119.7

Experimental results show

- PPL improvements from the baseline 4-layer (124.6) to the 5-layer Highway (119.7)

Neural networks with multiplicative gates

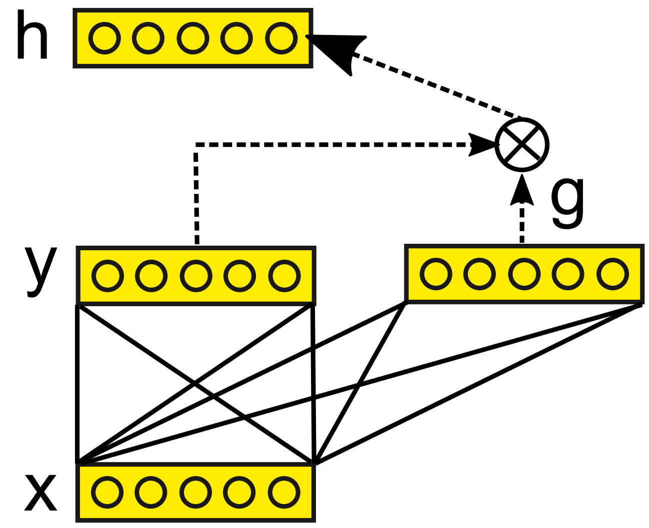
Lateral/Tensor networks

[Yu et al. 2013, Devlin et al. EMNLP 2015]

Input x, Output h:

$$
\begin{aligned}
y &= \sigma(W_y x + b_y) \\
g &= \sigma(W_g x + b_g) \\
h &= g \odot y
\end{aligned}
$$

W_y and W_g are weight matrices, b_y and b_g are biases



- Minimalistic gating mechanism.
- Can be seen as variant of maxout network
  (max operation instead of multiplication, 2 populations)
- "Highway without highway connection"

Perplexity results

- 20-gram feedforward models with 600 nodes per layer
- Perplexities on the development text

	Number of Layers		
Topology	2	3	4
Baseline FFNN	126.4	124.9	124.6
Lateral	123.4	122.0	122.2

Observations:

- PPL improvements from the baseline 4-layer (124.6) to 3-layer Lateral (122.0)
- Worse than 5-layer Highway (119.7)
- Illustrates the effect of linear connection $(1 - g) \odot x$

LSTM vs. GRU

- On Treebank LSTM outperforms GRU [Jozefowicz et al. ICML 2015]
- LSTM PPL from [Sundermeyer et al. 2015]

		LSTM		GRU	
		depth			
		1	2	1	2
size	100	147.0	139.6	143.9	136.4
	200	127.7	117.7	121.9	116.8
	300	117.6	109.1	115.7	110.7
	400	112.8	104.6	114.7	110.0
	500	109.2	101.8	112.6	108.1
	600	107.8	100.5	112.2	108.9

- GRU performs similar to LSTM for small model size
- LSTM gives better PPL

LSTM vs. GRU

After fine-tuning

- Further fine-tuning on 2 M in-domain data

Fine-tuning	LSTM	GRU
no	100.5	108.1
yes	98.3	104.7

- GRU performs about 7% worse than the LSTM

Neural networks with multiplicative gates

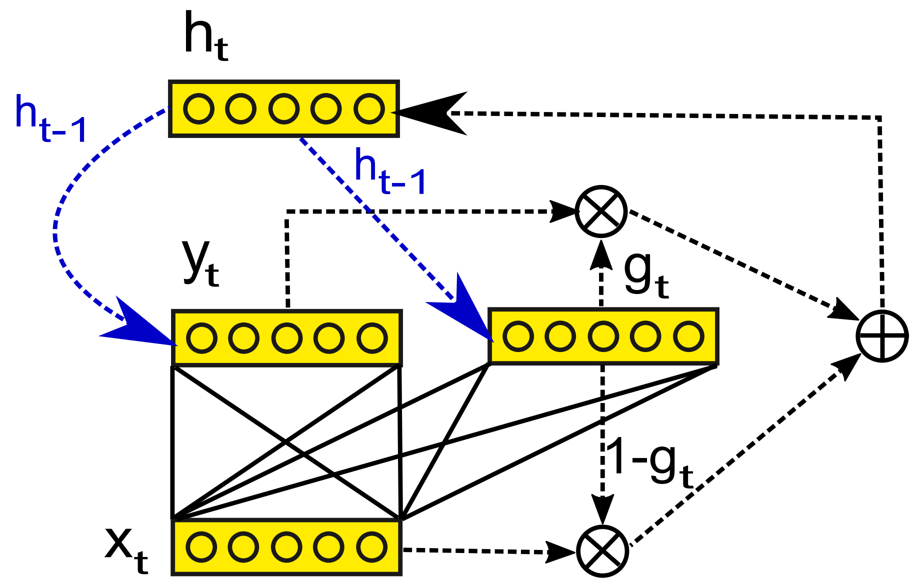
Highway connections in RNNs

- Motivations of highway connection is not limited to the MLP $\Rightarrow$ also applies to deep RNNs
- Extension specific to the LSTM has been proposed [Zhang et al. ICASSP 2016]
- More generic approach: replace the transformation in highway network by a recurrent transformation (can be LSTM or GRU)

