Revisiting the Architectures like Pointer Networks to Efficiently Improve the Next Word Distribution, Summarization Factuality, and Beyond

Haw-Shiuan Chang*, Zonghai Yao*, Alolika Gon, Hong Yu, Andrew McCallum
Can Large LM Learn to Output Arbitrary Next Word Distribution?

No
A Simple Example

• There are *plates*, *keys*, *scissors*, *toys*, and *balloons* in front of me, and I pick up the …

• Ideal distribution
  • *plates*   ~0.2
  • *keys*    ~0.2
  • *scissors* ~0.2
  • *toys*    ~0.2
  • *balloons* ~0.2
There are plates, keys, scissors, toys, and balloons in front of me, and I pick up the scissors.

I pick up the scissors and

There are toys, plates, scissors, keys, and balloons in front of me, and I pick up the keys.

The keys are cold and metallic.
Hallucination and Repetition

There are plates, keys, scissors, toys, and balloons in front of me, and I pick up the ...

- phone (from GPT-2)?

- Hallucination
  - Should copy but not copy

- I like tennis, baseball, golf, basketball, and ...

- tennis (from GPT-2)?

- Repetition
  - Should not copy but copy

<table>
<thead>
<tr>
<th></th>
<th>Softmax (GPT-2)</th>
<th>Pointer Network</th>
<th>Unlikelihood Training</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hallucination</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Repetition</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Why is GPT Unable to Learn to Copy Properly?
GPT-2 cannot predict both “woman” and “king” as the next word.

After debating whether to bow to the king or the woman first, the jester decided on the
After debating whether to bow to the **king** or the **woman** first, the jester decided on the **man**. 
Mixture of Softmax (MoS)

After debating whether to bow to the king or the woman first, the jester decided on the...

……

GPT-2

1 Softmax
2 vocab
3 Softmax

Weighted Sum

Softmax

vocab

f_{c,1}

f_{c,2}

f_{c,3}

facets

\pi_c

Chang, Haw-Shiu, and Andrew McCallum. "Softmax bottleneck makes language models unable to represent multi-mode word distributions." In ACL 2022.

After debating whether to bow to the king or the woman first, the jester decided on the
After debating whether to bow to the king or the woman first, the jester decided on the
After debating whether to bow to the king or the woman first, the jester decided on the...
After debating whether to bow to the king or the woman first, the jester decided on the
Softmax CEPR

Local Word Embeddings
- Encoder
- Decoder

Global Word Embeddings
- Decoder
- Encoder
- Top n
- Rest of Vocabulary

Context Partition (C)
- Encoder Context
- Decoder Context

Pointer Network (P)
- $f_{CLE}$
- $f_{CLD}$
- $f_{CLPE}$
- $f_{CLPD}$
- $f_{CLC}$
- $f_{CLE}$
- $f_{CLR}$
- $f_{CLV}$

