Word Vectorization

- Implementing “word2vec”
using PyTorch -

Brought to you by

Md. Ishtiak Hossain
011 152 056

Rifaz Nahiyan
011 152 054

Md. Golam Farzad
011 152 0548

A. M. Ishtiaque Ali
011 152 059

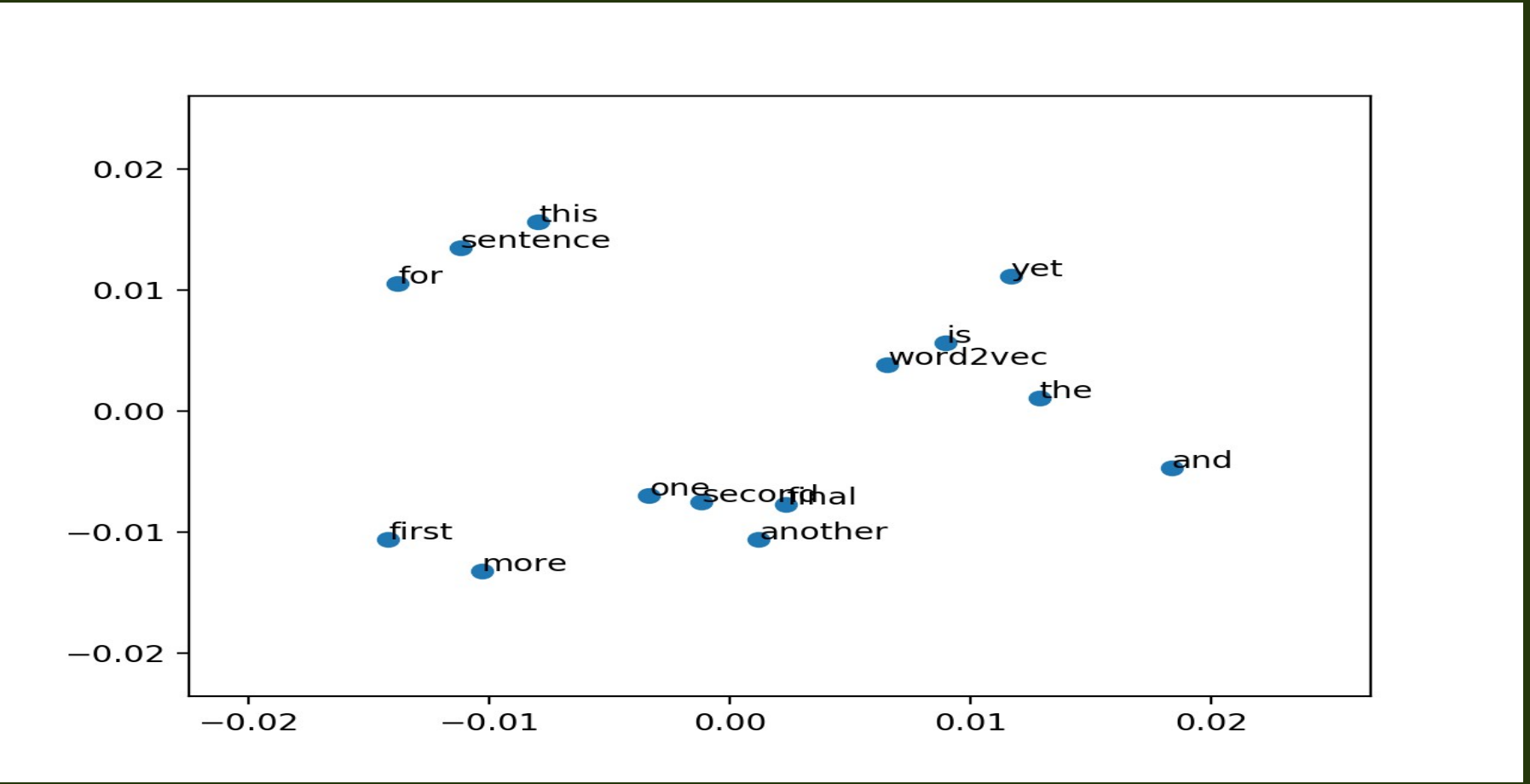
Word Vectorization

… what?