Input x_t, Output h_t:

$$\begin{aligned} y_t &= \mathrm{GRU}(x_t, h_{t-1}) \\ g_t &= \sigma(W_g x_t + R_g h_{t-1} + b_g) \\ h_t &= g_t \odot y_t + (1 - g_t) \odot x_t \end{aligned}$$

W_g and R_g are weight matrices, b_g is a bias



Highway connections in RNNs

Perplexity results

Size	Topology	Fine-tuning	Number of layers 2	3	4
300	GRU	no	110.7	114.5	116.4
	GRU-Highway	no	109.1	106.3	106.6
		yes	105.5	102.9	103.3
500	GRU-Highway	yes	101.5	100.3	99.1

- Similar improvements as for feedforward models
- Highway connections allow to benefit from the depth
- Overall improvement from 104.7 to 99.1 (about 5% rel.)

Overall ASR results

Lattice rescoring results with neural models interpolated with KN4

Language model	Topology (NxL)	DEV PPL	DEV WER[%]	EVAL PPL	EVAL WER[%]
4-gram KN	-	132.7	12.3	131.2	10.5
Baseline FFNN	600x3	106.1	11.3	106.0	9.5
Sigm-Highway	600x5	103.9	11.2	103.1	9.5
Lateral	600x3	104.8	11.3	104.5	9.7
LSTM	600x2	89.8	10.7	90.5	9.0
GRU	500x2	93.0	10.8	94.2	9.4
GRU-Highway	500x4	90.7	10.6	91.4	9.2

Observations:

- For feedfoward models, gains from gating mechanism are not significant
- Confirms the effectiveness of LSTM for language modeling
- Improvements from the highway connection and the depth for RNN

Can we make use of an attention mechanism for language modeling?

Motivation

- Are all predecessor words equally important for this prediction?

Thanks for taking the time to download this BBC radio five live $\boxed{\textbf{podcast}}$

- An application for language modeling would be to make **word triggers** explicit
- Initial experiments by considering a minimalistic recurrent attention layer
  Word/Context vectors (outputs of the predecessor layer): $x_1, \ldots, x_t$

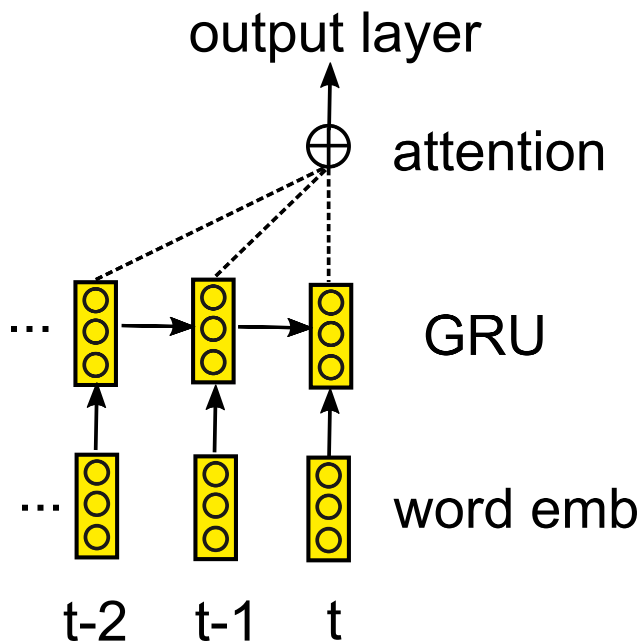
$$\begin{aligned} \forall i \in \{1, .., t\} \quad s_{t,i} &= w^{\intercal} \tanh(Wx_i + Rh_{t-1} + b) \\ \alpha_t &= \text{softmax}(s_t) \\ h_t &= \sum_{i=1}^{t} \alpha_{t,i} x_i \end{aligned}$$

