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 $h$:

$$y = \sigma(W_y x + b_y)$$
$$g = \sigma(W_g x + b_g)$$
$$h = g \odot y + (1 - g) \odot x$$

$W_y$ 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

<table>
<thead>
<tr>
<th>Topology</th>
<th>Number of Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Baseline FFNN</td>
<td>126.4</td>
</tr>
<tr>
<td>Sigmoid-Highway</td>
<td>126.5</td>
</tr>
</tbody>
</table>

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 $h$:

\[
\begin{align*}
y &= \sigma(W_y x + b_y) \\
g &= \sigma(W_g x + b_g) \\
h &= g \odot y
\end{align*}
\]

$W_y$ 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"
Lateral/Tensor networks

Perplexity results

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

<table>
<thead>
<tr>
<th>Topology</th>
<th>Number of Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline FFNN</td>
<td>126.4 124.9 124.6</td>
</tr>
<tr>
<td>Lateral</td>
<td>123.4 122.0 122.2</td>
</tr>
</tbody>
</table>

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 \((1 - g) \odot x\)
Neural networks with multiplicative gates

LSTM vs. GRU

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

<table>
<thead>
<tr>
<th>size</th>
<th>LSTM</th>
<th>GRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>147.0</td>
<td>139.6</td>
</tr>
<tr>
<td>200</td>
<td>127.7</td>
<td>117.7</td>
</tr>
<tr>
<td>300</td>
<td>117.6</td>
<td>109.1</td>
</tr>
<tr>
<td>400</td>
<td>112.8</td>
<td>104.6</td>
</tr>
<tr>
<td>500</td>
<td>109.2</td>
<td>101.8</td>
</tr>
<tr>
<td>600</td>
<td>107.8</td>
<td>100.5</td>
</tr>
</tbody>
</table>

• 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

<table>
<thead>
<tr>
<th>Fine-tuning</th>
<th>LSTM</th>
<th>GRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>100.5</td>
<td>108.1</td>
</tr>
<tr>
<td>yes</td>
<td>98.3</td>
<td>104.7</td>
</tr>
</tbody>
</table>

- 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 ⇒ 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 $h_t$:

$$y_t = \text{GRU}(x_t, h_{t-1})$$
$$g_t = \sigma(W_g x_t + R_g h_{t-1} + b_g)$$
$$h_t = g_t \odot y_t + (1 - g_t) \odot x_t$$

$W_g$ and $R_g$ are weight matrices, $b_g$ is a bias
### Highway connections in RNNs

#### Perplexity results

<table>
<thead>
<tr>
<th>Size</th>
<th>Topology</th>
<th>Fine-tuning</th>
<th>Number of layers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>300</td>
<td>GRU</td>
<td>no</td>
<td>110.7</td>
</tr>
<tr>
<td></td>
<td>GRU-Highway</td>
<td>yes</td>
<td>105.5</td>
</tr>
<tr>
<td>500</td>
<td>GRU-Highway</td>
<td>yes</td>
<td>101.5</td>
</tr>
</tbody>
</table>

- 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

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

Observations:
- For feedforward 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 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): $x_1, \ldots, x_t$

$$ \forall i \in \{1, \ldots, t\} \quad s_{t,i} = w^T \tanh(W x_i + R h_{t-1} + b) $$
$$ \alpha_t = \text{softmax}(s_t) $$
$$ h_t = \sum_{i=1}^{t} \alpha_{t,i} x_i $$

• Insert this in an RNN LM

$W$ and $R$ are weight matrices, $w$ is a vector weight and $b$ is a bias
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$^8$ podcast

$^{22}$ In$^4$ this$^7$ book$^{17}$ there$^7$ are$^5$ things$^{13}$ that$^7$ are$^5$ very$^{14}$ 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

References