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

Haw-Shiuan Chang　　Andrew McCallum

UMassAmherst | Manning College of Information & Computer Sciences

Outline

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

Outline

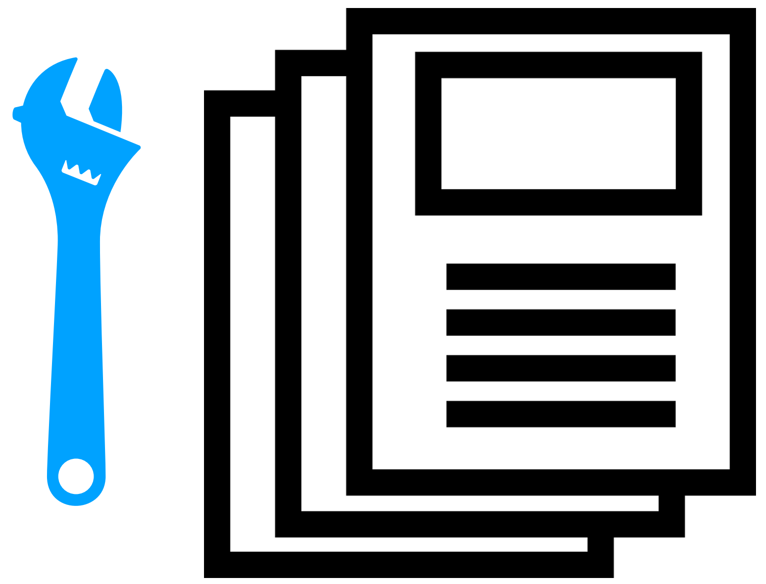
- Introduction
- Theoretical Analysis
- Method
- Experiments
- Conclusion and Future Work

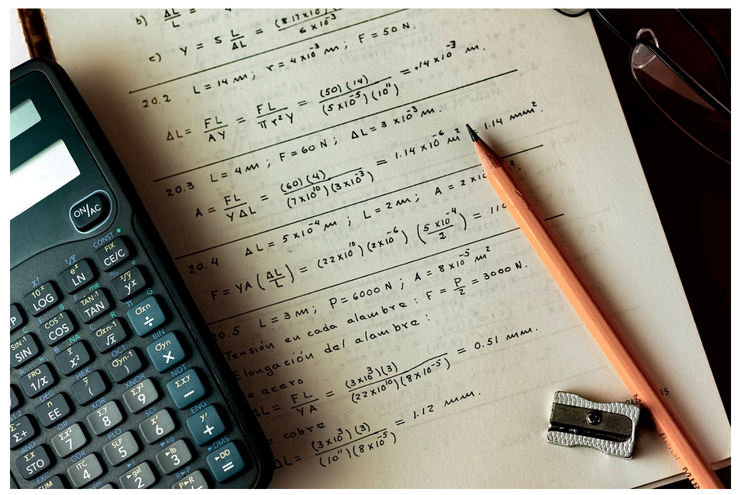
Sampling Distributions from Large LMs

Input → **Next Word Distribution**

John goes to → Large Language Model (LM) → bed 0.32, work 0.1, … ← Sampling

- 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)





……

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In EMNLP

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. arXiv preprint arXiv:2110.07178.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, R mi Leblond, Tom Eccles, James Keeling, et al. 2022. Competition-level code generation with alphacode. arXiv preprint arXiv:2203.07814.

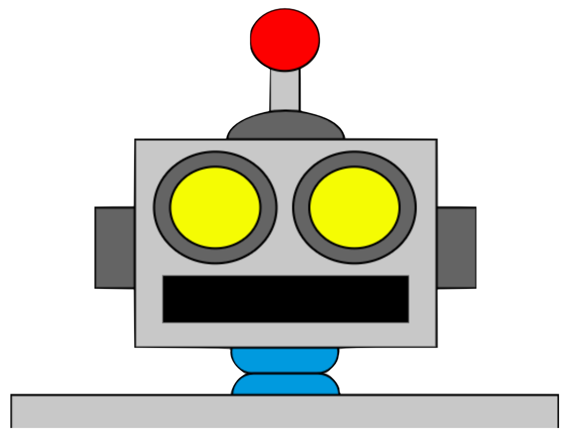
Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

Can large LMs learn any distribution over the next word?