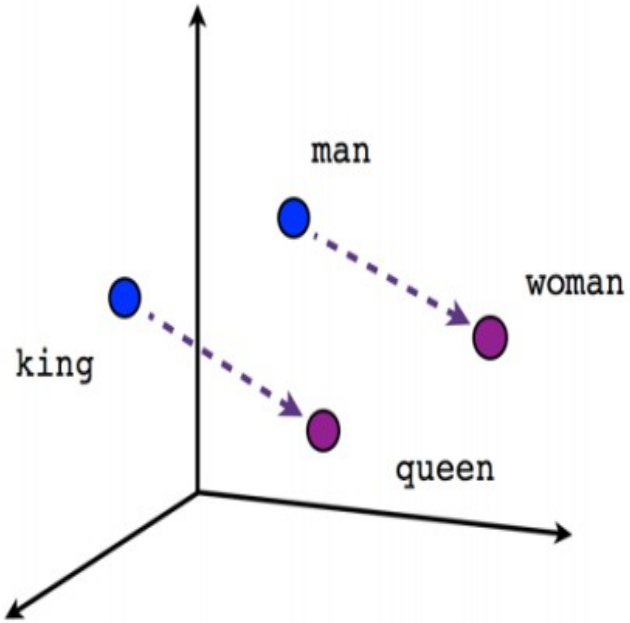
*The alchemy of making **very useful** vectors from words*

Why?

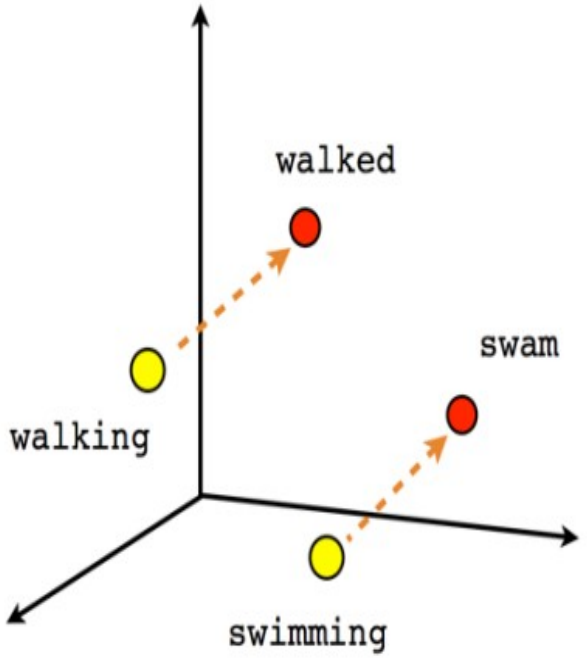
So that words can float in a geometric plane...



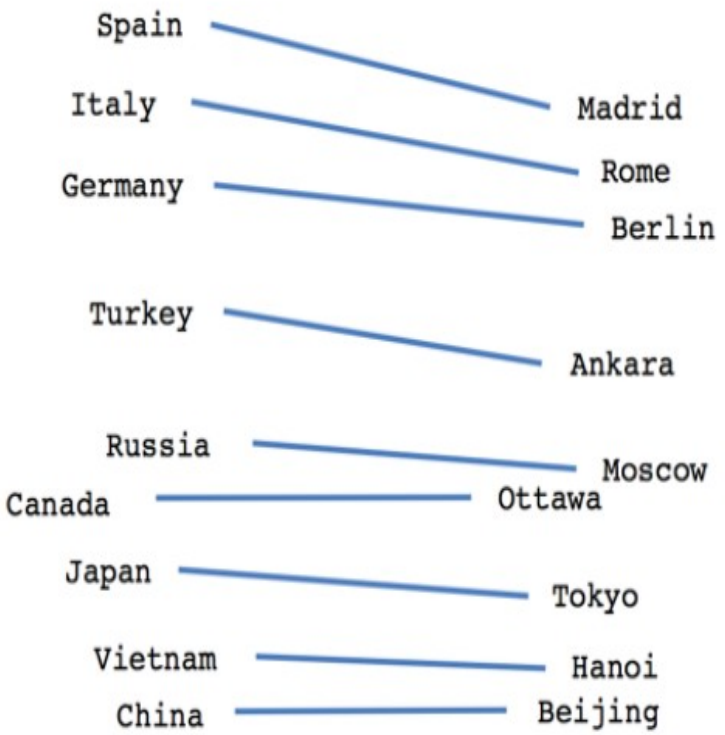
But why...



