Softmax Bottleneck Makes Language Models Unable to Represent Multi-mode Word Distributions

Haw-Shiuan Chang Andrew McCallum

UMass Amherst Manning College of Information & Computer Sciences
Outline

- Introduction
- Theoretical Analysis
- Method
- Experiments
- Conclusion and Future Work
Outline

• Introduction
  • Theoretical Analysis
  • Method
  • Experiments
• Conclusion and Future Work
Sampling Distributions from Large LMs

- Assist creative writing (Akoury et al., 2020)
- Reduce the cost of building datasets (West et al., 2021)
- Generate codes (Li et al., 2022)
- Solve math problems (Cobbe et al., 2021)

Input

John goes to

Large Language Model (LM)

Next Word Distribution

bed 0.32
work 0.1
...

Sampling


Can large LMs learn any distribution over the next word?
After debating whether to bow to the **king** or the **woman** first, the jester decided on the
Most of Existing Approaches

After debating whether to bow to the king or the woman first, the jester decided on the woman. The output next word probability is:

- woman: 0.5
- queen: 0.2
- man: 0.1

The ideal next word probability is:

- woman: 0.4
- king: 0.4
- lady: 0.03

The GPT-2 Encoder outputs the hidden state $h_t$ which is used for language generation.
Predicting “woman” as the Next Word
Could GPT-2 Predict Both “woman” and “king” as the Next Word?
No, if there are some words between them and GPT-2 has only one hidden state
Outline

• Introduction
• Theoretical Analysis
• Method
• Experiments
• Conclusion and Future Work
Softmax Bottleneck (Yang et al., 2018)

- If $V>D$, we cannot output arbitrary probabilities over $V$ words
- Limitations
  - Serious among which words?
  - Affect the top words? If yes, when?
  - Disappears after making $D>V$?

Our Theoretical Improvements

- If N words are in a small subspace, we cannot rank N words arbitrarily
- Improvements
  - Serious among which words? -> Among words in a small subspace
  - Affect the top words? If yes, when? -> Yes. When the ideal distribution is multi-mode
  - Disappears after making D>V? -> No, if some words are in a small subspace
A Limitation of Single Embedding

**Theorem 1 (simplified):** If many word embeddings are linearly dependent, the softmax in a LM cannot rank the words arbitrarily.

**Example:** If “woman - man = queen - king”, GPT-2 cannot rank the word woman and king as the top 2 words.

**Example:** If “UMass = 0.2 University + 0.2 Massachusetts”, GPT-2 cannot rank a rare word UMass on top of the similar popular words University and Massachusetts (Demeter et al., 2020).

**Linear Algebra Intuition:** N+1 words are linear dependent ➔
They are in subspace with d < N ➔ cannot have arbitrary probabilities.
Theorem 2 (simplified): If many word embeddings are approximately linearly dependent and the magnitude of the hidden state has an upperbound, the softmax in a LM cannot assign very small probabilities to some words.

Example: If “woman + king = queen + man + ε”, GPT-2 cannot make the logits of queen and man much smaller than the logits of king and woman.

Example: If “woman = man + ε”, GPT-2 cannot make the logits of man much smaller than the logits of woman.

Intuition: \( h^T \text{king} + h^T \text{woman} = h^T \text{queen} + h^T \text{man} + h^T \varepsilon \), and we can ignore \( h^T \varepsilon \) if \(||h||\) and \(||\varepsilon||\) are both small.
Outline

• Introduction

• Theoretical Analysis

• Method

• Experiments

• Conclusion and Future Work
After debating whether to bow to the king or the woman first, the jester decided on the...
Mixture of Softmax (Yang et al., 2018)

Mixture of Softmax (MoS) is one of the few effective modifications for Transformer (Narang et al., 2021)

Facet embeddings

A new bottleneck

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


Sharan Narang, et al. Do transformer modifications transfer across implementations and applications? EMNLP 2021
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
Multi-facet Softmax (MFS)

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

• Introduction
• Theoretical Analysis
• Method
• Experiments
• Conclusion and Future Work
Multi-facet Softmax (MFS) Perplexity