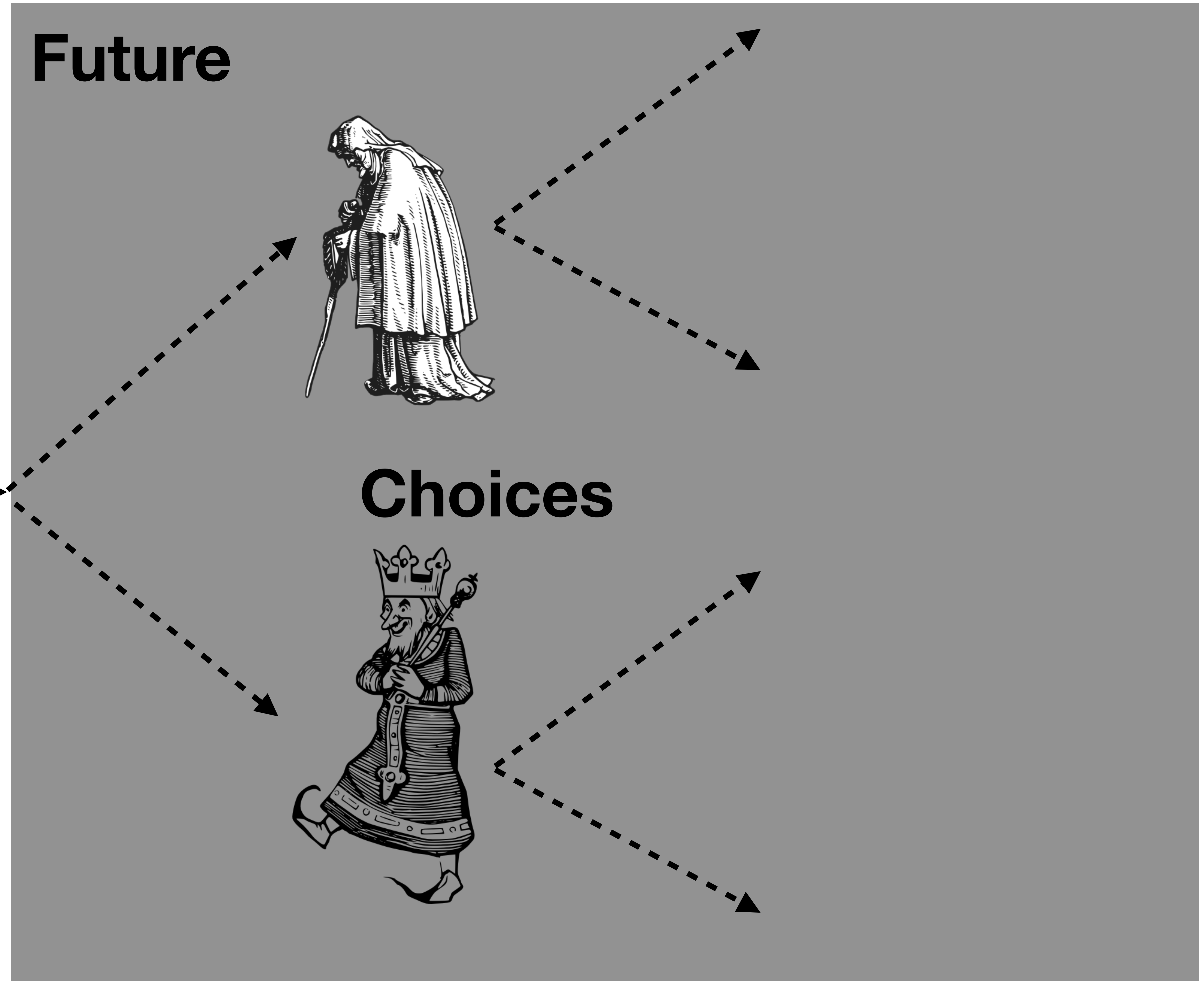
An Ambiguous Context

Context

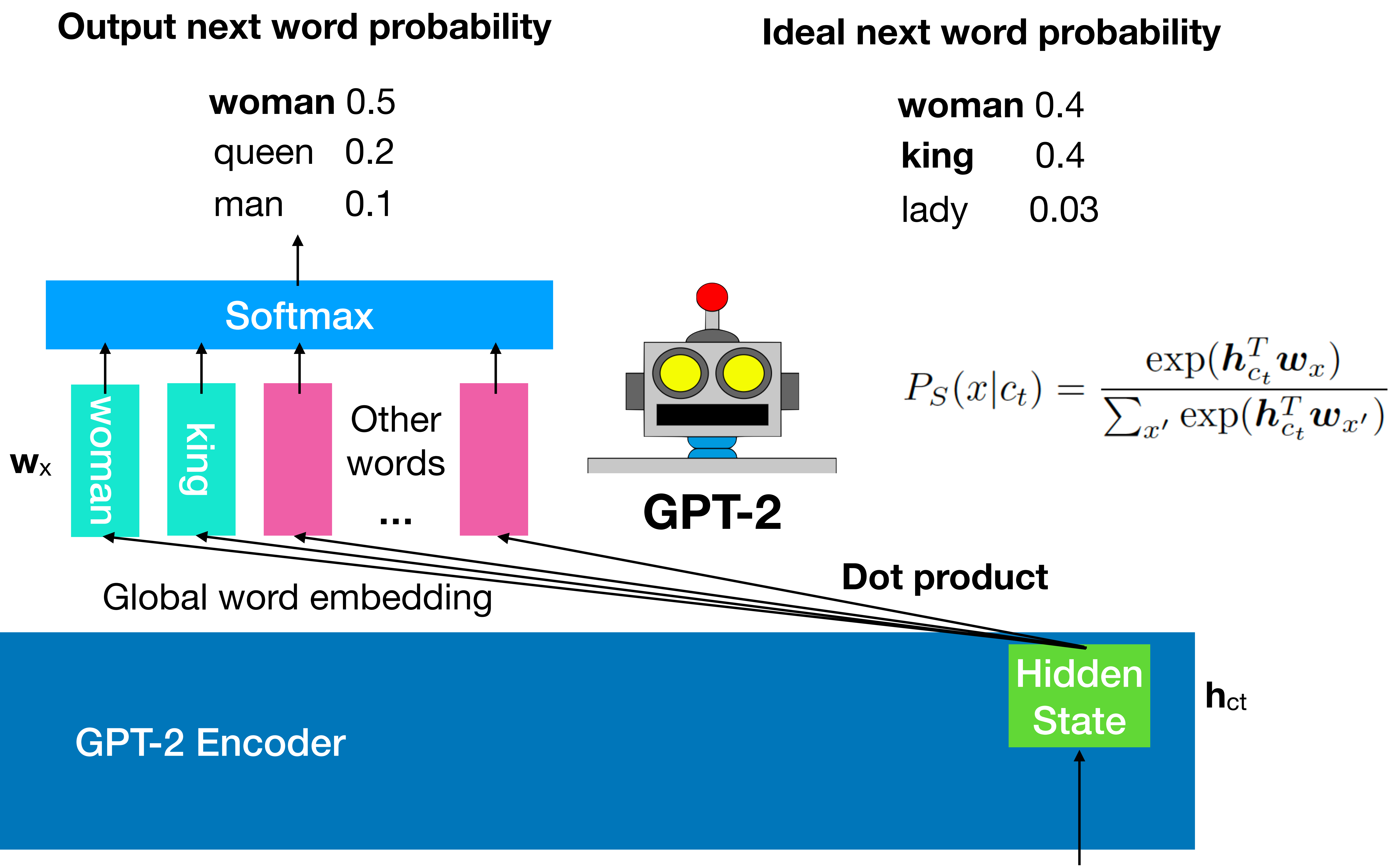
After debating whether to bow to the **king** or the **woman** first, the jester decided on the



