IS883: Deploying Generative AI

Mohannad Elhamod

Neural Networks

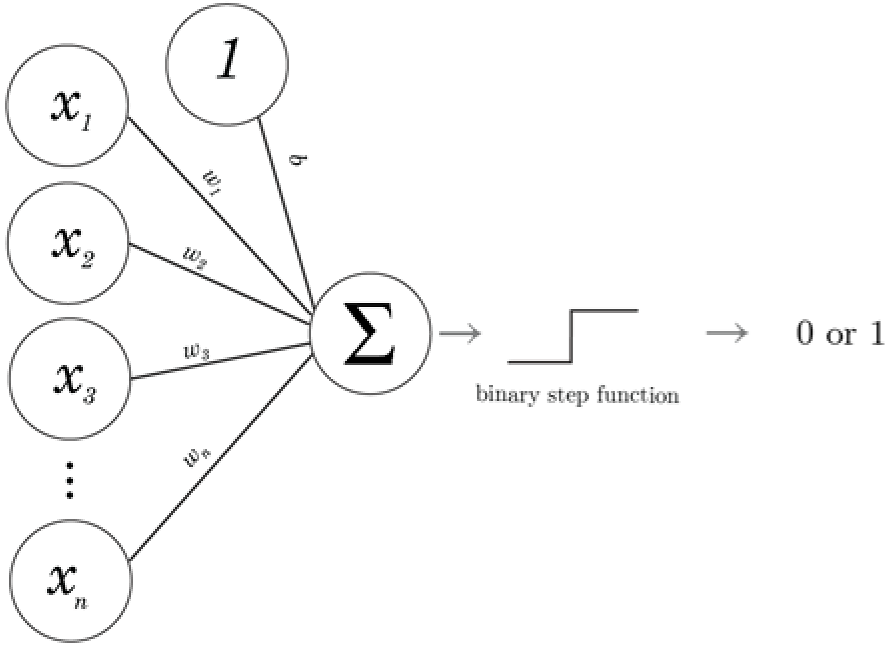
BOSTON UNIVERSITY

Boston University Questrom School of Business

The building block: The Perceptron

- Demo

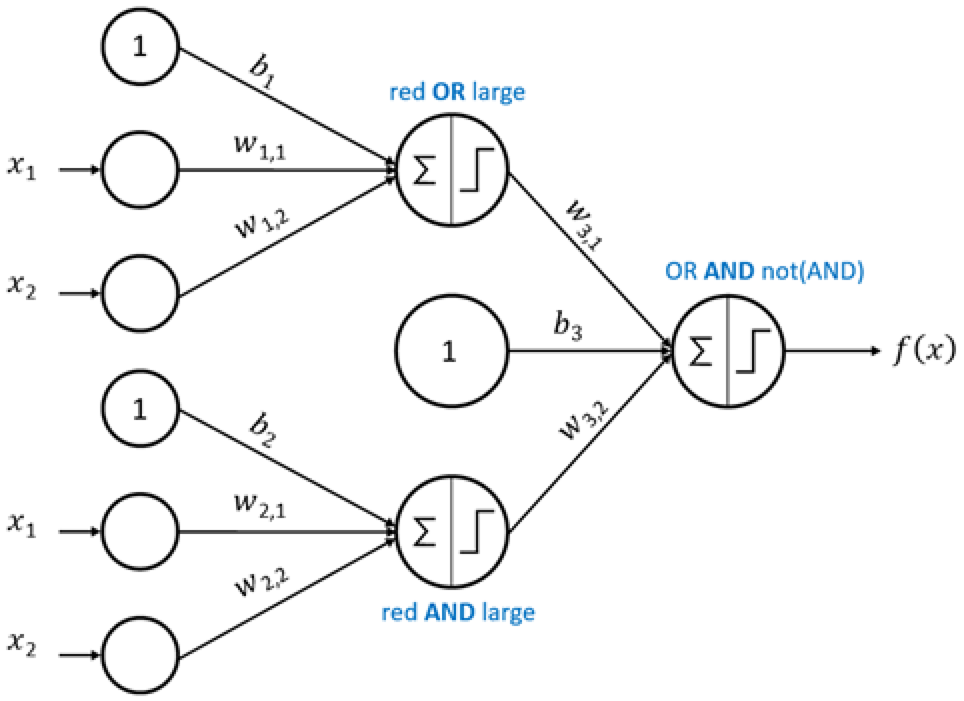
- Can we work around the “linear separability” issue”?



