

# Basics in Language Models

**Kun Yuan**

**Center for Machine Learning Research @ Peking University**

Feb. 29, 2024

- Word embedding
- Language models
- Recurrent neural network
- Sequence-to-sequence model

Some of the materials are from a great course [1]

---

[1] The deep learning specialization

# 1-hot word representation

- Vocabulary set = {a, aaron, ..., zulu}; the size is typically on the order of 10,000
- 1-hot representation is the most natural idea to represent word

Man  
(5391)

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

Woman  
(9853)

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

King  
(4914)

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Queen  
(7159)

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

Apple  
(456)

$$\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Orange  
(6527)

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

# 1-hot word representation

- 1-hot representation ignores semantic relationship

I like orange juice.  $\longrightarrow$  I like apple \_\_\_\_.

- “Orange” and “apple” should be close to each other. Language model should fill in “juice”
- But in one-hot representation, apple and orange are not close to each other

Man (5391)	Woman (9853)	King (4914)	Queen (7159)	Apple (456)	Orange (6527)
$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$

# Semantic representation

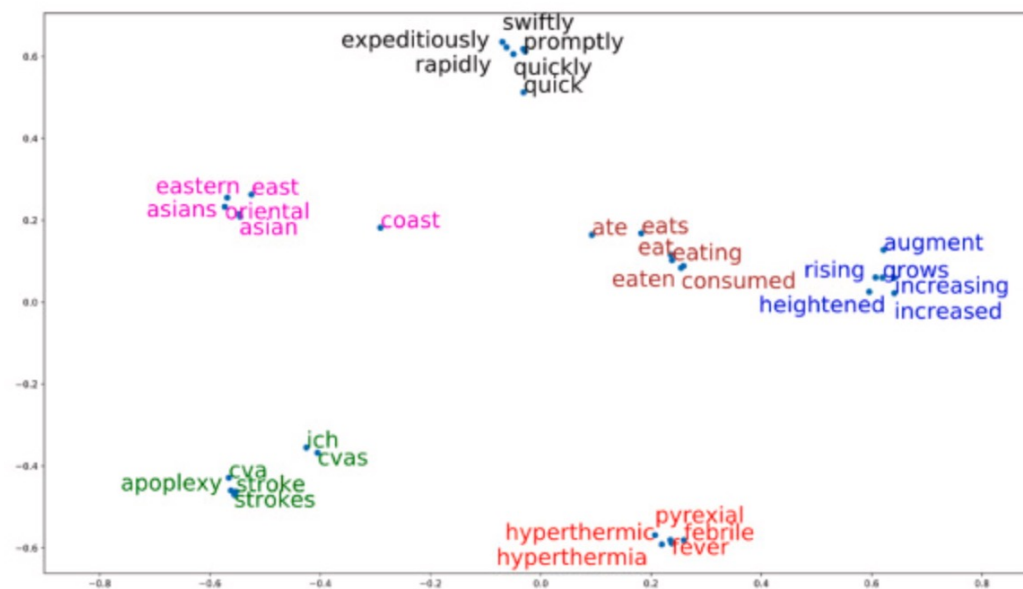
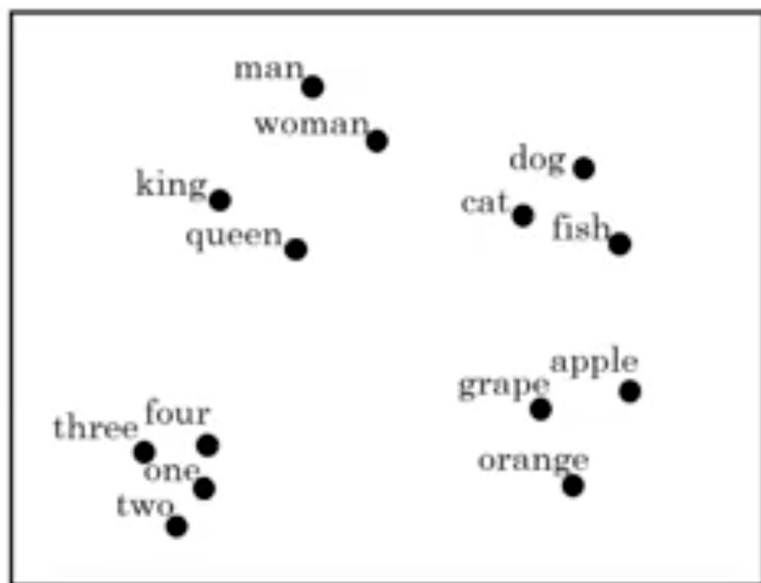
- Each word is represented with vectors that involve semantics

	<b>Man (5391)</b>	<b>Woman (9853)</b>	<b>King (4914)</b>	<b>Queen (7159)</b>	<b>Apple (456)</b>	<b>Orange (6527)</b>
Gender	-1.00	1.00	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.51	0.47	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97

- “Man” is close to “Woman”, “King” is close to “Queen”, and “Apple” is close to “Orange”
- Semantic representation can be much shorter than 1-hot representation

# Semantic representation

- Visualization in semantic representation



- Semantic representation helps the down-stream tasks

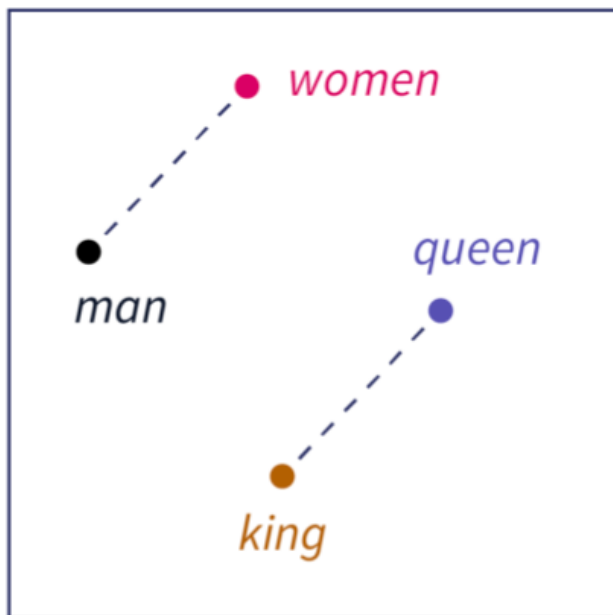
I like orange juice.            I like apple \_\_\_\_.