GPT-2



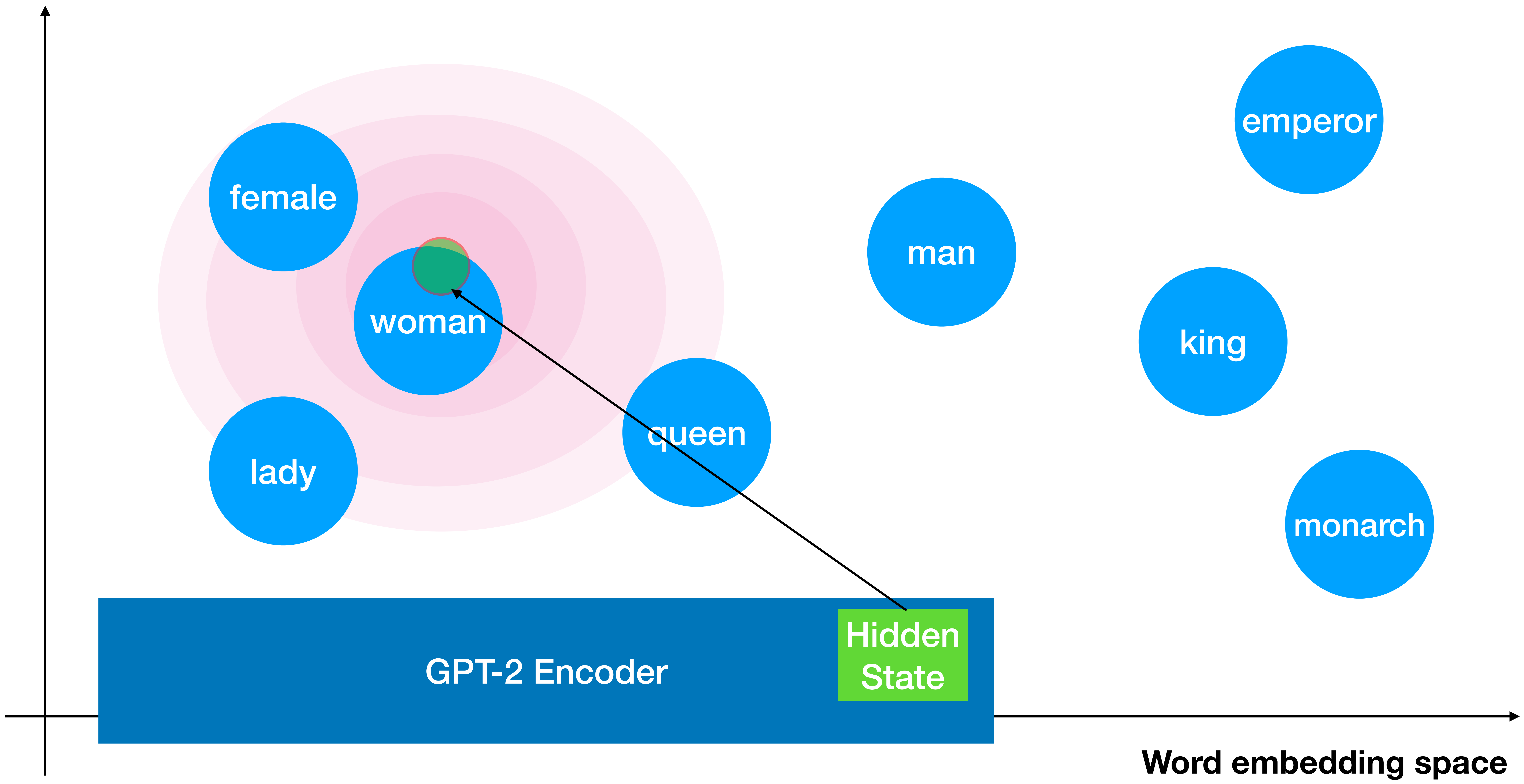
Most of Existing Approaches



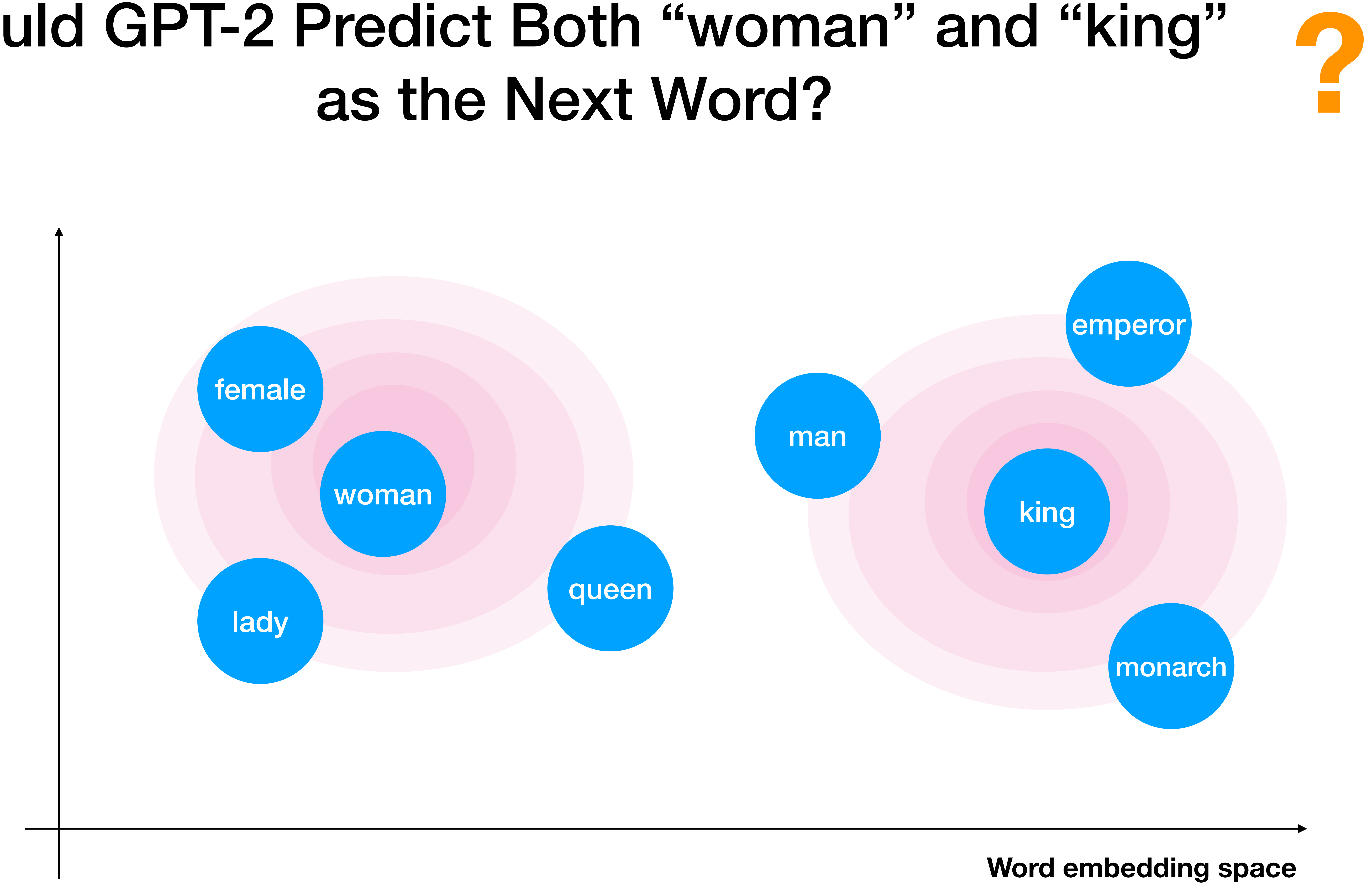
$$P_S(x|c_t) = \frac{\exp(\boldsymbol{h}_{c_t}^T \boldsymbol{w}_x)}{\sum_{x'} \exp(\boldsymbol{h}_{c_t}^T \boldsymbol{w}_{x'})}$$

