NATURAL LANGUAGE PROCESSING (PRACTICE)

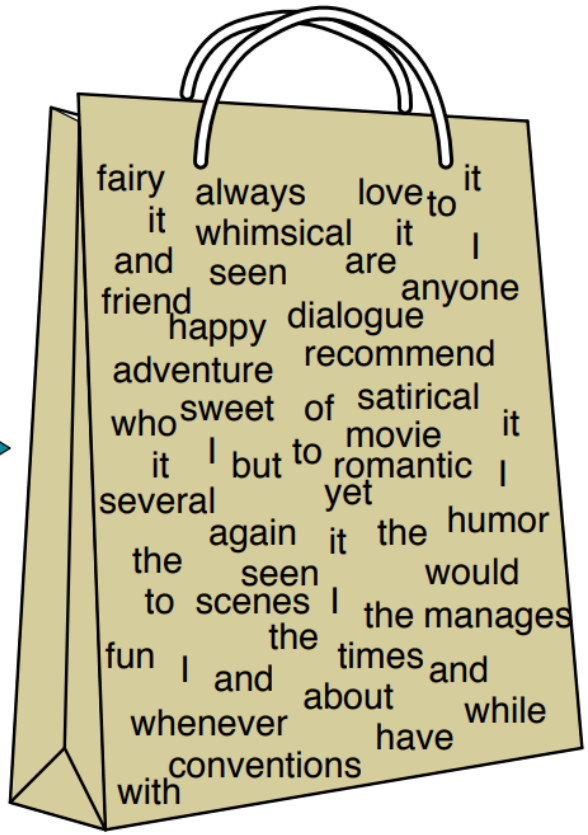
NLP 251 - Lab 5: Word Embeddings

Department of Computer Science and Engineering
Ho Chi Minh University of Technology, VNU-HCM

Review Knowledge

Bag of Words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Co-occurrence Matrix

- I like deep learning.
- I like NLP.
- I enjoy flying.

	I	like	enjoy	deep	learning	NLP	flying
I	0	2	1	0	0	0	0
like	2	0	1	0	1	1	0
enjoy	1	1	0	0	0	0	1
deep	0	0	0	0	1	0	0
learning	0	1	0	1	0	0	0
NLP	0	1	0	0	0	0	1
flying	0	0	1	0	0	1	0

TF-IDF: Weighing terms in the vector

$$w_{i,j} = \mathbf{tf}_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$$tf_{i,j} = \text{number of occurrences of } i \text{ in } j$$
$$df_i = \text{number of documents containing } i$$
$$N = \text{total number of documents}$$

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

One-hot Vector

For Example: Man, woman, boy, girl, prince, princess, queen, king, monarch.

Each word gets a 1×9 vector representation:

	1	2	3	4	5	6	7	8	9
Man	1	0	0	0	0	0	0	0	0
Woman	0	1	0	0	0	0	0	0	0
Boy	0	0	1	0	0	0	0	0	0
Girl	0	0	0	1	0	0	0	0	0
Prince	0	0	0	0	1	0	0	0	0
Princess	0	0	0	0	0	1	0	0	0
Queen	0	0	0	0	0	0	1	0	0
King	0	0	0	0	0	0	0	1	0
Monarch	0	0	0	0	0	0	0	0	1

Table: One-hot encoding representation of words.

Word2vec

Sparse versus dense vectors

tf-idf (or PMI) vectors are:

- **long** ($|V| = 20,000$ to $50,000$)
- **sparse** (most elements are zero)