Adam Dhala

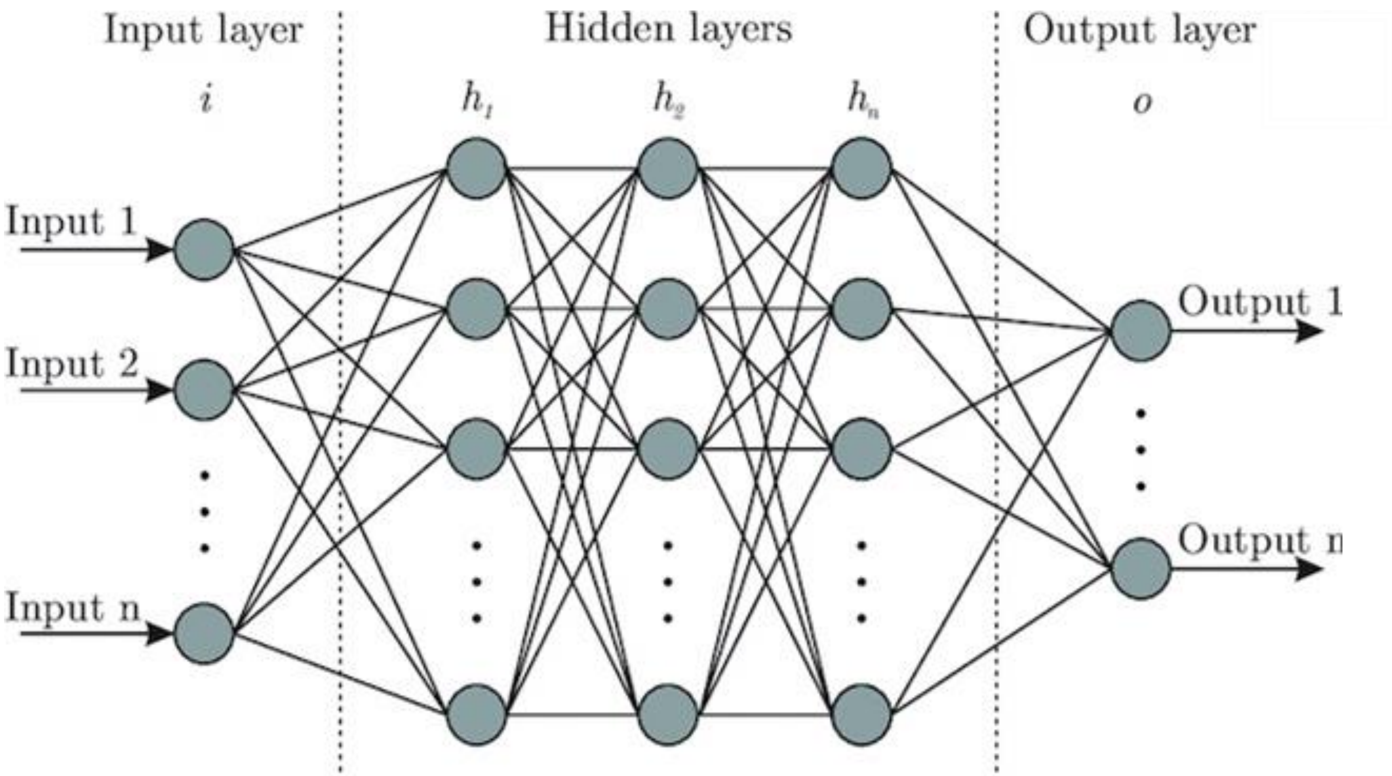
Power in Numbers: Multiple Perceptrons

- Demo



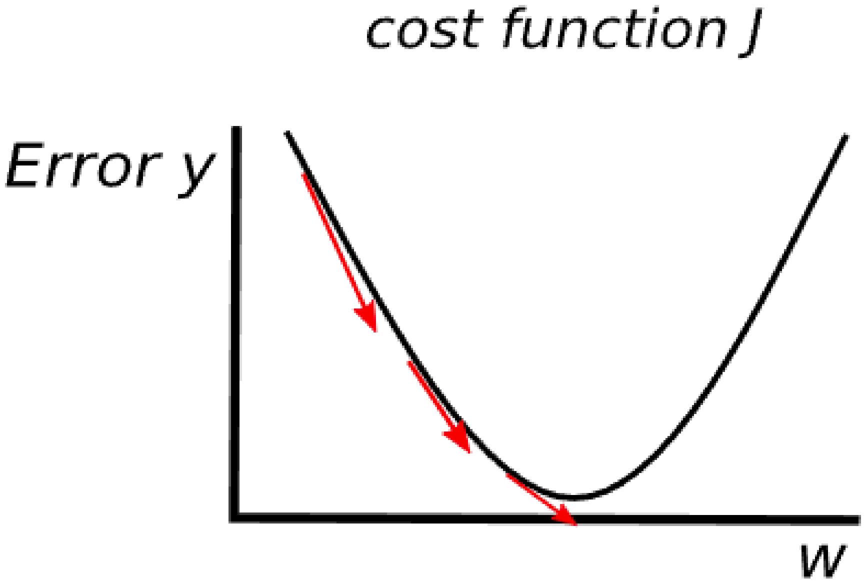
western-neuralnets.ca

Neural Newtorks



Optimization

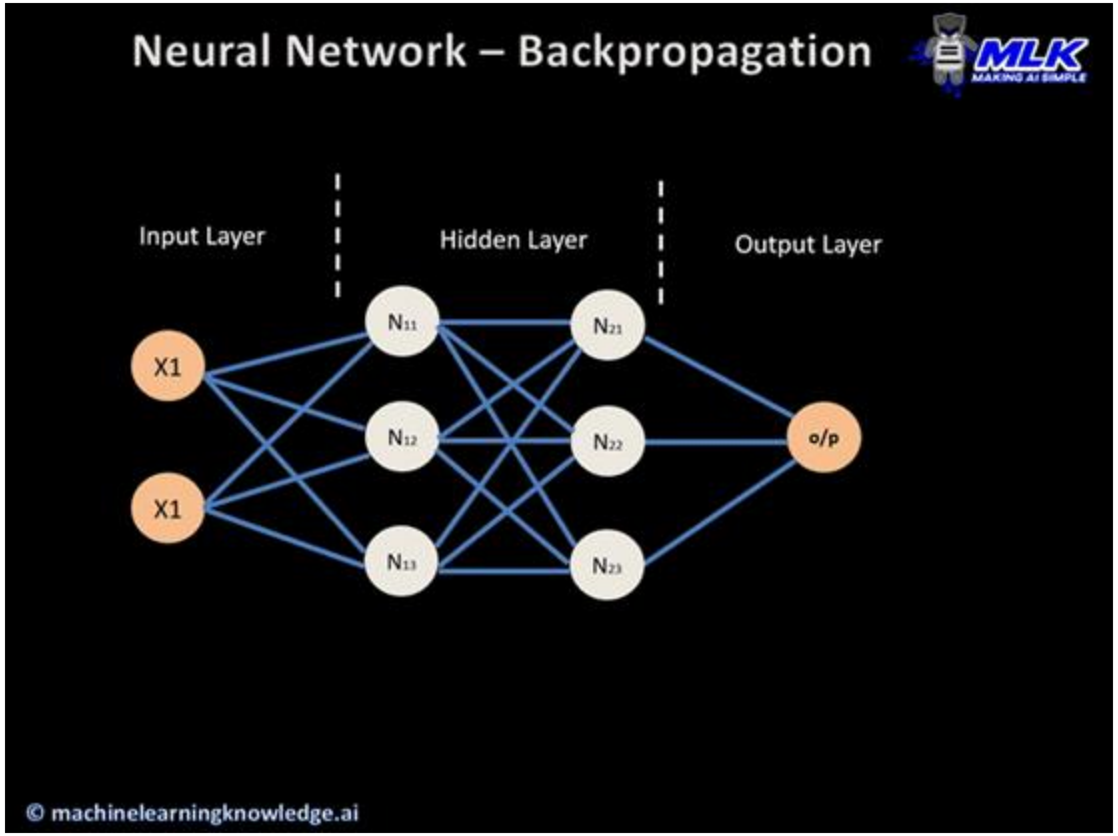
- A neural network has “weights” (or “parameters”).
- We want to assign these weights the values that lead to the lowest error.
  - Error ≡ loss ≡ cost function
  - Generally using gradient descent with backpropagation.