Softmax
- CEPR
Experiments
# GPT-2 Perplexity Comparison

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Size</th>
<th>GPT-2 Small</th>
<th>OWT (↓)</th>
<th>Wiki (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Softmax (GPT-2)</td>
<td>125.0M</td>
<td>82.9</td>
<td>18.96</td>
<td>24.28</td>
</tr>
<tr>
<td>Softmax + Mi</td>
<td>130.9M</td>
<td>85.6</td>
<td>18.74</td>
<td>24.08</td>
</tr>
<tr>
<td>Mixture of Softmax (MoS) (Yang et al., 2018)</td>
<td>126.2M</td>
<td>130.2</td>
<td>18.97</td>
<td>24.10</td>
</tr>
<tr>
<td>MoS + Mi (Chang and McCallum, 2022)</td>
<td>133.3M</td>
<td>133.2</td>
<td>18.68</td>
<td>23.82</td>
</tr>
<tr>
<td>Pointer Generator (PG) (See et al., 2017)</td>
<td>126.2M</td>
<td>106.0</td>
<td>18.67</td>
<td>23.70</td>
</tr>
<tr>
<td>Pointer Sentinel (PS) (Merity et al., 2017)</td>
<td>126.2M</td>
<td>94.1</td>
<td>18.70</td>
<td>23.79</td>
</tr>
<tr>
<td>Softmax + R:20 + Mi</td>
<td>132.1M</td>
<td>90.4</td>
<td>18.67</td>
<td>24.03</td>
</tr>
<tr>
<td>Softmax + R:20,100 + Mi</td>
<td>133.3M</td>
<td>101.1</td>
<td>18.69</td>
<td>23.93</td>
</tr>
<tr>
<td>Softmax + C + Mi</td>
<td>132.1M</td>
<td>94.8</td>
<td>18.48</td>
<td>23.56</td>
</tr>
<tr>
<td>Softmax + P + Mi</td>
<td>133.3M</td>
<td>99.1</td>
<td>18.58</td>
<td>23.66</td>
</tr>
<tr>
<td>PG + Mi</td>
<td>133.3M</td>
<td>111.2</td>
<td>18.43</td>
<td>23.43</td>
</tr>
<tr>
<td>PS + Mi</td>
<td>133.3M</td>
<td>98.0</td>
<td>18.48</td>
<td>23.53</td>
</tr>
<tr>
<td>Softmax + CR:20,100 + Mi</td>
<td>134.5M</td>
<td>113.3</td>
<td>18.46</td>
<td>23.48</td>
</tr>
<tr>
<td>Softmax + CPR:20,100 + Mi</td>
<td>136.8M</td>
<td>119.9</td>
<td>18.43</td>
<td>23.42</td>
</tr>
<tr>
<td>MoS + CPR:20,100 + Mi</td>
<td>139.2M</td>
<td>165.1</td>
<td><strong>18.39</strong></td>
<td><strong>23.29</strong></td>
</tr>
</tbody>
</table>

The king & woman example could be solved by pointer network or MoS.
Summarization Experiments

- Improve BookSum more
- Probably because the John in one book is different from the John in another book