Male-Female



Verb tense



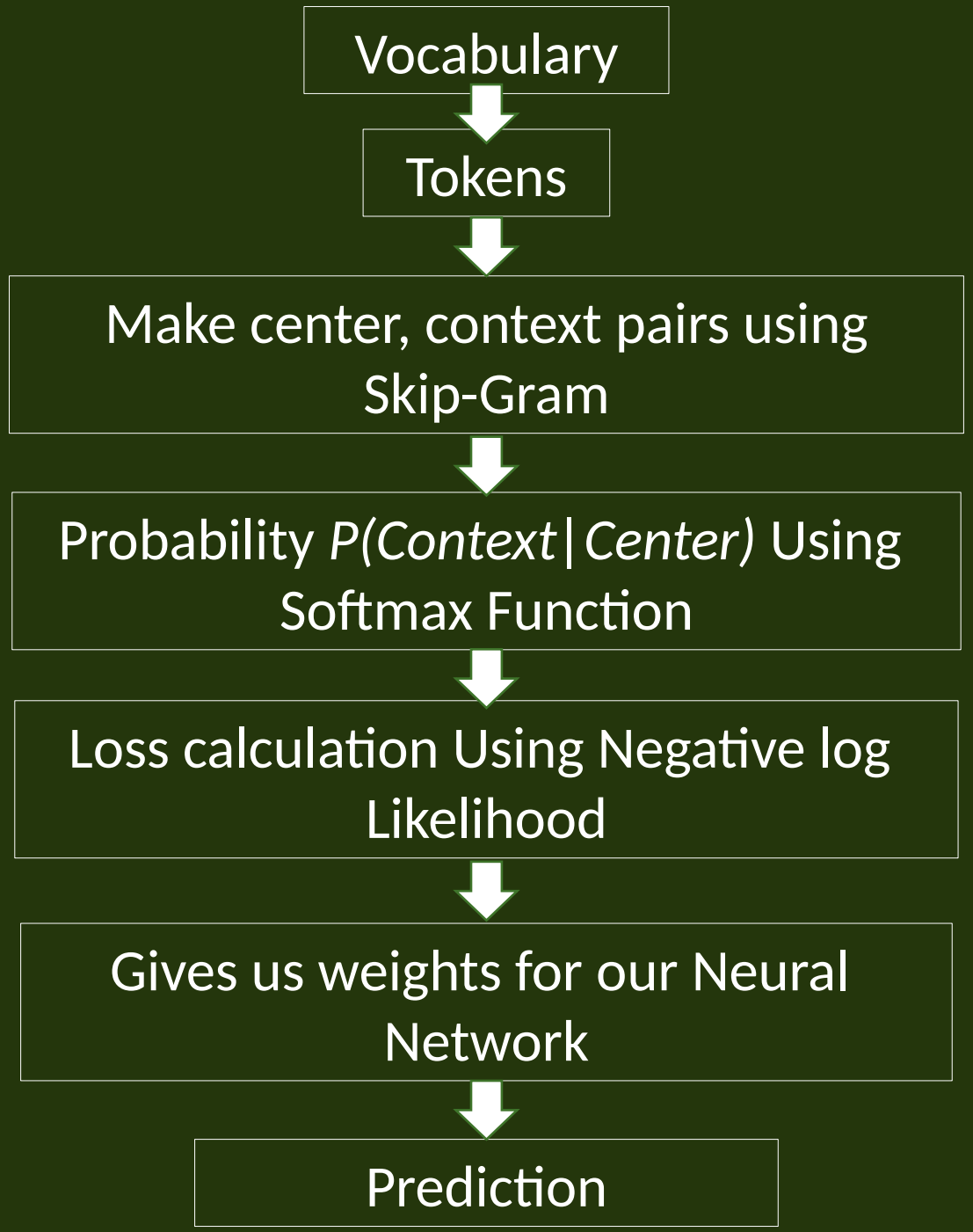
Country-Capital

Also because words are powerful

And when they can be used at the speed of computers they become even more powerful.