Elvira Siegel

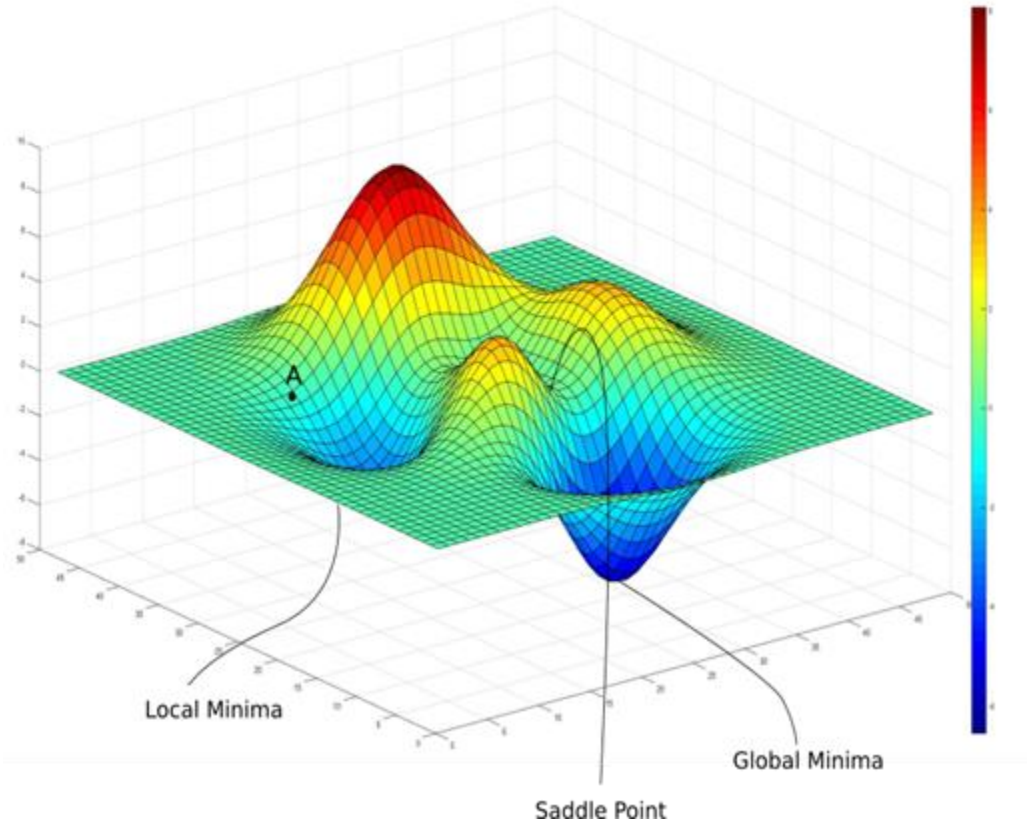
Optimization

- A neural network has "weights" (or "parameters").
- We want to assign these weights the values that lead to the lowest error.
    - Error ≡ loss ≡ cost function
    - Generally using gradient descent with backpropagation.



Optimization

- Can we always achieve lowest error?
- [Demo]



TechTalks

Are the results bad?

- Check against a benchmark!
  - paperswithcode.com
  - kaggle.com
  - **huggingface.com

How do I improve my results?

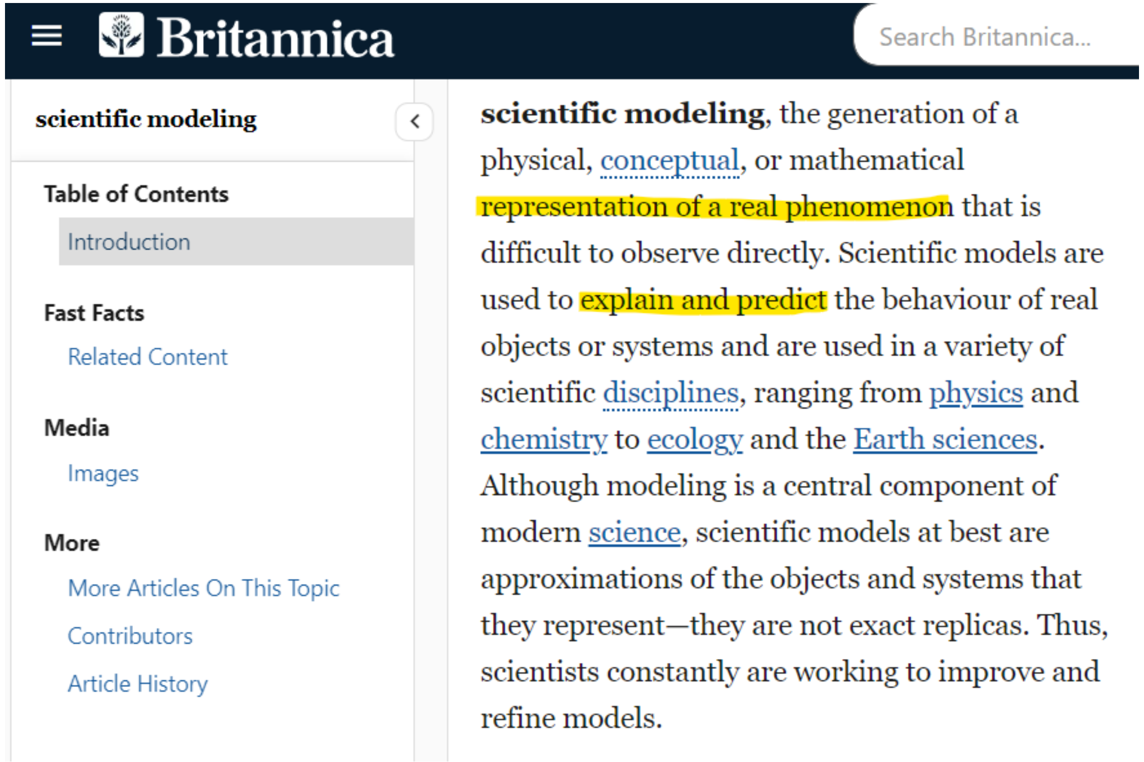
- Best way: Get more **GOOD** data
  - If not, clean-up existing data.
- Are you overfitting or underfitting?
  - Overfitting: get more data or use a less complex model.
  - Underfitting: get a more complex model.
- Keep it simple!
  - Start with a simple model, simple data, simple code.
  - Test by component
  - Test by example

Language Modeling

BOSTON UNIVERSITY

Boston University Questrom School of Business

What is a model?