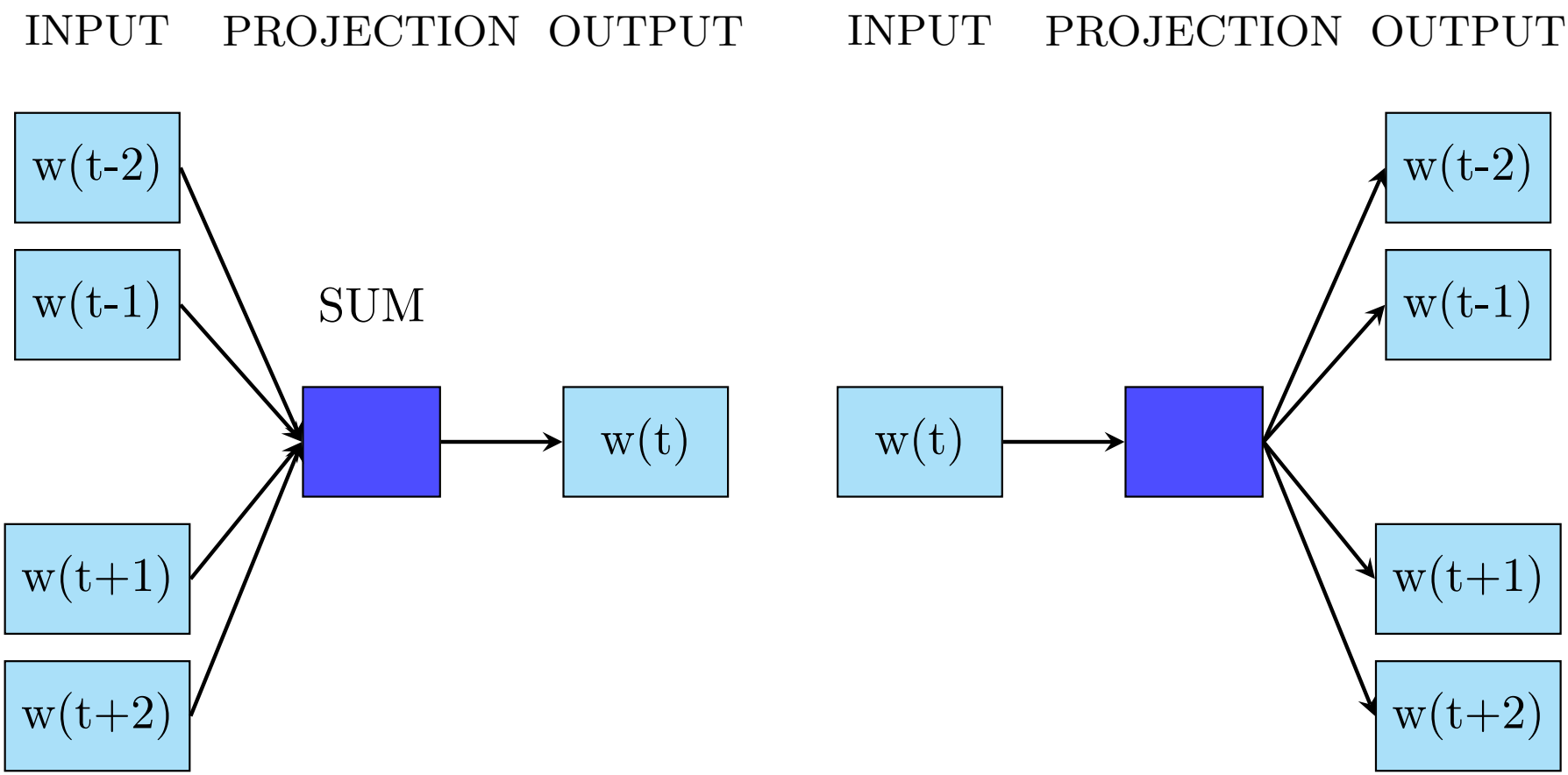
Alternative: learn vectors are:

- **short** (length 50-1000)
- **dense** (most elements are non-zero)

Why dense vectors?

- Short vectors may be easier to use as **features** in machine learning (fewer weights to tune)
- Dense vectors may **generalize** better than explicit counts
- Dense vectors may do better at capturing synonymy:
  - *car* and *automobile* are synonyms; but are distinct dimensions
  - a word with *car* as a neighbor and a word with *automobile* as a neighbor should be similar, but aren't
- **In practice, they work better**