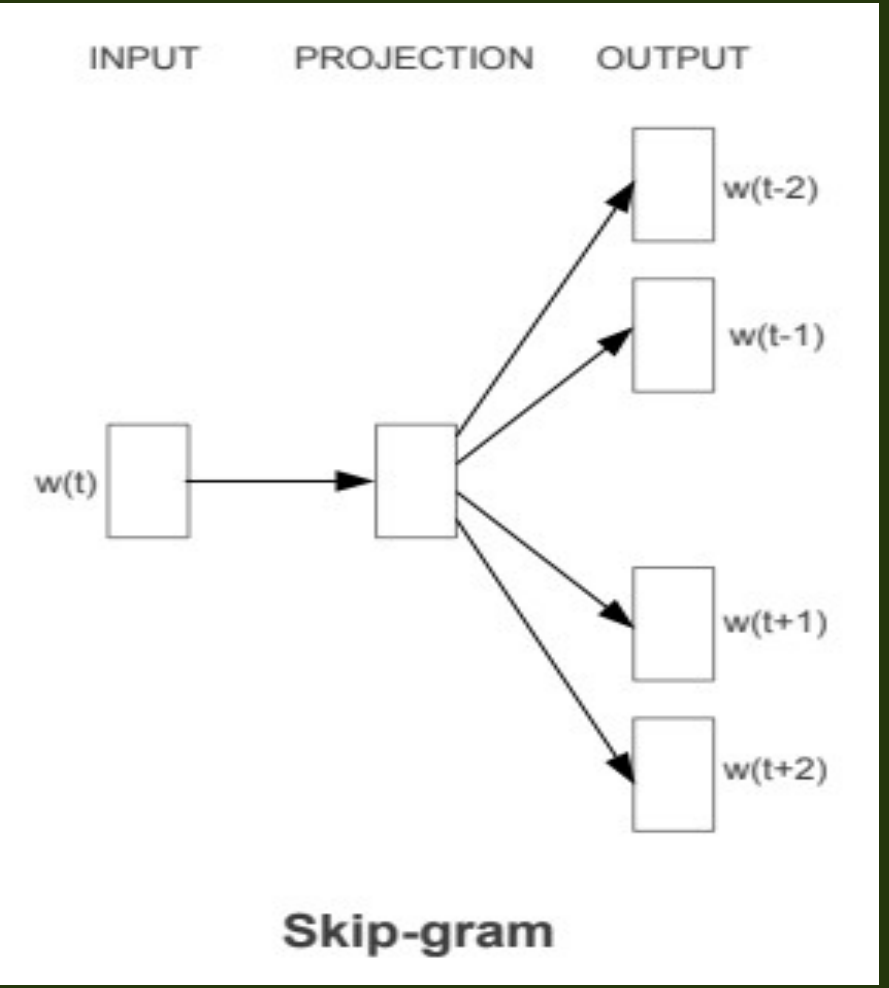
With all that covered…
Now we get to the details

Converting Words to Vectors


Vocabulary
Tokens
Make center, context pairs using Skip-Gram
Probability *P(Context|Center)* Using Softmax Function
Loss calculation Using Negative log Likelihood
Gives us weights for our Neural Network
Prediction

Models:
Skip-Gram

Skip-gram gives you the surrounding words to a given center word.