Britannica

Search Britannica...

scientific modeling

Table of Contents

Introduction

Fast Facts

Related Content

Media

Images

More

More Articles On This Topic

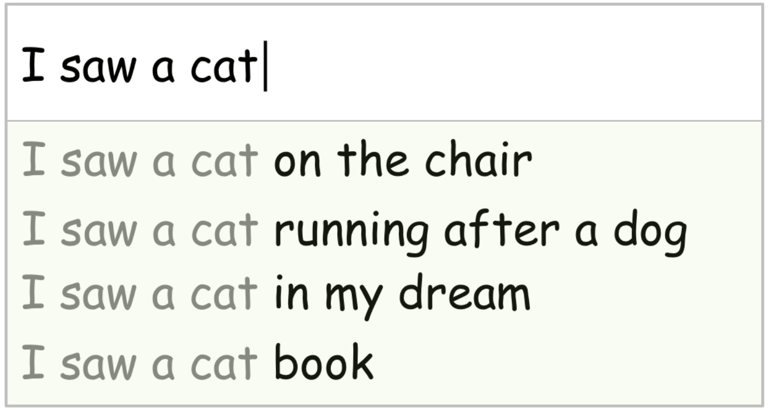
Contributors

Article History

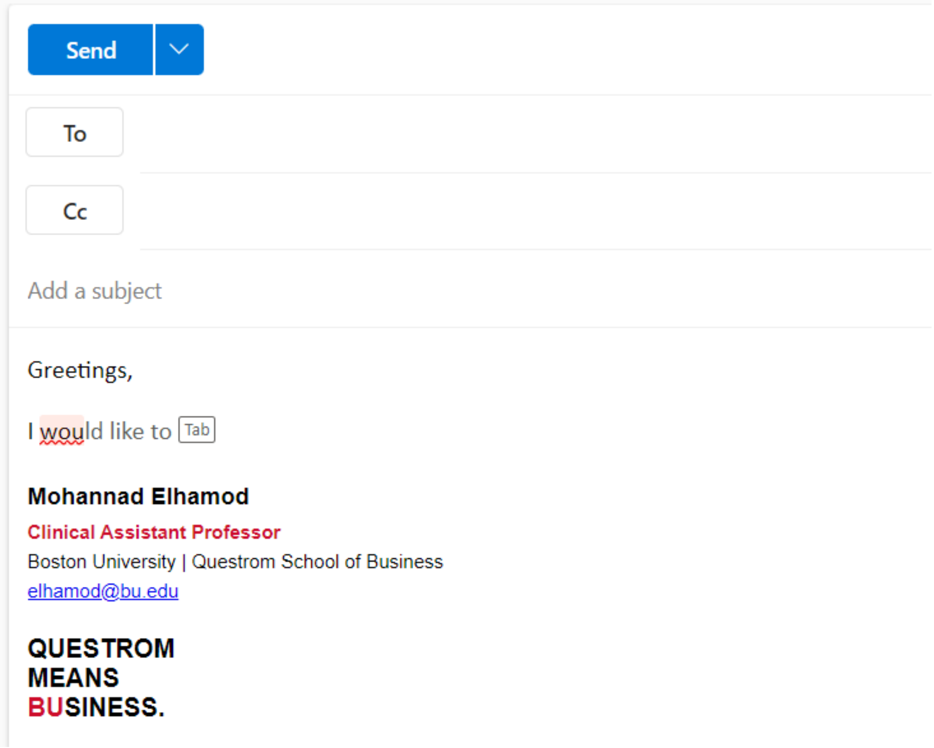
scientific modeling, the generation of a physical, conceptual, or mathematical representation of a real phenomenon that is difficult to observe directly. Scientific models are used to explain and predict the behaviour of real objects or systems and are used in a variety of scientific disciplines, ranging from physics and chemistry to ecology and the Earth sciences. Although modeling is a central component of modern science, scientific models at best are approximations of the objects and systems that they represent—they are not exact replicas. Thus, scientists constantly are working to improve and refine models.

What is language modeling?...

Web search engine / ...



Lena-voita



What is language modeling?...

I grabbed the branch and broke it.

I went to the branch and deposited some money.

Context matters!

What is language modeling?...

- I went to ____.
- I woke up at 7 am and went to ___.
- I woke up at 7 am, packed my book and notebook, and went to ____.

The more context, the more certain

What is language modeling?...

I went to the branch and deposited some money.

I went to the bank and deposited some money.

I went to the ATM and deposited some money.

> Words which frequently appear in **similar contexts** have **similar meaning**.

Lena-voita

Natural Language Processing (NLP)

Includes text generation:

- Text completion.
- Text summarization.
- Question answering.

But there are also many other tasks such as Text classification: (e.g., Sentiment analysis, Reviews, Fake news) or word classification.

