Improving Neural Language Modeling via Adversarial Training

Dilin Wang*, Chengyue Gong* (equal contribution) Qiang Liu

Department of Computer Science
The University of Texas at Austin
Neural Language Modeling

- Example: the clouds are in the **sky**

\[
\begin{align*}
    h_t &= f_{NN}(x_{t-1}, h_{1:t-1}; \theta) \\
    p(x_t \mid x_{1:t-1}; \theta, w) &= \text{Softmax}(x_t, h_t; w) \\
    &= \frac{\exp(w_{x_t}^\top h_t)}{\sum_{\ell=1}^{|\mathcal{V}|} \exp(w_{\ell}^\top h_t)}
\end{align*}
\]

- Maximum log-likelihood estimation (MLE):

\[
\max_{\theta, w} \sum_t \log p(x_t \mid x_{1:t-1}; \theta, w)
\]
Overfitting

Existing overfitting preventing methods:
- Dropout [e.g., Gal & Ghahramani, 2016]
- Optimizer [e.g., Merity et al., 2017]
- Other: weight tying [Press & Wolf, 2016; Inan et al., 2017]; activation regulariza­tion [Merity et al., 2017], etc.
Adversarial MLE

- Idea: inject an adversarial perturbation on the word embedding vectors in the Softmax layer, and maximize the worst-case performance,

\[
\max_{\theta, w} \min_{\delta_t} \sum_t \log \left( \frac{\exp((w_t + \delta_t)^T h_t)}{\exp((w_t + \delta_t)^T h_t) + \sum_{j \neq t} \exp(w_j^T h_t)} \right)
\]

s.t. \( \|\delta_t\| \leq \epsilon \).

A closed-form solution

\[
\delta_t^* = \arg \min_{\|\delta_t\| \leq \epsilon} (w_t + \delta_t)^T h_t = -\epsilon \frac{h_t}{\|h_t\|}.
\]
Adversarial MLE

Idea: inject an adversarial perturbation on the word embedding vectors in the Softmax layer, and maximize the worst-case performance,

$$\max_{\theta, w} \min_{\delta_t} \sum_t \log \left( \frac{\exp((w_t + \delta_t)^\top h_t)}{\exp((w_t + \delta_t)^\top h_t) + \sum_{j \neq t} \exp(w_j^\top h_t)} \right)$$

$s.t \quad ||\delta_t|| \leq \epsilon.$

A closed-form solution

$$\delta_t^* = \arg \min_{||\delta_t|| \leq \epsilon} (w_t + \delta_t)^\top h_t = -\epsilon \frac{h_t}{||h_t||}.$$
Adversarial MLE Promotes Diversity

- If $w_i$ dominates all the other words under $\epsilon$-adversarial perturbation, in that

$$\min_{||\delta_i|| \leq \epsilon} (w_i + \delta_i)^\top h = (w_i^\top h - \epsilon ||h||)$$

$$> w_j^\top h, \quad \forall j \neq i,$$

then we have,

$$\min_{j \neq i} ||w_j - w_i|| > \epsilon,$$

that is, $w_i$ is separated from the embedding vectors of all other words by at least $\epsilon$ distance.
# Improving on Language Modeling

<table>
<thead>
<tr>
<th>Method</th>
<th>Params</th>
<th>Valid</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWD-LSTM</td>
<td>24M</td>
<td>51.60</td>
<td>51.10</td>
</tr>
<tr>
<td><strong>AWD-LSTM + Ours</strong></td>
<td>24M</td>
<td><strong>49.31</strong></td>
<td><strong>48.72</strong></td>
</tr>
<tr>
<td>AWD-LSTM + MoS</td>
<td>22M</td>
<td>48.33</td>
<td>47.69</td>
</tr>
<tr>
<td><strong>AWD-LSTM + MoS + Ours</strong></td>
<td>22M</td>
<td><strong>47.15</strong></td>
<td><strong>46.52</strong></td>
</tr>
</tbody>
</table>

**Table: PTB**

<table>
<thead>
<tr>
<th>Method</th>
<th>Params</th>
<th>Valid</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWD-LSTM</td>
<td>33M</td>
<td>46.40</td>
<td>44.30</td>
</tr>
<tr>
<td><strong>AWD-LSTM + Ours</strong></td>
<td>33M</td>
<td><strong>42.48</strong></td>
<td><strong>40.71</strong></td>
</tr>
<tr>
<td>AWD-LSTM + MoS</td>
<td>35M</td>
<td>42.41</td>
<td>40.68</td>
</tr>
<tr>
<td><strong>AWD-LSTM + MoS + Ours</strong></td>
<td>35M</td>
<td><strong>40.27</strong></td>
<td><strong>38.65</strong></td>
</tr>
</tbody>
</table>

**Table: WT2**
## Improving on Machine Translation

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer Base</td>
<td>27.30</td>
</tr>
<tr>
<td><strong>Transformer Base + Ours</strong></td>
<td><strong>28.43</strong></td>
</tr>
<tr>
<td>Transformer Big</td>
<td>28.40</td>
</tr>
<tr>
<td><strong>Transformer Big + Ours</strong></td>
<td><strong>29.52</strong></td>
</tr>
</tbody>
</table>

**Table: WMT2014 En→De**

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer Small</td>
<td>32.47</td>
</tr>
<tr>
<td><strong>Transformer Small + Ours</strong></td>
<td><strong>33.61</strong></td>
</tr>
<tr>
<td>Transformer Base</td>
<td>34.43</td>
</tr>
<tr>
<td><strong>Transformer Base + Ours</strong></td>
<td><strong>35.18</strong></td>
</tr>
</tbody>
</table>

**Table: IWSLT2014 De→En**
Conclusions

Proposed an adversarial training mechanism for language modeling

- A Closed-form solution & easy to implement
- Diversity Promotion
- Strong empirical results

Thank You

*Poster #105, Today 06:30 PM – 09:00 PM @ Pacific Ballroom*