National Research
Tomsk
State
University

Natural Language Processing. Embeddings. Part 3
(A Very Short Introduction)

Sergey V. Axyonov,

PhD, Associate Professor,

Department of Fundamental Computer Science, Institute of Applied Math & Computer Science,

Tomsk State University

Tomsk-2024

Summary from the Previous Presentations

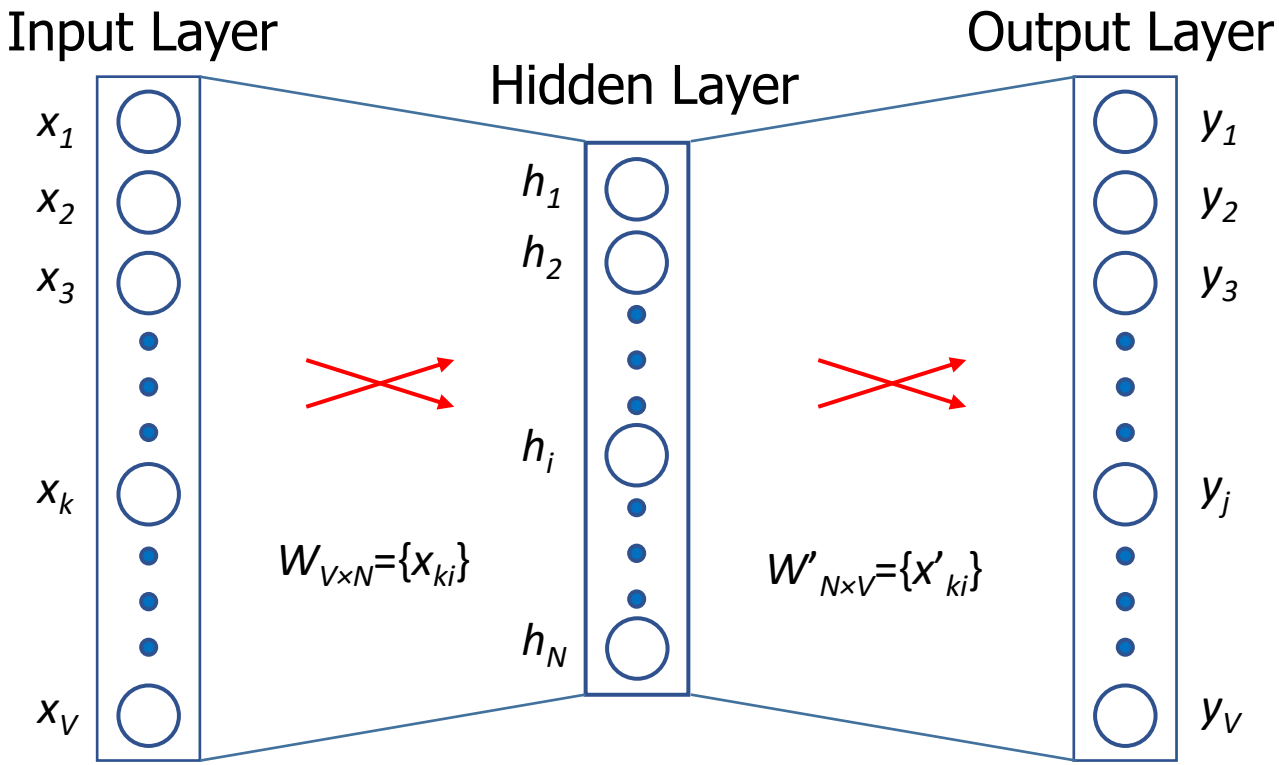
1. Neural networks can detect difficult dependencies in our data.

2. If we convert text data into numbers we can find relationships between tokens.

3. We can classify texts, make images and videos from text description and vice versa with this tool.

The Next Word Prediction

- The task of predicting a word from the previous word
- Input: one-hot vector of the previous word
- Output: the probability of the next word
- *W* или *W*– Embedding Matrices



V - The Size of Vocabulary

N - Embedding size

The Context (Surrounding Tokens)

"There was a table set out under a tree in front of the house, and the March Hare and the Hatter were having tea at it..." ("Alice's Adventures in Wonderland" by Lewis Carrol)

There was a **table** set out under $m = 3$

Hare and the Hatter **were** having tea at it $m = 4$

table set out under a **tree** in front of the house $m = 5$

m - The number of words to the left and right of the target word.