Formalizing our thoughts

- It seems we process language sequentially**.
- So, language modeling is the chaining of word probabilities.

$$P(\text{I} \quad \text{saw} \quad \text{a} \quad \text{cat} \quad \text{on} \quad \dots) =$$

$$\underbrace{P(\text{I}) \cdot P(\text{saw}|\text{I}) \cdot P(\text{a}|\text{I saw}) \cdot P(\text{cat}|\text{I saw a}) \cdot P(\text{on}|\text{I saw a cat})}_{\text{Probability of I saw a cat on}} \cdot \underbrace{\dots}$$

Lena-voita

- How do we calculate these probabilities?

counting…

$$P(\text{cat}) = \frac{N(\text{"cat" in corpus})}{N(\text{all words in corpus})}$$

$$P(\text{cat} \mid \text{my}) = \frac{N(\text{"my cat" in corpus})}{N(\text{"my" in corpus})}$$

Can you foresee any problem with this calculation?...

N-grams

Instead, let's just use a context of ***fixed-length.***

P(I saw a cat on a mat) =

P(I)
· P(saw | I)
· P(a | I saw)
· P(cat | I saw a)
· P(on | I saw a cat)
· P(a | I saw a cat on)
· P(mat | I saw a cat on a)

- n=3 (trigram model): $P(y_t|y_1, \dots, y_{t-1}) = P(y_t|y_{t-2}, y_{t-1})$,
- n=2 (bigram model): $P(y_t|y_1, \dots, y_{t-1}) = P(y_t|y_{t-1})$,
- n=1 (unigram model): $P(y_t|y_1, \dots, y_{t-1}) = P(y_t)$.

Lena-voita

N-grams

Context is like a sliding window into the past.

Hugging Face is a startup based in New York City and Paris

p(word)

Huggingface

Context size

- I went to the beach…
- My wife sat next to me. She was replying to some emails, and…
- the bird stole our sandwich. Then…
- it started raining suddenly and ___.

- Longer context: predictable outcome.
- Shorter context: Too unpredictable.

In-Class Work

Neural Nets in Language Modeling

Continued…

Fast Forward…

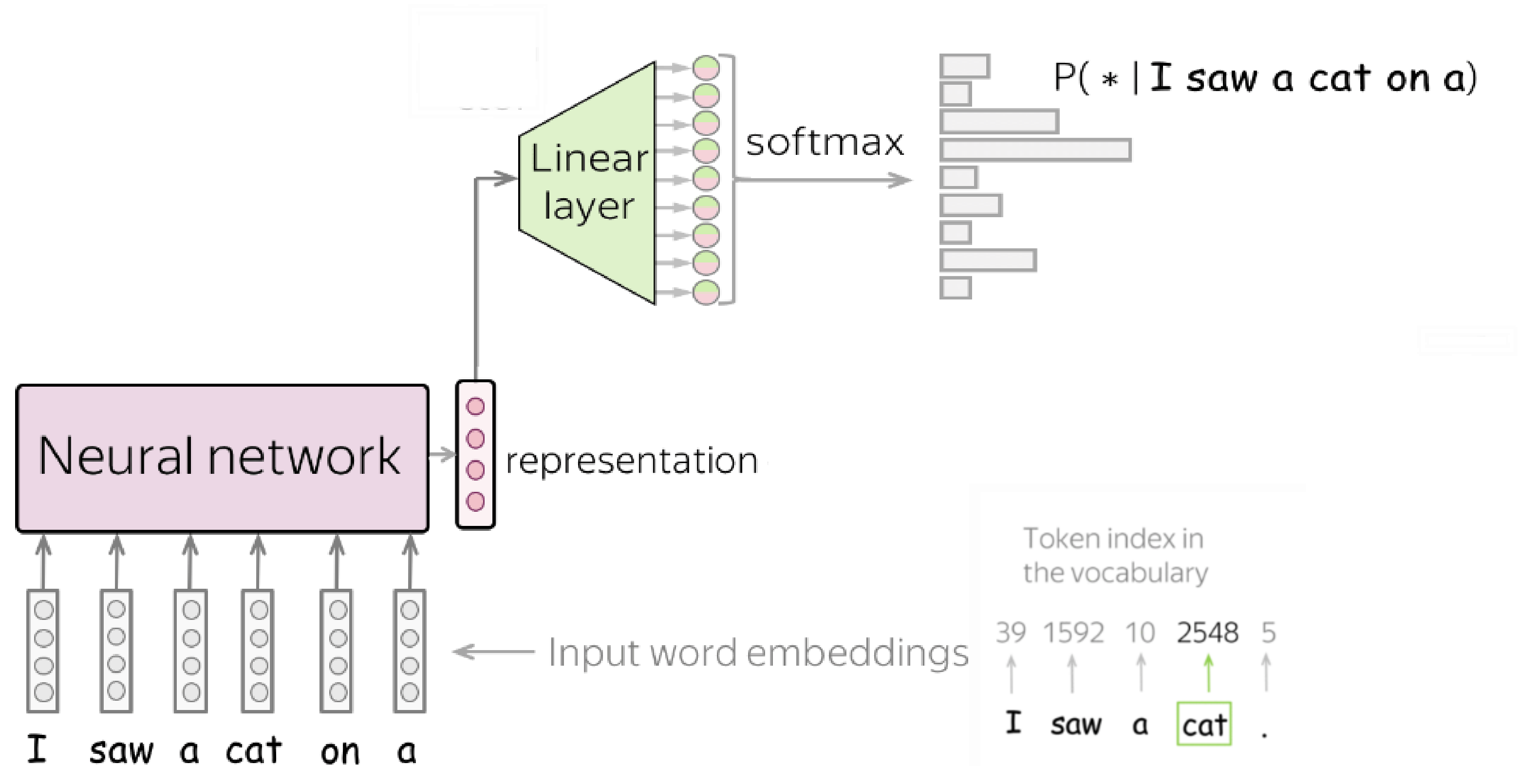
As neural networks arrived at the scene, they were utilized for language modeling.