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

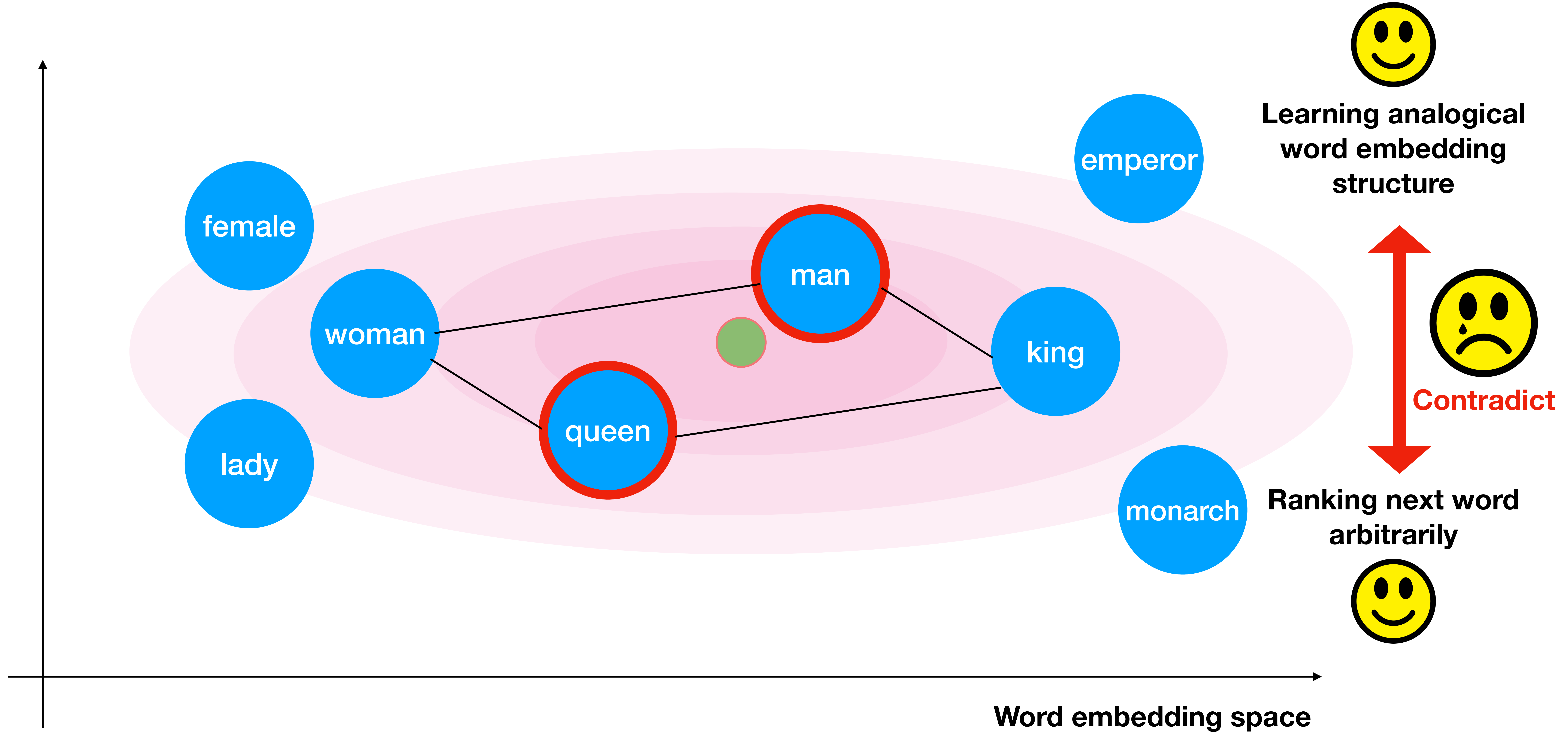
Predicting “woman” as the Next Word



Could GPT-2 Predict Both "woman" and "king" as the Next Word?



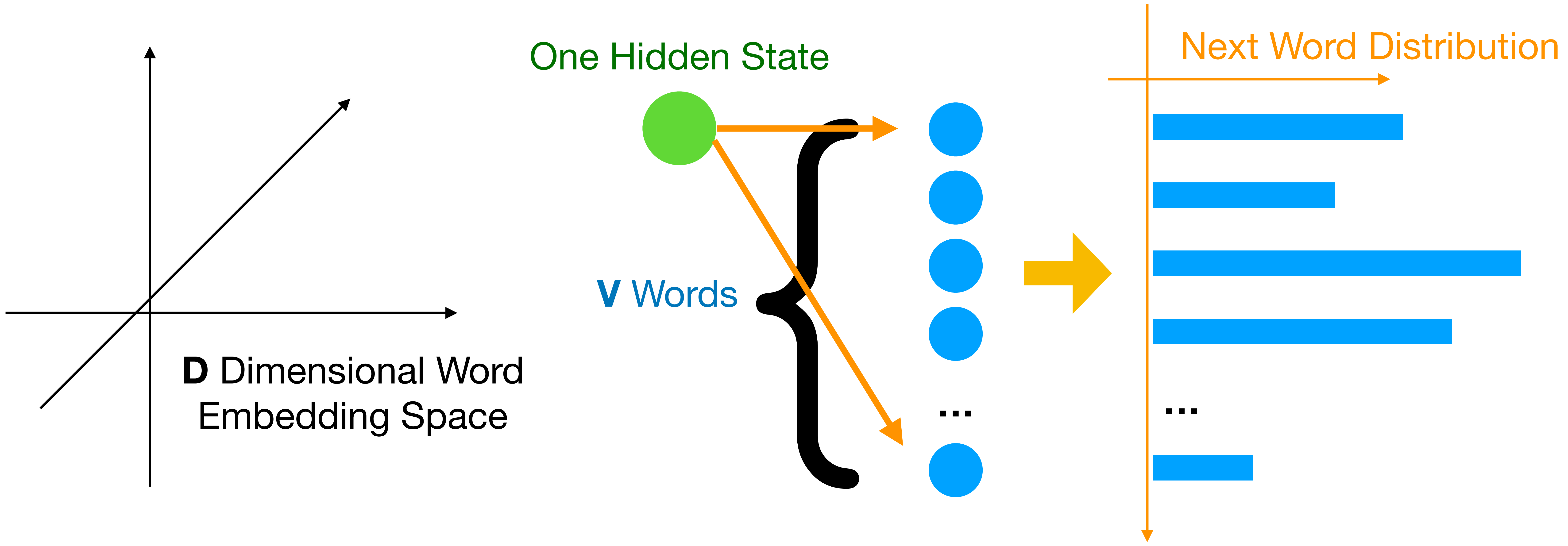
No, if there are some words between them and GPT-2 has only one hidden state



Outline

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

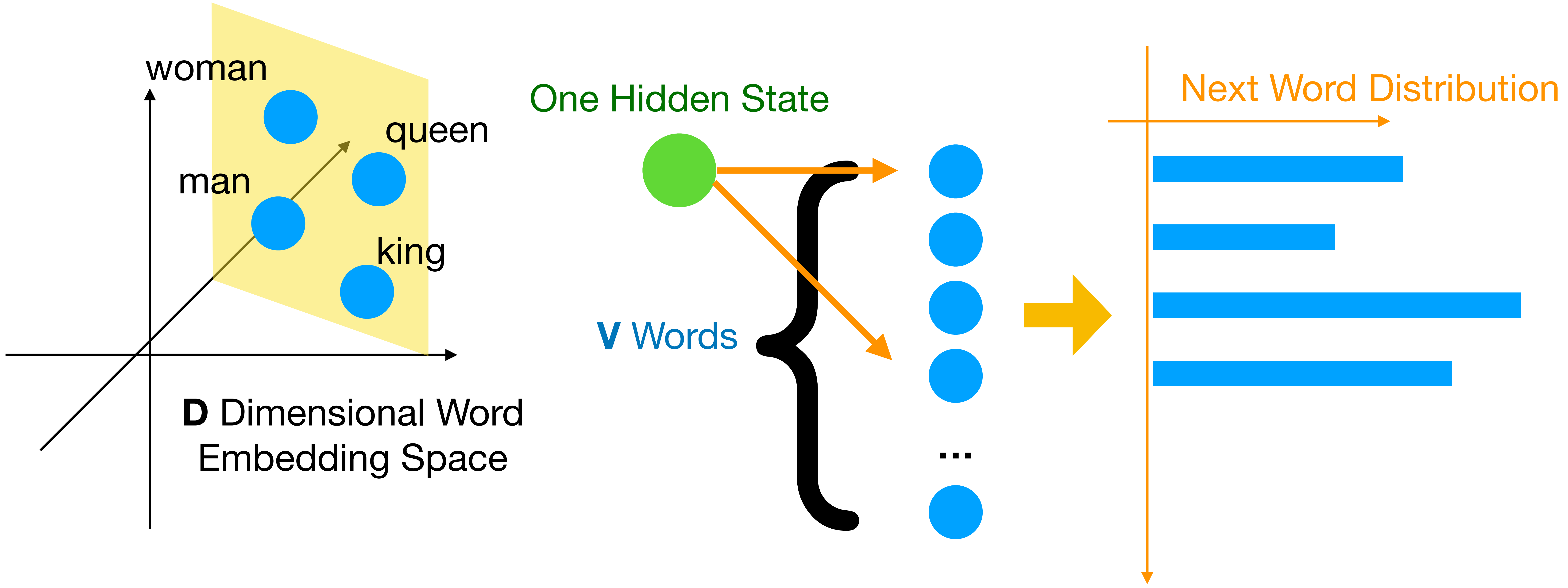
Softmax Bottleneck (Yang et al., 2018)



- 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?