Some Examples. Same Contexts

The kitten watched the fish.
My kitten is drinking milk from the bowl.
The kitten slept cutely in that basket.
Ellie and Anna played with a kitten in the bedroom.

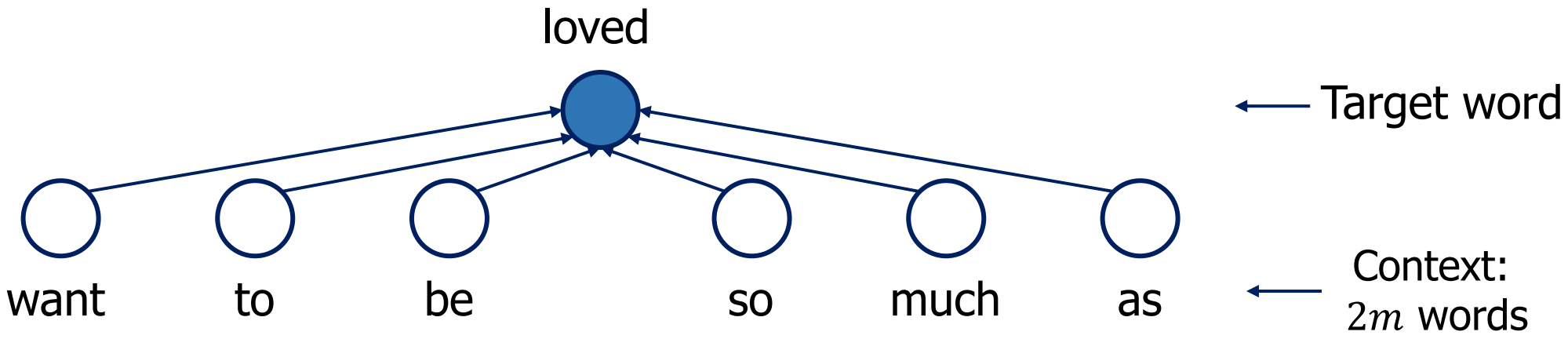
The cat watched the fish.
My cat is drinking milk from the bowl.
The cat slept cutely in that basket.
Ellie and Anna played with a cat in the bedroom.

Marta saw a beautiful bag in a store window.
She was the most beautiful bride in the world.

Martha saw a glamorous bag in the store window.
She was the most charming bride in the world.