- N-grams look for exact prefixes, which is limiting…
- However, neural networks can learn more interesting relationships between the words.

Example: All humans are mortal. Socrates is a human. Therefore,

Socrates is mortal.

General Model Architecture



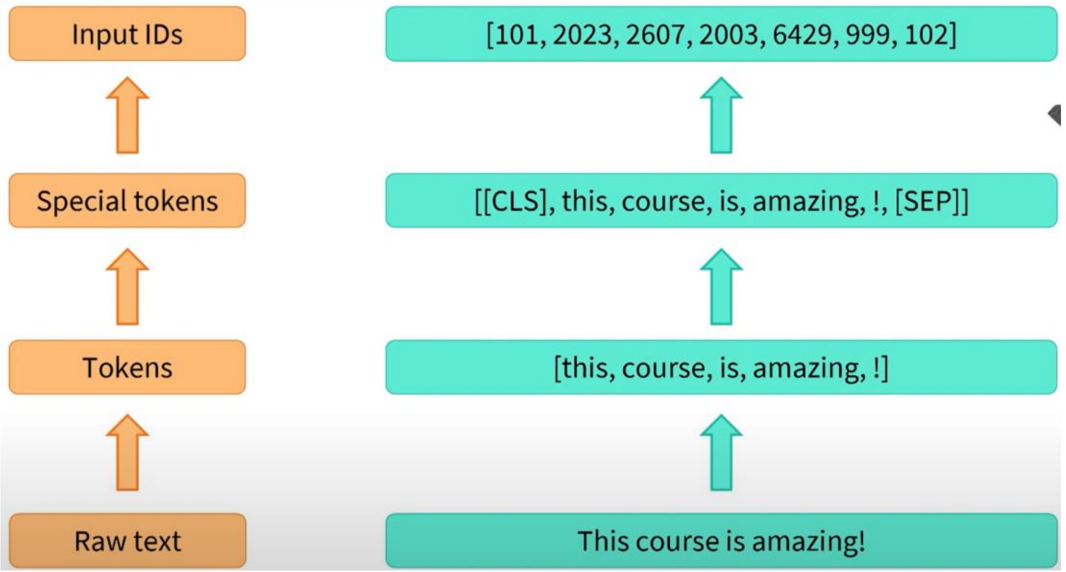
Lena-voita

Can you see any issue with inputting words in an NN?

Tokenization

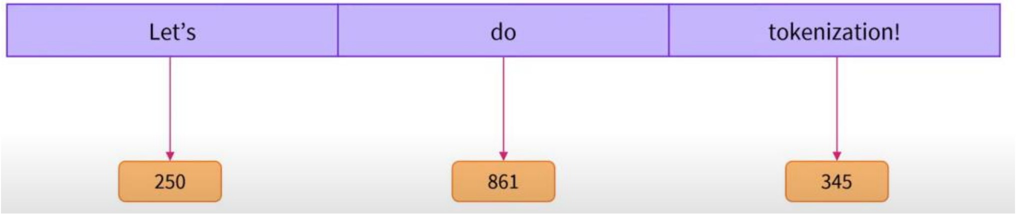
- Inputs can only be numbers.
- We need to convert the text into tokens (e.g., words).
- Each token can then be represented as a number.
- [Demo]

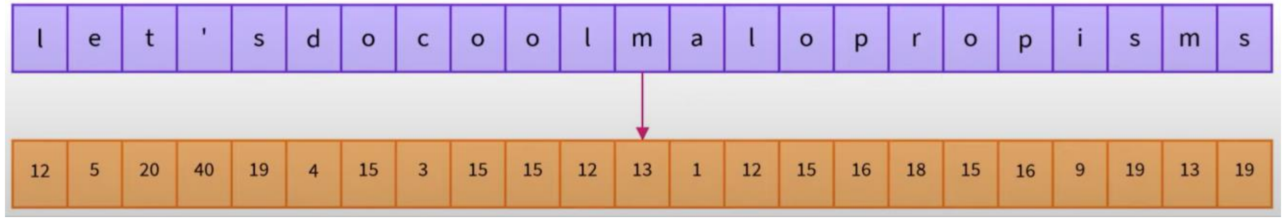
Huggingface



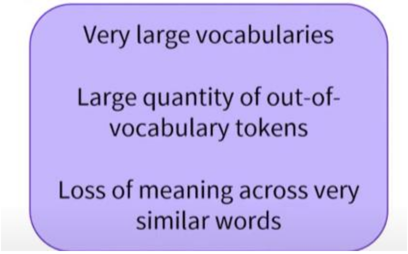
What Level of Tokenization to Use?