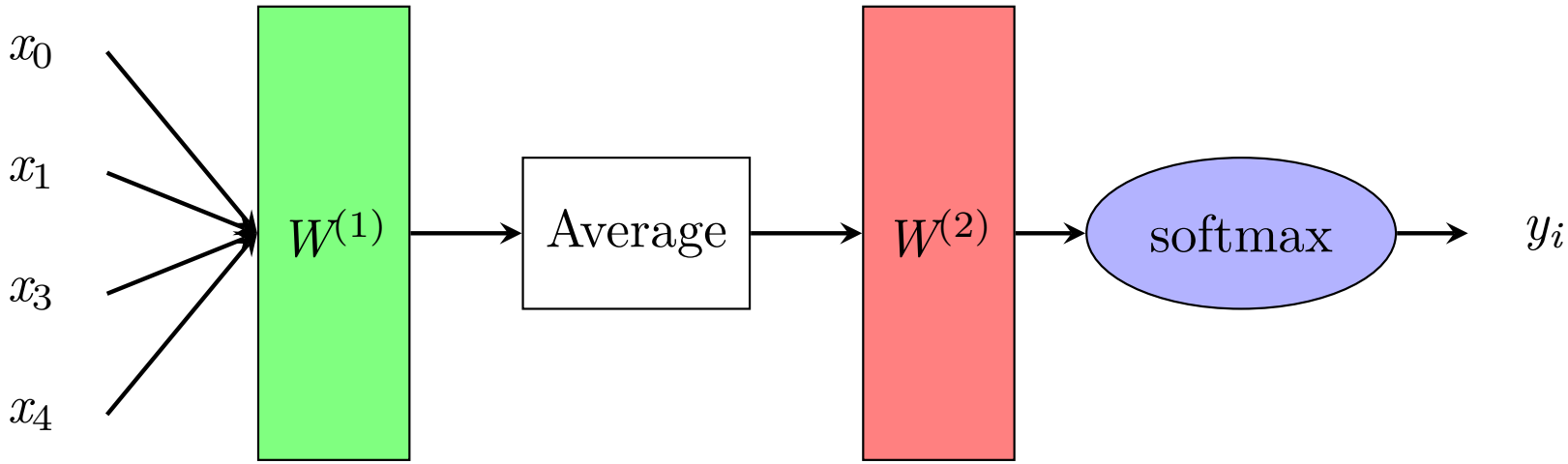
CBOW and Skip-gram



(a) CBOW (b) Skip-gram

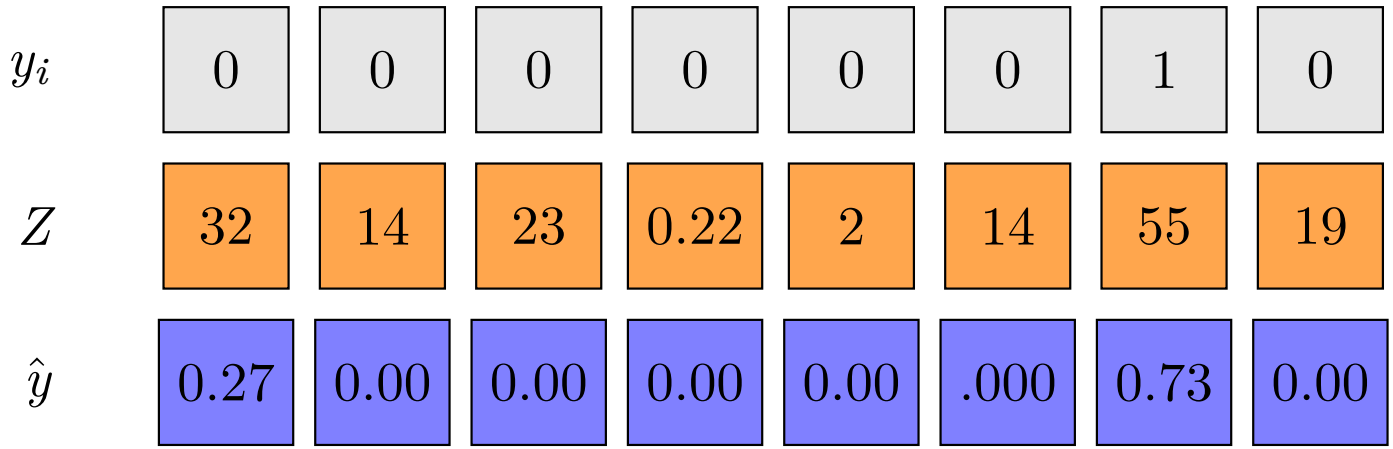
Efficient Estimation of Word Representations in Vector Space (Tomas et al., 2013)

CBOW Model



- $|V|$ is the vocabulary size.
- $x_i \in \mathbb{R}^{1\times|V|}$ is the one-hot representation for a word.
- $y_i \in \mathbb{R}^{|V|\times 1}$ is the one-hot representation for the correct word, in the middle (the target word).