Skip-gram closer look

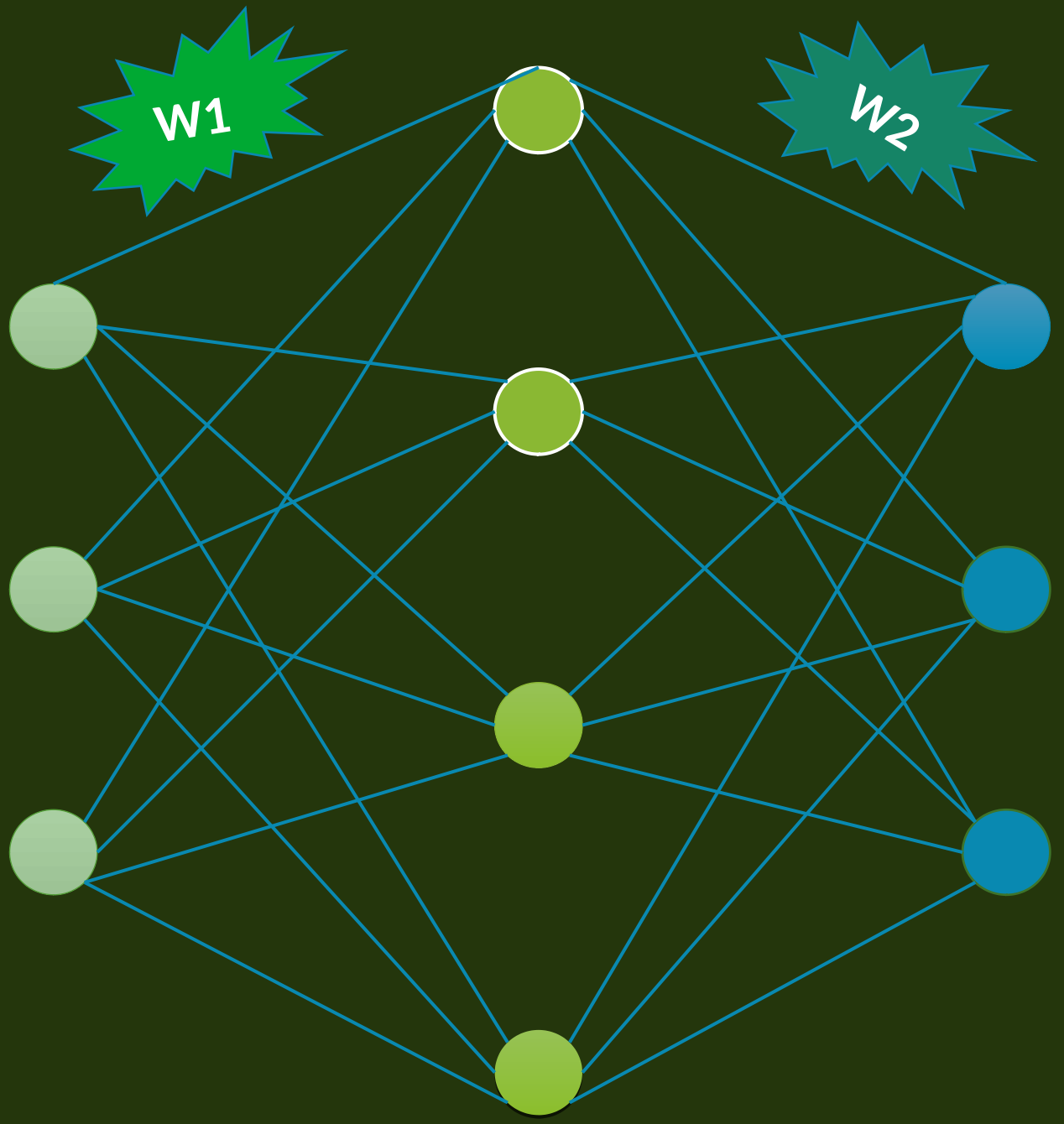
Probability P(Context|Center) Using Softmax:

$$P(\text{context}|\text{center}) = \frac{\exp(u_{\text{context}}^T v_{\text{center}})}{\sum_{w \in \text{vocab}} \exp(u_w^T v_{\text{center}})}$$