Yang, Zhilin, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. "Breaking the Softmax Bottleneck: A High-Rank RNN Language Model." In *ICLR*. 2018.

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

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

Approximately Linearly Dependent

Theorem 2 (simplified): If many word embeddings are approximately linearly dependent and the magnitude of the hidden state has a 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*

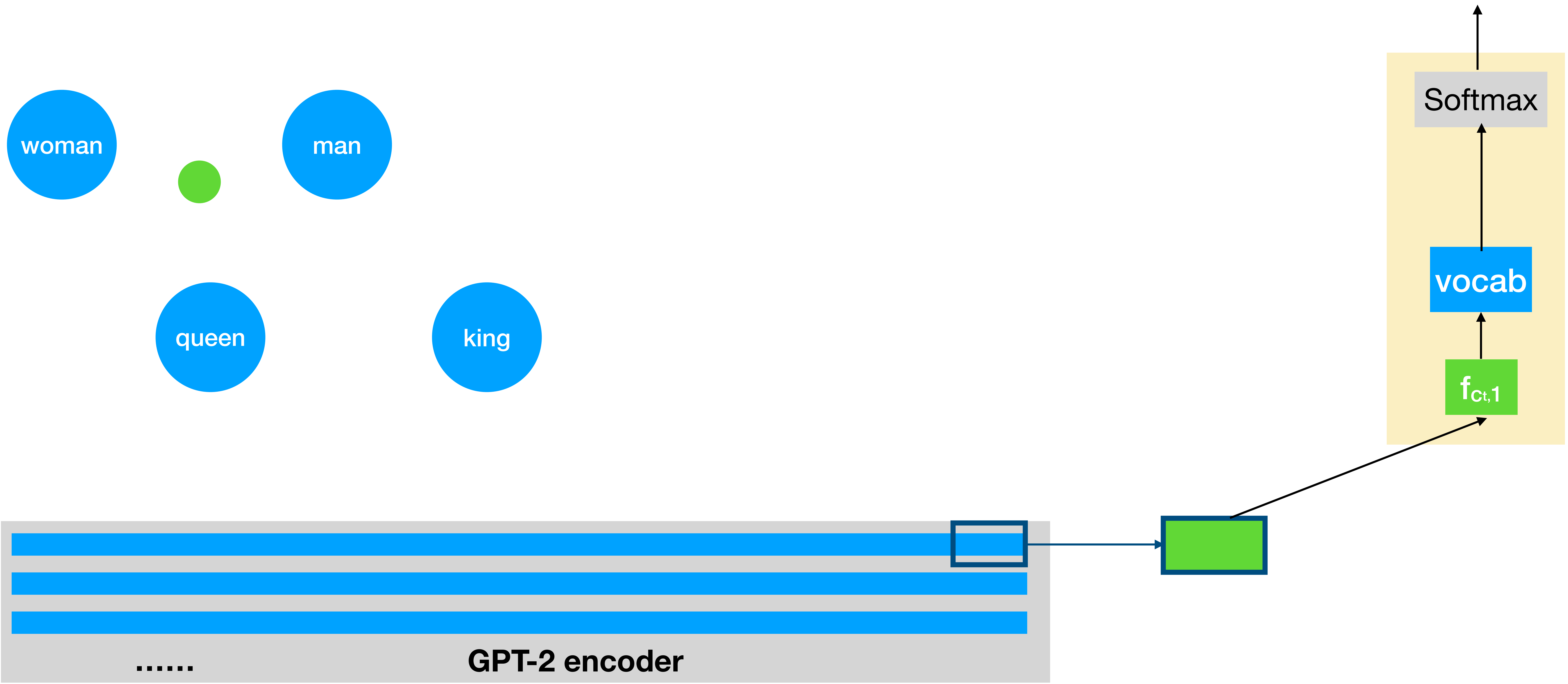
Example: If "*woman* = *man* + *ε*", GPT-2 cannot make the logits of *man* much smaller than the logits of *woman*

Intuition: $\boldsymbol{h}^T$ *king* + $\boldsymbol{h}^T$ *woman* **=** $\boldsymbol{h}^T$ *queen* + $\boldsymbol{h}^T$ *man* + $\boldsymbol{h}^T$ *ε*, and we can ignore $\boldsymbol{h}^T$ *ε* if $||\boldsymbol{h}||$ and $||\varepsilon||$ are both small

Outline

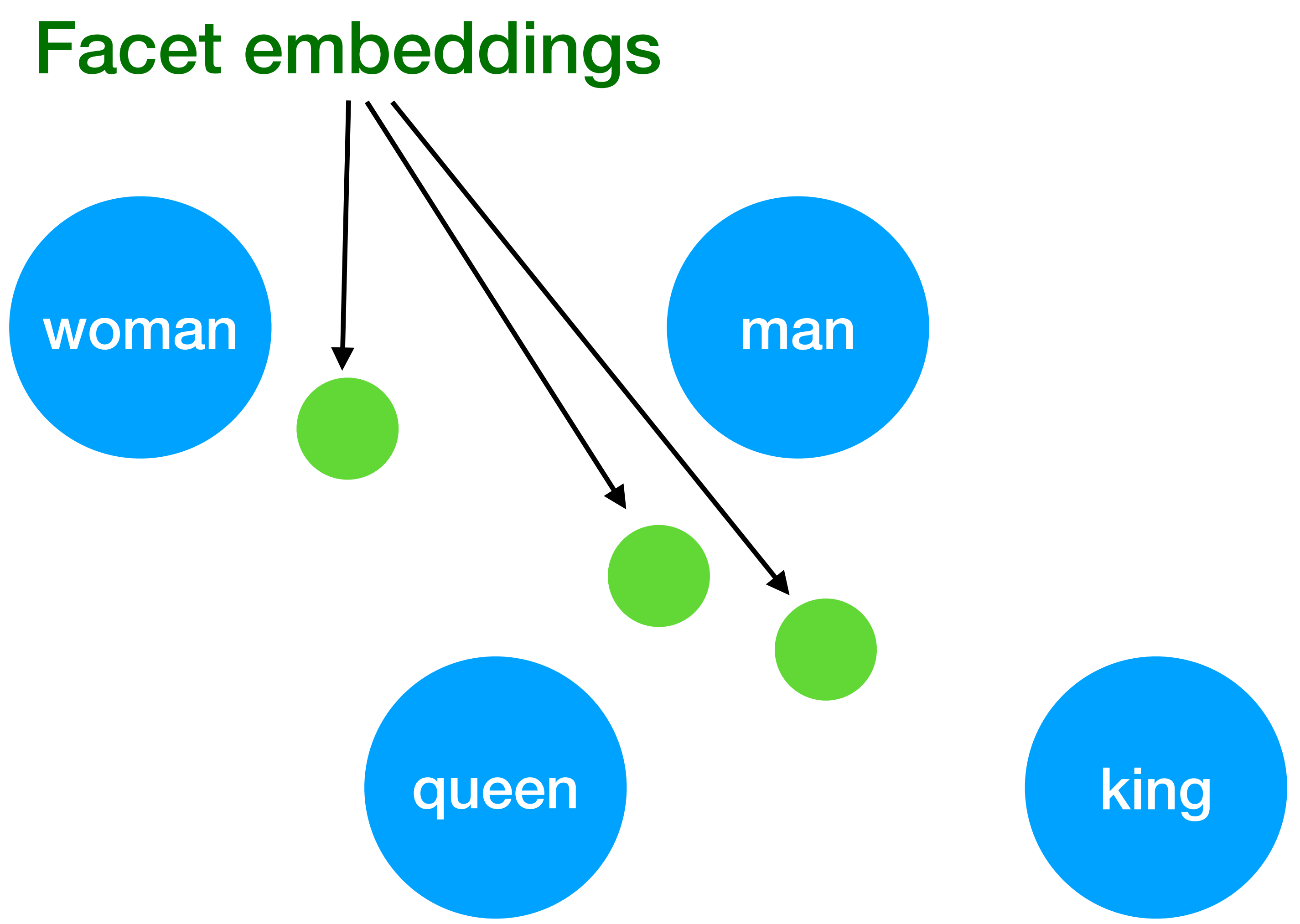
- Introduction
- Theoretical Analysis
- Method
- Experiments
- Conclusion and Future Work