CBOW Model - Softmax Function

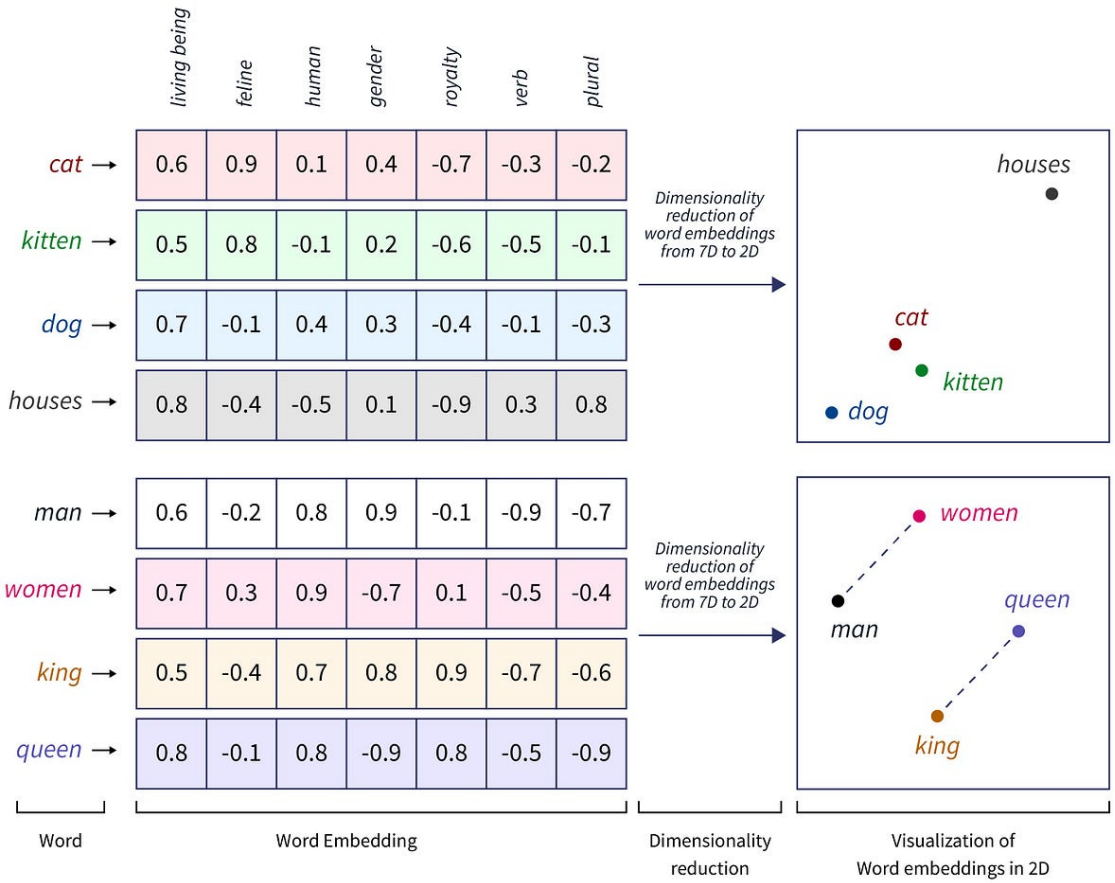


$$\hat{y} = \text{softmax}(Z) \quad \text{and} \quad \sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}};$$

$y_i \in \mathbb{R}^{|V| \times 1}$ is the one-hot representation for the correct word, in the middle (the target word).

Word Vectors

After the training process, each row or column in the matrix $W^{(1)}$ contains the values of words and Word representation vectors better capture syntactic and semantic aspects, Achieves good results in NLP tasks.



Other Representation Methods

Glove

- **Advantages:** Fast training, Stable quality on large datasets, Good results even with small datasets and vectors.
- **Disadvantages:** Requires a lot of memory and Can be affected by the initialization process of the ”learning rate”.

FastText

- **Advantages:** Performs well with rare words due to the use of n-grams
- **Disadvantages:** Requires more memory due to character-based storage

ELMo

- **Advantages:** Better semantic representation compared to word2vec and Glove. (Two similar words may have different vectors depending on the context) and Handles rare word issues well due to character-based representation.
- **Disadvantages:** Requires more memory due to character-based storage.

THANKS FOR LISTENING!