<table>
<thead>
<tr>
<th>Models ↓</th>
<th>Configuration</th>
<th>Size</th>
<th>GPT-2 Small</th>
<th>GPT-2 Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#S</td>
<td>#I</td>
<td>#P</td>
<td>Time</td>
</tr>
<tr>
<td>Softmax (GPT-2)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>163.6M</td>
</tr>
<tr>
<td>SigSoftmax (Kanai et al., 2018)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>163.6M</td>
</tr>
<tr>
<td>Softmax + Multi-input</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>169.5M</td>
</tr>
<tr>
<td>Softmax + Multi-partition</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>165.4M</td>
</tr>
<tr>
<td>MoS (Yang et al., 2018) (4)</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>164.8M</td>
</tr>
<tr>
<td>MoS (Yang et al., 2018) (3)</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>164.8M</td>
</tr>
<tr>
<td>DOC (Takase et al., 2018)</td>
<td>3</td>
<td>9</td>
<td>1</td>
<td>171.9M</td>
</tr>
<tr>
<td>MFS w/o Multi-partition</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>166.6M</td>
</tr>
<tr>
<td>MFS w/o Multi-input</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>175.4M</td>
</tr>
</tbody>
</table>

Table 1: Perplexity comparison between MFS (Ours) and baselines. #S, #I, #P are the number of softmaxes (i.e., \(K\)), input hidden states, and partitions, respectively. The top four baselines use a single softmax. OWT and Wiki are the test set perplexity of OpenWebText and Wikipedia 2021, respectively. The standard errors of all models are smaller than 0.02 perplexity. We also compare the number of parameters and the inference time on one batch.

Improvement of MFS over Softmax is around 15% between GPT-2 Small and GPT-2 Medium (with 3x parameters)


Dwarak Govind Parthiban, Yongyi Mao, and Diana Inkpen. On the softmax bottleneck of recurrent language models. In AAAI 2021
Examples

<table>
<thead>
<tr>
<th>Corpus →</th>
<th>OpenWebText</th>
<th>Wikipedia 2021</th>
<th>Analogy in Templates (Section 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Context</td>
<td>... The Elastic Endpoint Security and Elastic SIEM solutions mentioned in this post are now referred to as <strong>Elastic</strong>.</td>
<td>... law and chance working together cannot generate CSI, either. Moreover, he claims that <strong>CSI</strong>.</td>
<td>I went to Paris and Germany before, and I love one of the places more, which is <strong>Germany</strong>.</td>
</tr>
<tr>
<td>Softmax (GPT-2)</td>
<td>the 0.087, E 0.043, End 0.039, <strong>Elastic</strong> 0.220, the 0.089, EC 0.033.</td>
<td>the 0.174, this 0.054, if 0.038, CSI 0.186, the 0.140, there 0.033.</td>
<td>Paris 0.893, France 0.045, <strong>Germany</strong> 0.033, Paris 0.544, <strong>Germany</strong> 0.389, France 0.064.</td>
</tr>
<tr>
<td>MFS (Ours)</td>
<td>MFS Softmax 1: end 0.051, the 0.043, security 0.023, <strong>Elastic</strong> 0.652, EC 0.080, ES 0.046, the 0.193, E 0.040, a 0.014.</td>
<td>MFS Softmax 2: the 0.191, law 0.127, if 0.053, <strong>CSI</strong> 0.677, law 0.029, laws 0.019.</td>
<td>MFS Softmax 3: Paris 0.979, France 0.013, <strong>Germany</strong> 0.007, Paris 1.000 Berlin 0.000.</td>
</tr>
<tr>
<td>MFS Softmax 1</td>
<td>MFS Softmax 2</td>
<td>MFS Softmax 3</td>
<td></td>
</tr>
</tbody>
</table>

Worse word similarity (Non-English text in OpenWebText): 2.5x Improvement

Top 10% most diverse facets: 3x Improvement

Candidates have an analogical relation: Perplexity 2.3 -> 1.7
Outline

• Introduction
• Theoretical Analysis
• Method
• Experiments
• Conclusion and Future Work
Conclusion