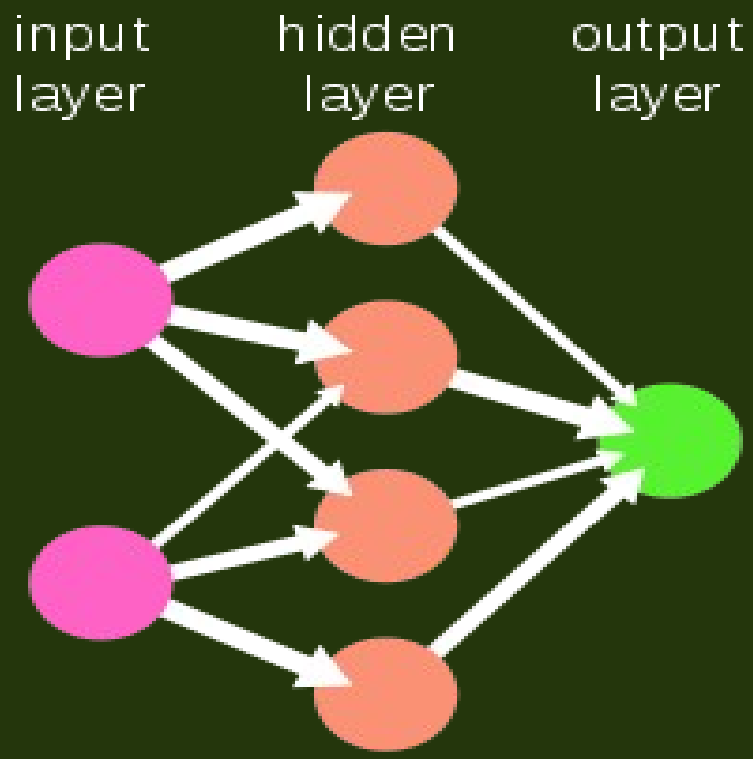
Loss function: Negative Log Likelihood

$$\text{loss} = -\frac{1}{T} \sum_{\text{center}} \sum_{\text{context}} \log P(\text{context} | \text{center}, \theta)$$

And the process repeats
to give us weights like…


W1
W2

A simple neural network



Now comes the coding part

The code

Python/numpy data structures like:

- Numpy ndarray
- Lists
- Dictionaries
- PyTorch Tensors

Python/numpy data structures like:

- Numpy ndarray
- Lists
- Dictionaries
- PyTorch Tensors ...?

PyTorch

Extremely powerful library for

- Data Science
- Neural Networks
- Provides Tensors

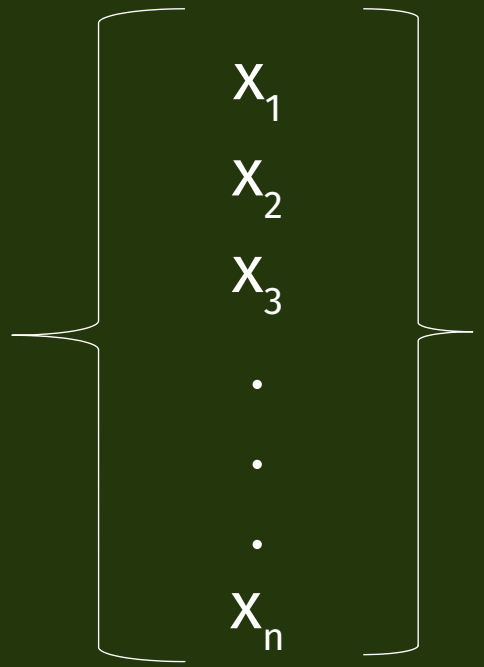
Can use also ~~steroids~~ GPU

Tensors

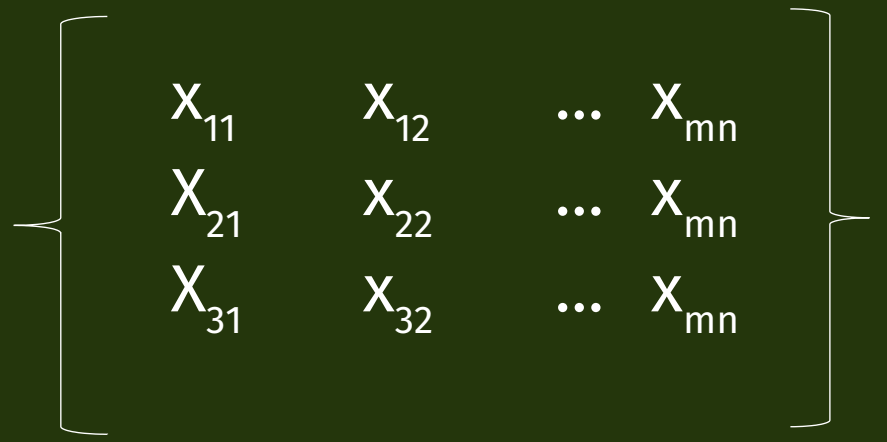
~~The muscles for PyTorch~~

Basically... multi-dimensional vectors

Vectors



Tensor



Demo

Thank you for your patience…

Questions?

We prefer none