GPT-2 (Softmax)

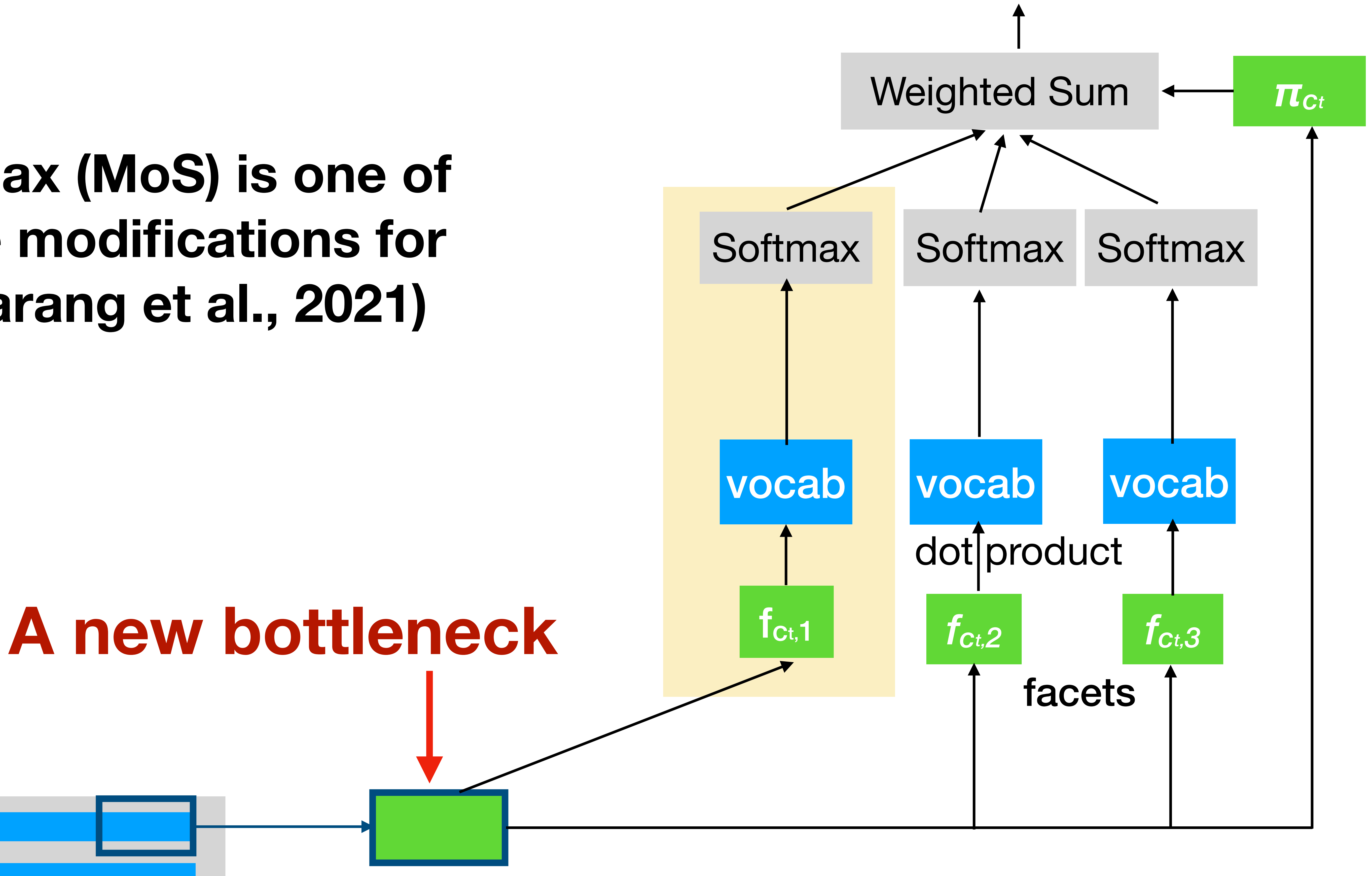


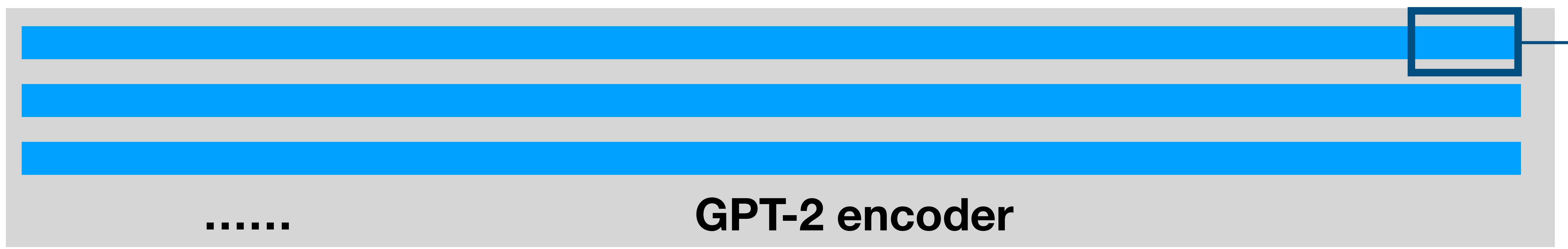
After debating whether to bow to the **king** or the **woman** first, the jester decided on the

Mixture of Softmax (Yang et al., 2018)