- Since “orange” and “orange” are close to each other, language model should fill in juice
- [Play semantic games in ChatGPT]



# Semantic representation

- Semantic representation helps find synonyms or antonyms
- Example: given “man” vs “women”, fill in “king” vs “\_\_\_\_\_”



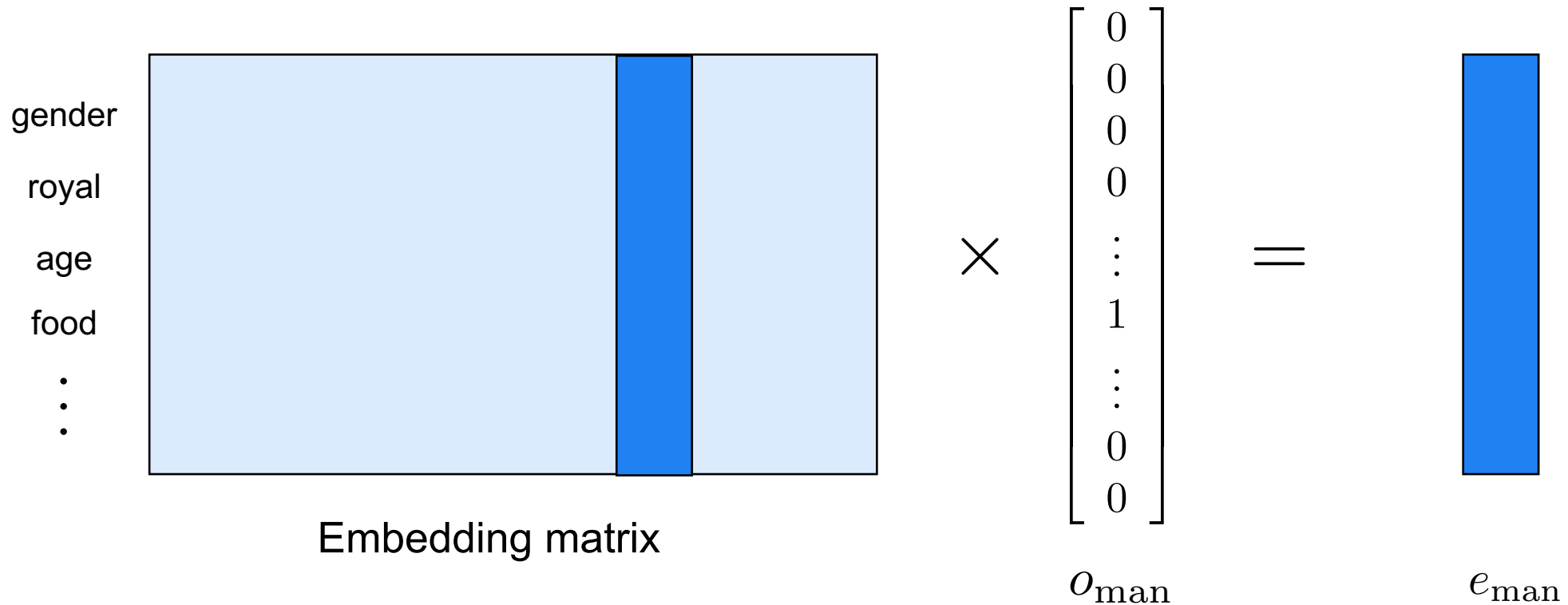
$$e_{\text{man}} - e_{\text{woman}} \approx e_{\text{king}} - e_w$$

$$\longrightarrow e_w \approx e_{\text{king}} - e_{\text{man}} + e_{\text{woman}}$$

$$\longrightarrow w = \arg \max_u \{e_u, e_{\text{king}} - e_{\text{man}} - e_{\text{woman}}\}$$

# 1-hot representation to semantic representation

- Given the embedding matrix, we can easily achieve the semantic representation as follows



The diagram illustrates the process of generating a semantic representation from an embedding matrix and a 1-hot vector. On the left, an "Embedding matrix" is shown as a light blue rectangle with a vertical blue stripe. To its left, the words "gender", "royal", "age", "food", and a vertical ellipsis are listed, corresponding to the rows of the matrix. In the center, a multiplication symbol  $\times$  is followed by a vertical vector  $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$ , which is labeled  $o_{\text{man}}$  below it. To the right of the vector is an equals sign  $=$ , followed by a vertical blue rectangle labeled  $e_{\text{man}}$  below it.

$$\begin{bmatrix} \text{gender} \\ \text{royal} \\ \text{age} \\ \text{food} \\ \vdots \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \vdots \end{bmatrix}$$

Embedding matrix  $o_{\text{man}}$   $e_{\text{man}}$

## Embedding matrix

- How to get the embedding matrix?
- We first collect the dataset from the corpus. Many ways; we use the simplest one to highlight the idea
- Given any word in a sentence, find a nearby (say, with window 2) word to construct the word pair

I want a glass of orange juice



(orange, juice)

(orange, glass)

He likes watching TV



(watching, TV)

(watching, likes)

corpus

# dataset

# Embedding matrix

- Given the dataset, we use the first word to predict the second word

