Machine Learning Applications

Word Embeddings

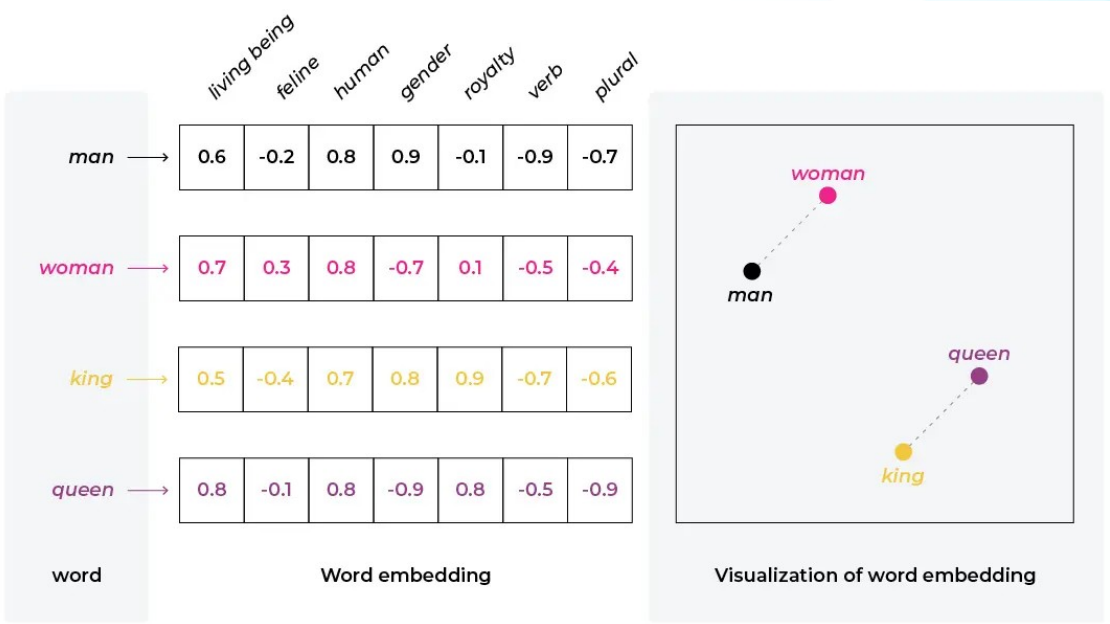
Tan Cher Wah (cherwah@nus.edu.sg)

Word Embeddings

- Word Embeddings represent **words** as **numerical vectors** in a high-dimensional space
- By measuring the **distance** between **word vectors**, we can determine how similar or different words are. For example, "cat" and "dog" might be closer together than "cat" and "apple"
- Word embeddings can be used as input to various machine learning models, such as neural networks, for tasks like text classification, sentiment analysis, and machine translation

Word Embeddings

- Word-Embeddings are **unique** and of **equal length**
- Each dimension of a word-embedding represents a **feature** of the word's representation (e.g. gender)
- Words with **similar semantics** have **similar embeddings**

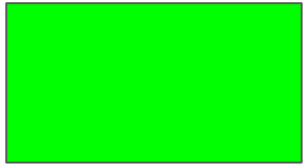


Representing Colors with Embeddings

- **RGBA**-embedding describes the intensity of each color component
- The **A**lpha value blends the color with whatever is behind it, affecting how much of the underlying content is visible
- More dimensions, more information

[199, 21, 133]

[55, 192, 203]



[0, 255, 0]

[199, 21, 133, 102]

Alpha = **0** denotes **full transparency**
Alpha = **255** denotes **full opacity**

Why not One-Hot Encoding?

- One-hot encoding creates a vector as long as the **vocabulary size**, leading to very high-dimensional representations, which can be inefficient for large vocabularies
- One-hot encoded vectors are **sparse**, meaning they contain mostly zeros. This can lead to inefficiencies in storage and computation, especially when dealing with large datasets
- One-hot encoding does not capture any **semantic relationships** between words. Each word is treated as independent, losing context and meaning

Sample Training Data

- Consider the string “Mary has a little lamb”
- A context size of 1 is to consider 1 word left and word right with respect to the **target** word

Context Size = 1	Target Word	Context
[Mary has]	Mary	has
[Mary has a]	has	Mary, a
[has a little]	a	has, little
[a little lamb]	little	a, lamb
[little lamb]	lamb	little