Word2Vec: Continuous Bag-of-Words

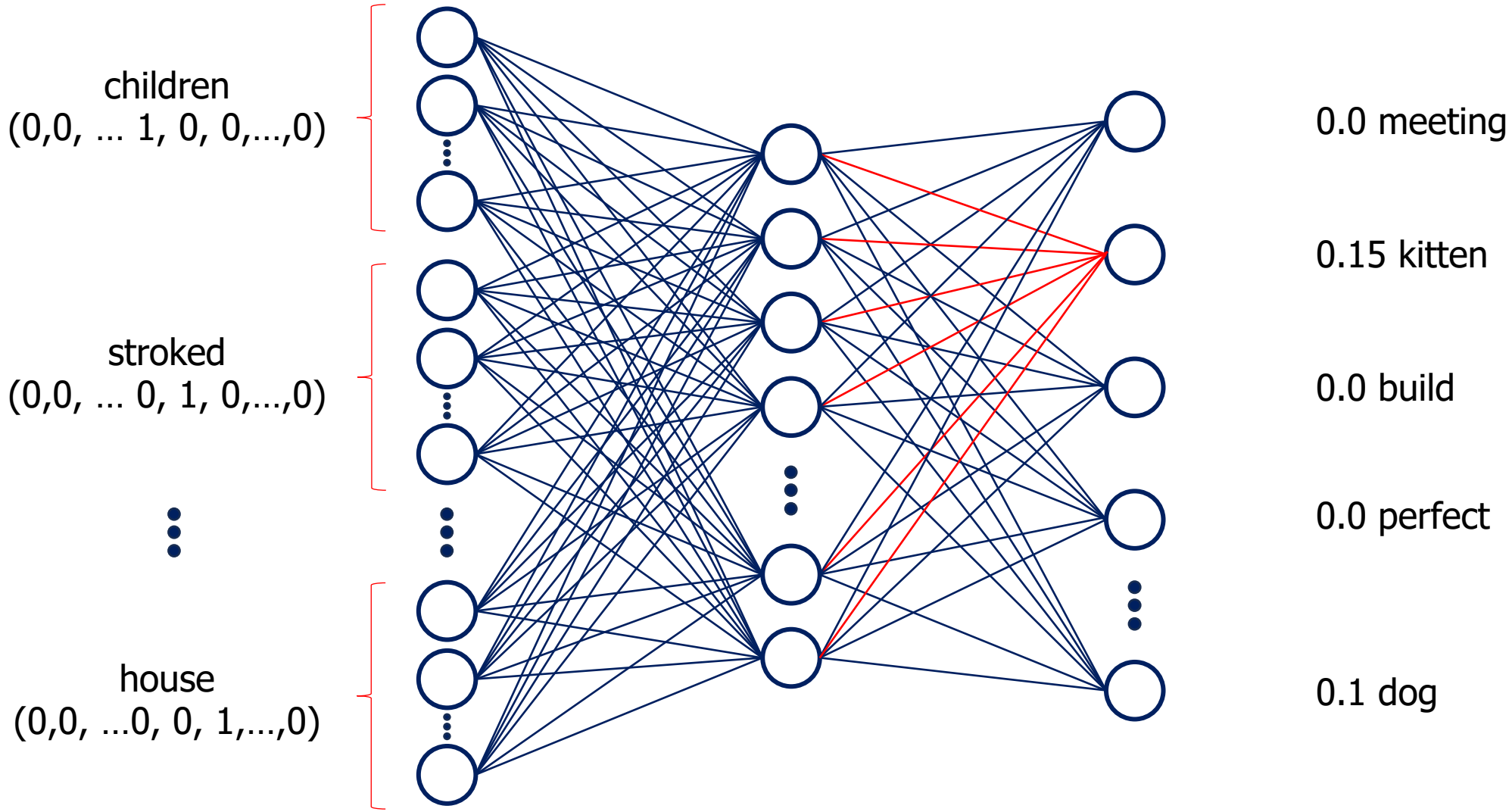
"Perhaps one did not want to be loved so much as to be understood." ("Nineteen Eighty-Four" by George Orwell)



The CBOW model architecture tries to predict the current target word (the center word) based on the source context words.

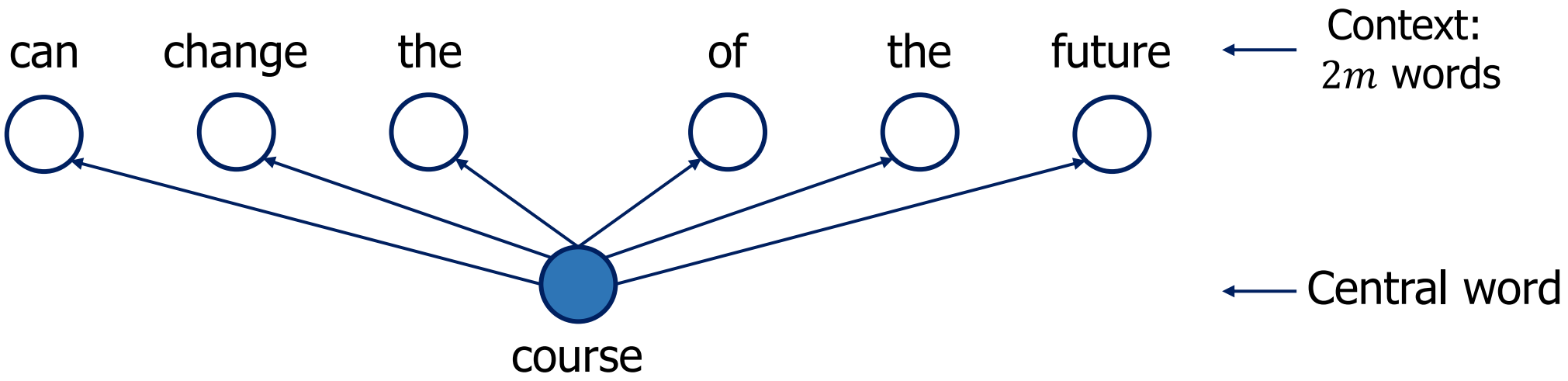
Continuous Bag-of-Words: Weights

Children stroked a lovely fluffy kitten on the lawn near the house.