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





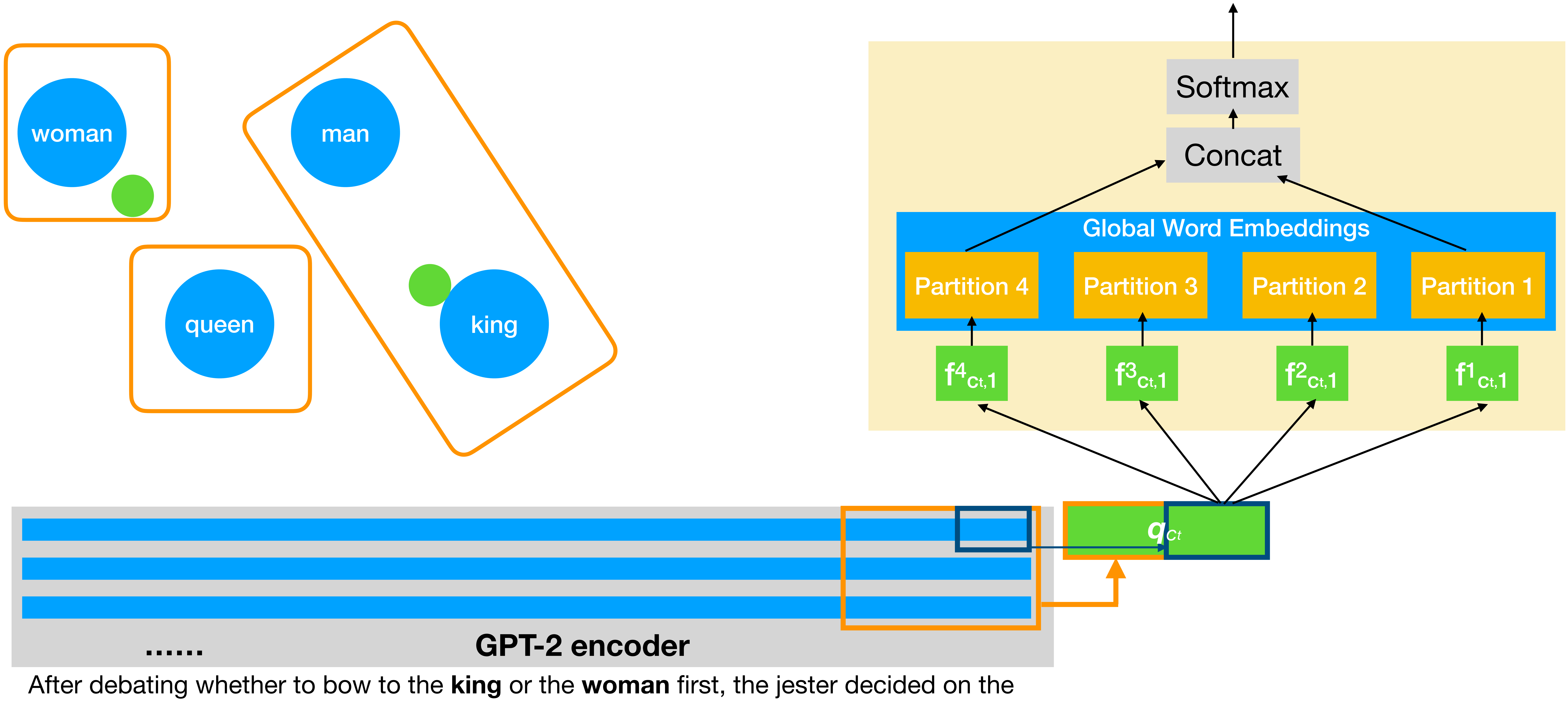
After debating whether to bow to the **king** or the **woman** first, the jester decided on the

Yang, Zhilin, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. "Breaking the Softmax Bottleneck: A High-Rank RNN Language Model." In *International Conference on Learning Representations*. 2018.

Sharan Narang, et al. Do transformer modifications transfer across implementations and applications? *EMNLP 2021*

MoS + Multi-input

Multiple Partitions



Multi-facet Softmax (MFS)

Outline

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

Multi-facet Softmax (MFS) Perplexity

Only adding nonlinearity is not enough (Parthiban et al., 2021)

Models ↓	Configuration #S	#I	#P	GPT-2 Small Size	Time	OWT	Wiki	GPT-2 Medium Size	Time	OWT	Wiki
Softmax (GPT-2)	1	1	1	163.6M	84ms	18.72	24.06	407.3M	212ms	15.89	20.34
SigSoftmax (Kanai et al., 2018)	1	1	1	163.6M	91ms	18.63	24.06	407.3M	221ms	16.07	20.65
Softmax + Multi-input	1	9	1	169.5M	87ms	18.50	23.89	417.8M	219ms	15.76	20.29
Softmax + Multi-partition	1	1	4	165.4M	88ms	18.77	24.08	410.5M	218ms	15.89	20.30
MoS (Yang et al., 2018) (4)	4	1	1	165.4M	152ms	18.61	23.77	410.5M	299ms	15.75	20.08
MoS (Yang et al., 2018) (3)	3	1	1	164.8M	130ms	18.63	23.81	409.4M	270ms	15.79	20.11
DOC (Takase et al., 2018)	3	3	1	164.8M	130ms	18.69	24.02	409.4M	270ms	15.88	20.34
MFS w/o Multi-partition	3	9	1	171.9M	133ms	18.37	23.56	422.0M	276ms	15.65	20.06
MFS w/o Multi-input	3	1	4	166.6M	134ms	18.60	23.72	412.6M	275ms	15.71	20.08
MFS (Ours)	3	9	4	175.4M	138ms	**18.29**	**23.45**	428.3M	283ms	**15.64**	**20.02**

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.

Multiple input hidden states help

Multiple partitions help

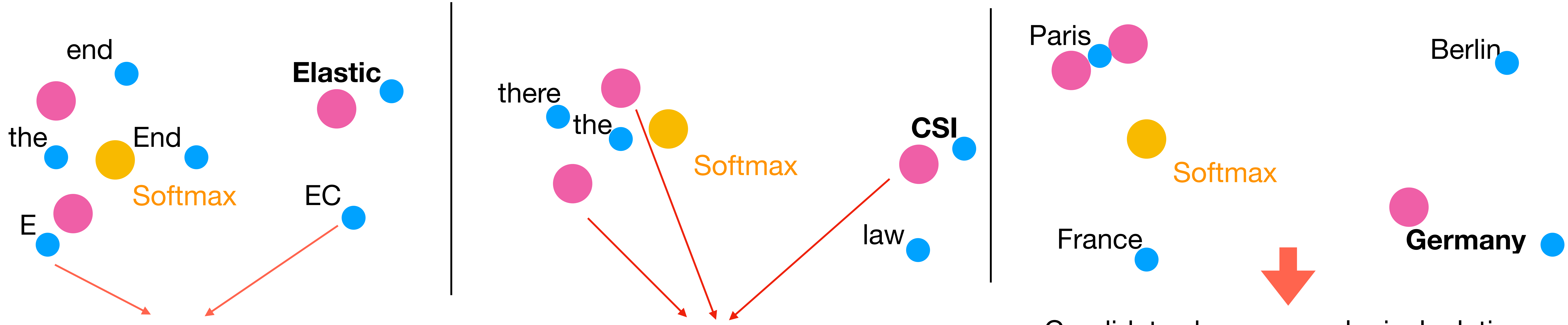
Improvement of MFS over Softmax is around 15% between GPT-2 Small and GPT-2 Medium (with 3x parameters)

Sekitoshi Kanai, Yasuhiro Fujiwara, Yuki Yamanaka, and Shuichi Adachi. Sigsoftmax: Reanalysis of the softmax bottleneck. In NeurIPS 2018