Theory

• Multi-mode distribution must exist if some word embeddings are in a small subspace

Method

• We propose two enhancements for mixture of softmax (MoS)

Analysis

• Our proposed method improves softmax layer in GPT-2 especially when the ideal next word distribution is multi-mode

Stolen Probability (Demeter et al., 2020)

Softmax Bottleneck (Yang et al., 2018)

MoS (Yang et al., 2018)

Adding Nonlinearity (Kanai et al., 2018)

Improvement

MFS (Ours)

Softmax bottleneck cannot explain the improvement of MoS (Parthiban et al., 2021)

Multi-mode is a good explanation (Ours)


David Demeter, Gregory Kimmel, and Doug Downey. Stolen probability: A structural weakness of neural language models. In ACL. 2020


Dwarak Govind Parthiban, Yongyi Mao, and Diana Inkpen. On the softmax bottleneck of recurrent language models. In AAAI 2021
Future Work

• How much MFS could help huge language models (e.g., GPT-3)
• Whether MFS could improve
  • NLU tasks
  • NLG tasks
  • Other extreme classification models using an output softmax layer
• The word similarity should be context dependent rather than globally fixed
Appendix
Perplexity improvement of MoS and MFS is much smaller compared to GPT-2

MFS still doubles the improvement of MoS over Softmax
Name something that people usually do before they leave for work.

Table 5: ProtoQA performances. All the numbers except perplexity are the percentages of the predictions that match the ground truth exactly on the crowdsourced development set. Max answers top k implies only evaluating the top k answers. Max incorrect top k indicates only evaluating the top answers that contain k errors. The best average performances are highlighted and the standard errors are reported as the confidence interval.
Proof Sketch

• Linearly dependent among \{w_{l1}, \ldots, w_{lr}, \ldots, w_{rl}\}

• 1 w_{king} - 1 w_{queen} = 1 w_{man} - 1 w_{woman}

• 1 w_{king} + 1 w_{woman} = 1 w_{queen} + 1 w_{man}

× h (hidden state) on both side

Large   Large   Small   Small

• 1 h^T w_{king} + 1 h^T w_{woman} = 1 h^T w_{queen} + 1 h^T w_{man}

If \exists h, s.t \min(h^T w_{king}, h^T w_{woman}) > \max(h^T w_{queen}, h^T w_{man})

Large   Large

• 1 h^T w_{king} + 1 h^T w_{woman} ≥
  2 \min(h^T w_{king}, h^T w_{woman}) ≥
  2 \max(h^T w_{queen}, h^T w_{man}) ≥
  1 h^T w_{queen} + 1 h^T w_{man} \quad (\rightarrow \leftarrow)

Small   Small

• Thus, the logits of LM cannot rank both \textit{king} and \textit{woman} on top of \textit{queen} and \textit{man}

• Linearly dependent among \{w_{l1}, \ldots, w_{lr}, \ldots, w_{rl}\}

• a_{l1} w_{l1} + \ldots + a_{lr} w_{lr} = a_{r1} w_{r1} + \ldots + a_{rn} w_{rn}

  • All coefficient a_{li} > 0, a_{ri} > 0

  • WLOG a_{li} + \ldots + a_{lr} ≥ a_{r1} + \ldots + a_{rn}

  • a_{l1} h^T w_{l1} + \ldots + a_{lr} h^T w_{lr} = a_{r1} h^T w_{r1} + \ldots + a_{rn} h^T w_{rn}

  • If \exists h, s.t \min(h^T w_{l1}, h^T w_{l2}) > \max(h^T w_{r1}, h^T w_{r2})

    • a_{l1} h^T w_{l1} + \ldots + a_{lr} h^T w_{lr} ≥
      (a_{l1} + \ldots + a_{lr}) \min(h^T w_{l1}, h^T w_{l2}) >
      (a_{r1} + \ldots + a_{rn}) \max(h^T w_{r1}, h^T w_{r2}) ≥
      a_{r1} h^T w_{r1} + \ldots + a_{rn} h^T w_{rn} \quad (\rightarrow \leftarrow)

• Thus, the logits of LM cannot rank all the left words on top of the right words.