<table>
<thead>
<tr>
<th>Model Name</th>
<th>CNN/DM R1</th>
<th>CNN/DM CIDEr</th>
<th>CNN/DM factCC</th>
<th>CNN/DM MAUVE</th>
<th>XSUM R1</th>
<th>XSUM CIDEr</th>
<th>XSUM factCC</th>
<th>XSUM MAUVE</th>
<th>BookSum Paragraph R1</th>
<th>BookSum Paragraph CIDEr</th>
<th>BookSum Paragraph factCC</th>
<th>BookSum Paragraph MAUVE</th>
<th>SAMSUM R1</th>
<th>SAMSUM CIDEr</th>
<th>SAMSUM factCC</th>
<th>SAMSUM MAUVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Softmax (S)</td>
<td>38.255</td>
<td>0.442</td>
<td>0.462</td>
<td>0.861</td>
<td>28.713</td>
<td>0.446</td>
<td>0.254</td>
<td>0.939</td>
<td>16.313</td>
<td>0.083</td>
<td>0.424</td>
<td>0.328</td>
<td>39.472</td>
<td>0.817</td>
<td>0.577</td>
<td>0.898</td>
</tr>
<tr>
<td>CopyNet (Gu et al., 2016)</td>
<td>37.990</td>
<td>0.439</td>
<td>0.482</td>
<td>0.865</td>
<td>28.573</td>
<td>0.442</td>
<td>0.274</td>
<td>0.940</td>
<td>16.666</td>
<td>0.092</td>
<td>0.439</td>
<td>0.402</td>
<td>39.525</td>
<td>0.853</td>
<td>0.579</td>
<td>0.924</td>
</tr>
<tr>
<td>PG (See et al., 2017)</td>
<td>37.913</td>
<td>0.442</td>
<td>0.467</td>
<td>0.874</td>
<td>28.777</td>
<td>0.450</td>
<td>0.257</td>
<td>0.931</td>
<td>16.432</td>
<td>0.088</td>
<td>0.429</td>
<td>0.376</td>
<td>32.451</td>
<td>0.585</td>
<td>0.552</td>
<td>0.153</td>
</tr>
<tr>
<td>PS (Merity et al., 2017)</td>
<td>38.058</td>
<td>0.444</td>
<td>0.435</td>
<td>0.876</td>
<td>29.155</td>
<td>0.464</td>
<td>0.275</td>
<td>0.934</td>
<td>16.357</td>
<td>0.086</td>
<td>0.431</td>
<td>0.390</td>
<td>38.731</td>
<td>0.817</td>
<td>0.578</td>
<td>0.865</td>
</tr>
<tr>
<td>S + R:20</td>
<td>37.881</td>
<td>0.435</td>
<td>0.440</td>
<td>0.856</td>
<td>29.076</td>
<td>0.465</td>
<td>0.232</td>
<td>0.941</td>
<td>16.542</td>
<td>0.090</td>
<td>0.435</td>
<td>0.399</td>
<td>39.073</td>
<td>0.752</td>
<td>0.579</td>
<td>0.847</td>
</tr>
<tr>
<td>S + E</td>
<td>38.137</td>
<td>0.441</td>
<td>0.441</td>
<td>0.872</td>
<td>29.327</td>
<td>0.479</td>
<td>0.275</td>
<td>0.948</td>
<td>16.628</td>
<td>0.093</td>
<td>0.436</td>
<td>0.403</td>
<td>40.055</td>
<td>0.835</td>
<td>0.583</td>
<td>0.943</td>
</tr>
<tr>
<td>S + CE</td>
<td>38.461</td>
<td>0.460</td>
<td>0.475</td>
<td>0.877</td>
<td>29.067</td>
<td>0.459</td>
<td>0.276</td>
<td>0.936</td>
<td>16.638</td>
<td>0.093</td>
<td>0.436</td>
<td>0.400</td>
<td>40.505</td>
<td>0.846</td>
<td>0.580</td>
<td>0.915</td>
</tr>
<tr>
<td>S + CE:R:20</td>
<td>38.346</td>
<td>0.456</td>
<td>0.482</td>
<td>0.890</td>
<td>29.067</td>
<td>0.459</td>
<td>0.276</td>
<td>0.942</td>
<td>16.894</td>
<td>0.098</td>
<td>0.440</td>
<td>0.418</td>
<td>40.127</td>
<td>0.891</td>
<td>0.582</td>
<td>0.946</td>
</tr>
<tr>
<td>S + CEPR:20</td>
<td>38.807</td>
<td>0.456</td>
<td>0.481</td>
<td>0.877</td>
<td>29.395</td>
<td>0.474</td>
<td>0.273</td>
<td>0.942</td>
<td>16.738</td>
<td>0.096</td>
<td>0.438</td>
<td>0.426</td>
<td>40.328</td>
<td>0.874</td>
<td>0.582</td>
<td>0.932</td>
</tr>
<tr>
<td>S + CEPR:20 + Mi</td>
<td>38.675</td>
<td>0.451</td>
<td>0.475</td>
<td>0.878</td>
<td>29.348</td>
<td>0.470</td>
<td>0.275</td>
<td>0.946</td>
<td>16.761</td>
<td>0.096</td>
<td>0.424</td>
<td>0.467</td>
<td>44.348</td>
<td>1.046</td>
<td>0.574</td>
<td>0.986</td>
</tr>
</tbody>
</table>

Comparable to some reranker methods

+ 30%
Conclusion

• Softmax bottleneck
  • -> hallucination and repetition problems

• Breaking the softmax bottleneck
  • -> improvements from pointer networks, rerankers, and mixture of softmax (MoS)

• Pointer networks + rerankers + MoS
  • -> softmax-CPR❤️ and softmax-CEPR
Our Other Work on Improving Single Embedding Representation

Future Work: Variable Assignment

- LM on code examples: Codex (OpenAI), AlphaCode (DeepMind)

- LM on math examples:

\[
a \leftarrow 1; \ b \leftarrow 3, \ x \leftarrow b; \text{ Which variable(s) > 2?}
\]
Future Work: Variable Assignment

- LM on code examples: Codex (OpenAI), AlphaCode (DeepMind)
- LM on math examples:
  \[ a \leftarrow 1; \quad b \leftarrow 3, \quad x \leftarrow b; \quad \text{Which variable(s) > 2?} \]