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

Examples

Corpus →	OpenWebText	Wikipedia 2021	Analogy in Templates (Section 5)
Input Context	... The Elastic Endpoint Security and Elastic SIEM solutions mentioned in this post are now referred to as **Elastic**	... law and chance working together cannot generate CSI, either. Moreover, he claims that **CSI**	I went to Paris and Germany before, and I love one of the places more, which is **Germany**
Softmax (GPT-2)	the 0.087, E 0.043, End 0.039	the 0.174, this 0.054, if 0.038	Paris 0.893, France 0.045, **Germany** 0.033
MFS (Ours)	**Elastic** 0.220, the 0.089, EC 0.033	CSI 0.186, the 0.140, there 0.033	Paris 0.544, **Germany** 0.389, France 0.064
MFS Softmax 1	end 0.051, the 0.043, security 0.023	the 0.191, law 0.127, if 0.053	Paris 0.979, France 0.013, **Germany** 0.007
MFS Softmax 2	**Elastic** 0.652, EC 0.080, ES 0.046	the 0.191, there 0.049, this 0.047	Paris 1.000 Berlin 0.000 ##Paris 0.000
MFS Softmax 3	the 0.193, E 0.040, a 0.014	**CSI** 0.677, law 0.029, laws 0.019	**Germany** 0.852, France 0.139, China 0.004



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

Stolen Probability (Demeter et al., 2020)

Softmax Bottleneck (Yang et al., 2018)

Generalization

Multi-mode Distribution (Ours)

Method

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

MoS (Yang et al., 2018)

Adding Nonlinearity (Kanai et al., 2018)

Improvement

MFS (Ours)

Analysis

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

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

Explanation

Multi-mode is a good explanation (Ours)

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. "Breaking the Softmax Bottleneck: A High-Rank RNN Language Model." In ICLR. 2018.
David Demeter, Gregory Kimmel, and Doug Downey. Stolen probability: A structural weakness of neural language models. In ACL. 2020
Sekitoshi Kanai, Yasuhiro Fujiwara, Yuki Yamanaka, and Shuichi Adachi. Sigsoftmax: Reanalysis of the softmax bottleneck. In NeurIPS 2018
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

BERT Experiment

Input	LM	Ideal Output
After debating whether to bow to the **king** or the **woman** first, the jester decided on the	GPT-2 Small	**king** or **woman**
After debating whether to bow to the **king** or the **woman** first, the jester decided on the [MASK], which makes him pleased.	BERT Base	**king**

- Perplexity improvement of MoS and MFS is much smaller compared to GPT-2
- MFS still doubles the improvement of MoS over Softmax

ProtoQA

Input

Name something that people usually do before they leave for work.

LM

GPT-2 Medium

Ideal Output

Shower 0.43
Breakfast 0.30
Dress 0.07
Lock door 0.07
…

Models ↓	Perplexity on Scraped Development Set	Max Answers Top 1	Max Answers Top 3	Max Answers Top 5	Max Answers Top 10	Max Incorrect Top 1	Max Incorrect Top 3	Max Incorrect Top 5
Softmax (GPT-2)	1.5432 ± 0.0003	34.1 ± 0.8	35.2 ± 0.5	37.8 ± 0.4	45.0 ± 0.5	18.3 ± 0.4	30.7 ± 0.5	38.5 ± 0.6
MoS (Yang et al., 2018) (3)	1.5407 ± 0.0004	33.9 ± 0.8	36.0 ± 0.6	37.7 ± 0.6	44.9 ± 0.4	18.3 ± 0.4	31.7 ± 0.6	38.2 ± 0.6
MFS w/o Multi-partition	1.5411 ± 0.0003	$\mathbf{34.3 \pm 0.7}$	$\mathbf{36.7 \pm 0.7}$	38.1 ± 0.5	45.2 ± 0.4	19.4 ± 0.4	32.0 ± 0.5	38.6 ± 0.3
MFS (Ours)	$\mathbf{1.5402 \pm 0.0005}$	34.1 ± 0.6	$\mathbf{36.7 \pm 0.5}$	$\mathbf{38.6 \pm 0.4}$	$\mathbf{45.4 \pm 0.5}$	$\mathbf{19.7 \pm 0.4}$	$\mathbf{32.1 \pm 0.4}$	$\mathbf{39.7 \pm 0.4}$

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

- $1\,\underline{\boldsymbol{w}_{king}} - 1\,\underline{\boldsymbol{w}_{queen}} = 1\,\underline{\boldsymbol{w}_{man}} - 1\,\underline{\boldsymbol{w}_{woman}}$
- $1\,\underline{\boldsymbol{w}_{king}} + 1\,\underline{\boldsymbol{w}_{woman}} = 1\,\underline{\boldsymbol{w}_{queen}} + 1\,\underline{\boldsymbol{w}_{man}}$

$\times$ $\underline{\boldsymbol{h}}$ (hidden state) on both side

- $\overset{\text{Large}}{1\,\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{king}}} + \overset{\text{Large}}{1\,\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{woman}}} = \overset{\text{Small}}{1\,\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{queen}}} + \overset{\text{Small}}{1\,\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{man}}}$
- If $\exists \underline{\boldsymbol{h}}$, s.t $\min(\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{king}}, \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{woman}}) > \max(\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{queen}}, \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{man}})$
  - $\overset{\text{Large}}{1\,\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{king}}} + \overset{\text{Large}}{1\,\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{woman}}} \geq$
    $2\min(\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{king}}, \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{woman}}) \boxed{>}$
    $2\max(\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{queen}}, \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{man}}) \geq$
    $\underset{\text{Small}}{1\,\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{queen}}} + \underset{\text{Small}}{1\,\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{man}}}\ \boxed{(\rightarrow \leftarrow)}$
- Thus, the logits of LM cannot rank both *king* and *woman* on top of *queen* and *man*

Generalization

- Linearly dependent among $\{\underline{\boldsymbol{w}_{l_1}}, \ldots, \underline{\boldsymbol{w}_{l_L}}, \underline{\boldsymbol{w}_{r_1}}, \ldots, \underline{\boldsymbol{w}_{r_R}}\}$
- $a_{l_1} \underline{\boldsymbol{w}_{l_1}} + \ldots + a_{l_L} \underline{\boldsymbol{w}_{l_L}} = a_{r_1} \underline{\boldsymbol{w}_{r_1}} + \ldots + a_{r_R} \underline{\boldsymbol{w}_{r_R}}$
  - All coefficient $a_{l_i} > 0$, $a_{r_j} > 0$
  - WLOG $a_{l_1} + \ldots + a_{l_L} \geq a_{r_1} + \ldots + a_{r_R}$
- $a_{l_1} \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{l_1}} + \ldots + a_{l_L} \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{l_L}} = a_{r_1} \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{r_1}} + \ldots + a_{r_R} \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{r_R}}$
- If $\exists \underline{\boldsymbol{h}}$, s.t $\min_i(\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{l_i}}) > \max_j(\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{R_j}})$
  - $a_{l_1} \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{l_1}} + \ldots + a_{l_L} \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{l_L}} \geq$
    $(a_{l_1} + \ldots + a_{l_L})\min_i(\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{l_i}}) >$
    $(a_{r_1} + \ldots + a_{r_R})\max_j(\underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{R_j}}) \geq$
    $a_{r_1} \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{r_1}} + \ldots + a_{r_R} \underline{\boldsymbol{h}}^T \underline{\boldsymbol{w}_{r_R}} (\rightarrow \leftarrow)$
- Thus, the logits of LM cannot rank all the left words on top of the right words.