Skip Gram

- The **Input** is the one-hot encoding of the **target word**, and the **Output** (**label**) is the one-hot encoding of **context words**
- At the end of training, the **Embedding Matrix** (in **red**) has our learned **word-embeddings**
- Each **ith row** of the Embedding Matrix corresponds to the **ith word** in our **vocabulary**

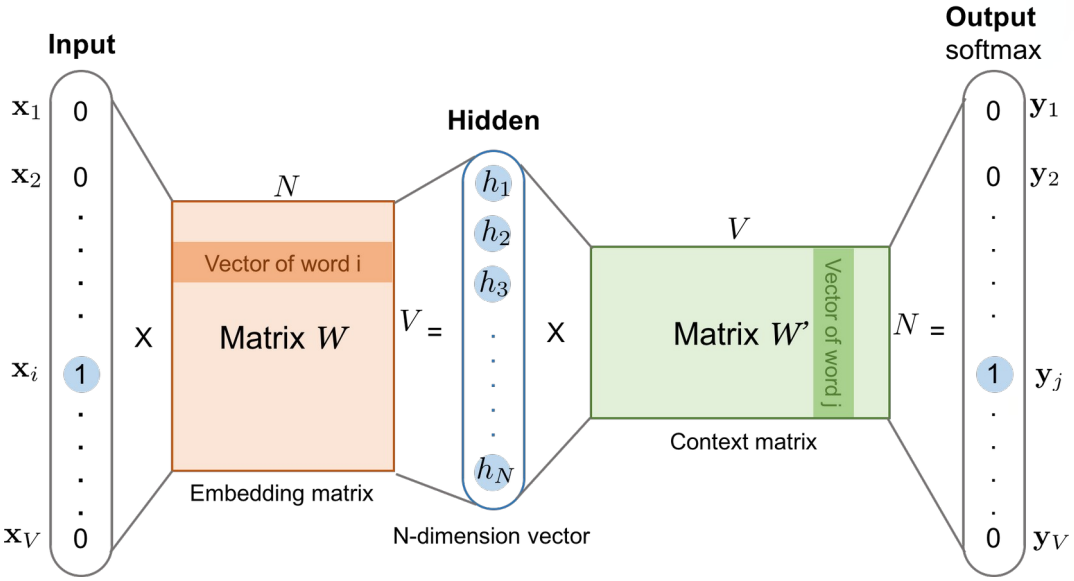


Image: https://lilianweng.github.io/posts/2017-10-15-word-embedding/

Continuous Bag-of-Words (CBOW)

- The **Input** is a **stacked** of one-hot encoded of **context words**, and the **Output** (**label**) is the one-hot encoding of the **target word**
- At the end of training, **Matrix W'** (in **green**) has our learned **word-embeddings**
- Each **ith column** of Matrix W' corresponds to the **ith word** in our **vocabulary**

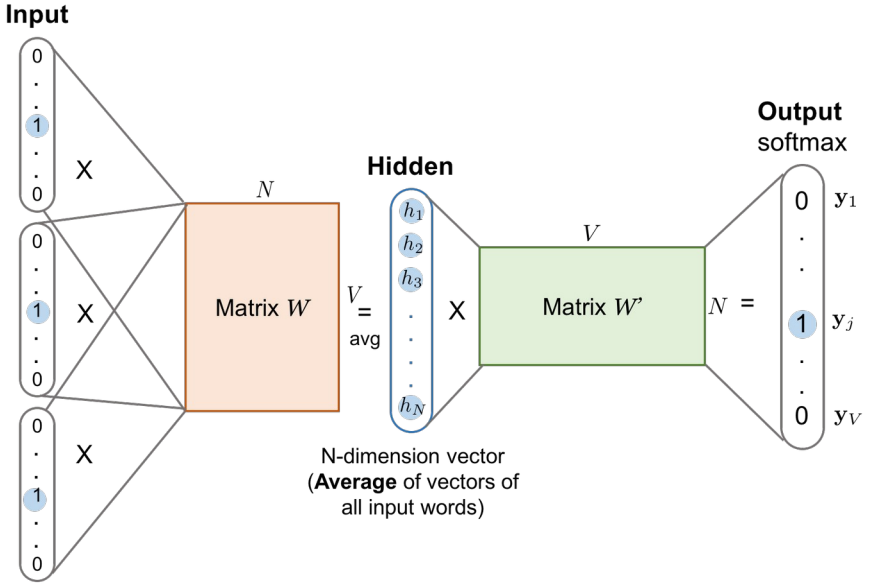


Image: https://lilianweng.github.io/posts/2017-10-15-word-embedding/

The End