Word2Vec: Skip-Gram

"Even the smallest person can change the course of the future." ("The Lord of the Rings" by John Ronald Reuel Tolkien)



It aims to predict the surrounding context words given a target word and captures semantic relationships between words.

FastText

fastText will break it into its n-gram components.

Example:

1. The bill has been sent back to comittee.
2. The decision was made by all the Northwest Wildlife Commitee members.
3. Jane chaired the Secondary Education Committe.

'comittee' → 'co', 'mi', 'tt', 'ee'
'committee' → 'co', 'mm', 'it', 'te', 'e'
'committe' → 'co', 'mm', 'it', 'te'
'committee' → 'committee', 'co', 'mm', 'it', 'te', 'e'

fastText learns the weights for every n-gram along with the entire word token. Each token/word will be expressed as the sum and an average of its n-gram components.

Embeddings

A latent space or embedding space is an embedding of a set of items within a manifold in which items resembling each other are positioned closer to one another.
Position within the latent space can be viewed as being defined by a set of latent variables that emerge from the resemblances from the objects.

As a rule of thumb, a dataset with less than 100,000 sentences may benefit from a lower-dimensional embedding (e.g., 50-100 dimensions), while a larger dataset may benefit from a higher-dimensional embedding (e.g., 200-300 dimensions).

A higher dimensional embedding can capture fine-grained relationships between words, but can take more data to learn.

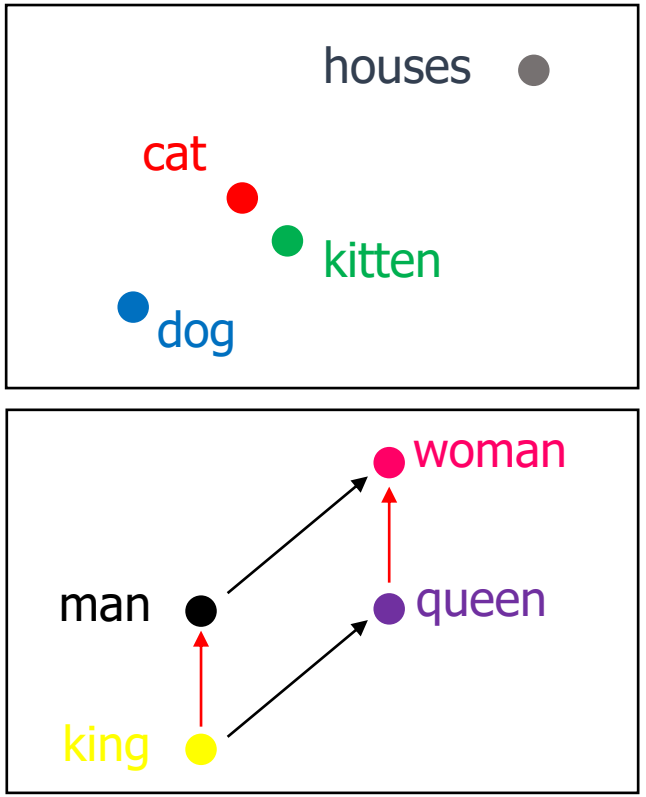
OpenAI has announced 2 new models, text-embedding-3-small and text-embedding-3-large, providing various dimensions 512 and 1536 and respectively 256, 1024 and 3072. (Jan 2024)

Embeddings: An Example '$Queen - Woman = King - Man$'

	living being	feline	human	gender	royalty	verb	plural
cat	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
kitten	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
dog	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
houses	-0.8	-0.4	-0.5	0.1	-0.9	-0.3	0.8
man	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
woman	0.7	0.3	0.9	-0.7	0.1	-0.5	-0.4
king	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
queen	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

7D→2D

7D→2D



Word

Embedding Vector

Dimensionality Reduction

2D Representations