W and R are weight matrices, w is a vector weight and b is a bias

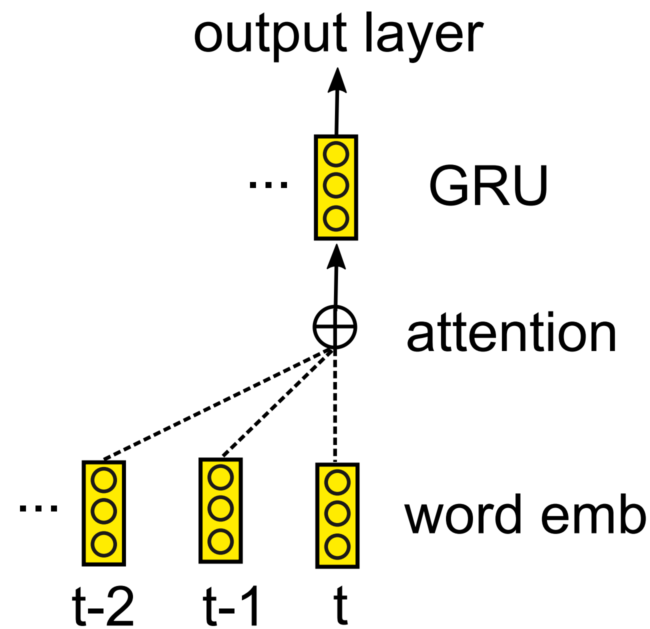
- Insert this in an RNN LM

Attention layer inside the RNN LM to learn word triggers

- Baseline GRU (WordEmb + GRU + Output) of PPL = 110.6
- Two possibilities considered to insert an attention layer



- WordEmb + **GRU + Attention**: PPL = 109.1
- No trigger is obtained, the model chooses the most recent context from the GRU



- WordEmb + **Attention + GRU**: PPL = 157.6 vs. KN4 (163.0)
- Trigger distribution can be observed

Attention layer inside the RNN LM to learn word triggers

Examples

- The numbers in the exponent to words show the weight (in %) of the word to predict the word in the box
- Top triggers are highlighted

\$^{6} **Thanks**^{10} for^{3} taking^{9} the^{2} time^{4} to^{3} **download**^{22} this^{5} **BBC**^{12} **radio**^{11} five^{4} live^{8} $\boxed{\textit{podcast}}$

\$^{22} In^{4} this^{7} **book**^{17} there^{7} are^{5} **things**^{13} that^{7} are^{5} **very**^{14} $\boxed{\textit{complicated}}$

- Qualitatively meaningful triggers could be observed
- Further investigation is necessary to improve the PPL
- Better architecture proposed in [Tran et al. NACCL 2016] with recurrent memory networks

Conclusion

Highway connections

- Help models to benefit from the depth
- Highway connection part is important (comparison to the lateral network)
- Can be also used in RNNs in a simple manner
- Slight improvements in PPL and WER could be obtained

LSTM vs GRU

- LSTM is a good default choice for language modeling

Finding word triggers from attention

- Difficult to get a good PPL from a simple approach
- Results limited to some qualitative observations
- More sophistication is necessary to get better PPL

Thank you for your attention

This work was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract no. W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

References

- M. Sundermeyer, R. Schlüter, and H. Ney, LSTM neural networks for language modeling. in Proc. Interspeech, Portland, OR, USA, Sep. 2012, pp. 194–197.
- M. Sundermeyer, H. Ney, and R. Schlüter, From feedforward to recurrent LSTM neural networks for language modeling, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 3, pp. 517–529, Mar. 2015.
- Y. Kim and Y. J. D. S. A. Rush, Character-aware neural language models, in Proc. AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, Feb. 2016.
- D. Yu, L. Deng, and F. Seide, The deep tensor neural network with applications to large vocabulary speech recognition, IEEE Trans. on Audio, Speech, and Language Processing, vol. 21, no. 2, pp. 388–396, 2013.
- J. Devlin, C. Quirk, and A. Menezes, Pre-computable multi-layer neural network language models, in Proc. of Conf. on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal, Sep. 2015, pp. 256–260.
- R. K. Srivastava, K. Greff, and J. Schmidhuber, Training very deep networks, in Advances in Neural Information Processing Systems (NIPS), Montreal, Canada, Dec. 2015, pp. 2368–2376.
- R. K. Srivastava, K. Greff, and J. Schmidhuber, Highway networks, in the Deep Learning workshop at Int. Conf. on Machine Learning (ICML), Lille, France, Jul. 2015.

References

- Y. Zhang, G. Chen, D. Yu, K. Yao, S. Khudanpur, and J. Glass, Highway long short-term memory RNNs for distant speech recognition, Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2016.
- S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
- R. Jozefowicz, W. Zaremba, and I. Sutskever, An empirical exploration of recurrent network architectures, in Proc. of Int. Conf. on Machine Learning (ICML), Lille, France, Jul. 2015, pp. 2342–2350.
- K. Cho, B. v. M. C. Gulcehre, D. Bahdanau, F. B. H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in Proc. of Conf. on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Oct. 2014, pp. 1724–1734.
- C. Tillmann and H. Ney, Word triggers and the EM algorithm, in Proc. Special Interest Group Workshop on Computational Natural Language Learning (ACL), Madrid, Spain, Jul. 1997, pp. 117–124.
- K. Tran, A. Bisazza, and C. Monz, Recurrent memory network for language modeling, in Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), San Diego, CA, USA, Jun. 2016.