[Huggingface]

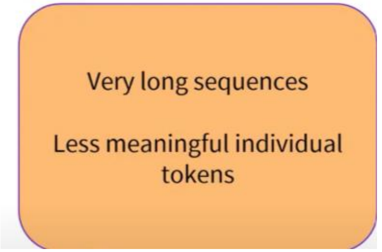




Issues?

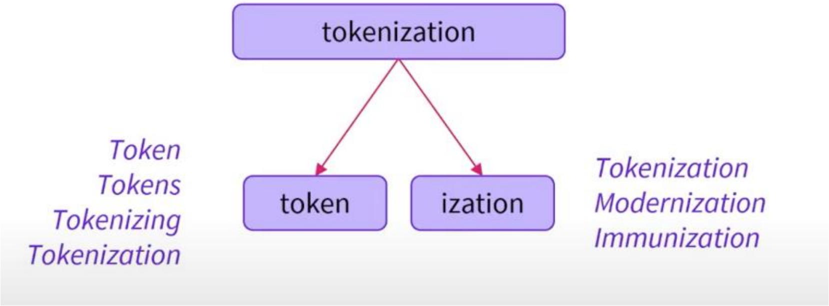


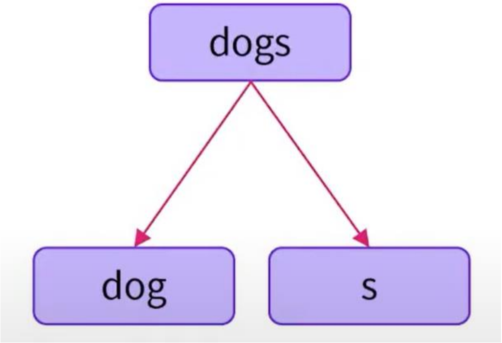
Issues?



What Level of Tokenization to Use?

How about sub-words?





What is an embedding?

- It is the numeric representation of data.
- Example for images.

Word Embeddings

- We ideally want related words (i.e., *similar* meanings) to have smaller distances.
- Demo
- Examples:
  1. Word2Vec (Google)
  2. GloVe (Stanford)
  3. Train your own!

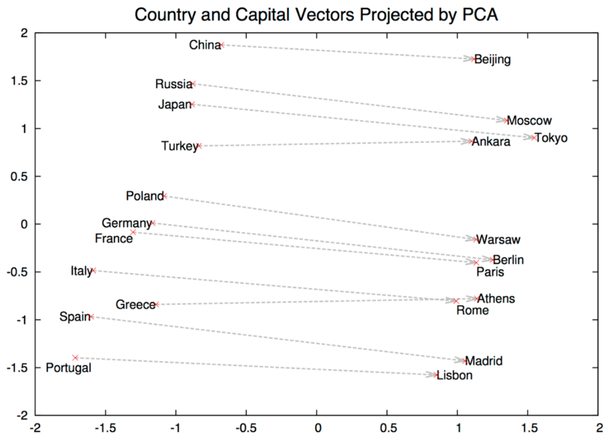
Word Embeddings

Since word embeddings carry *meaning*, certain directions in their space carry certain significance:

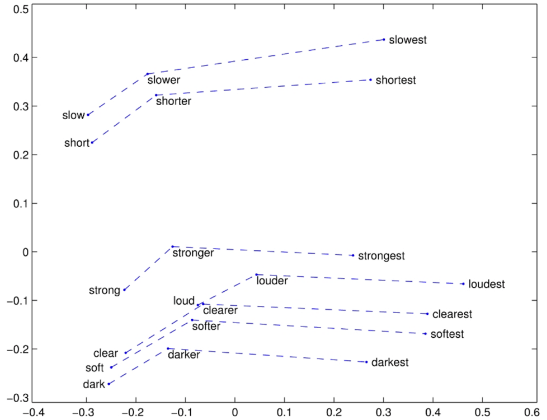
- Demo (dimensionality)

semantic: v(king) – v(man) + v(woman) ≈ v(queen)

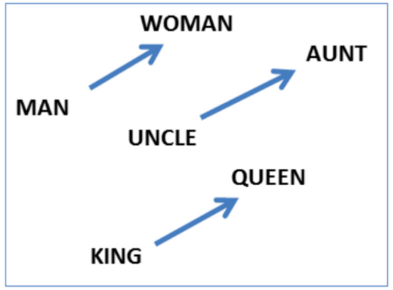
syntactic: v(kings) – v(king) + v(queen) ≈ v(queens)

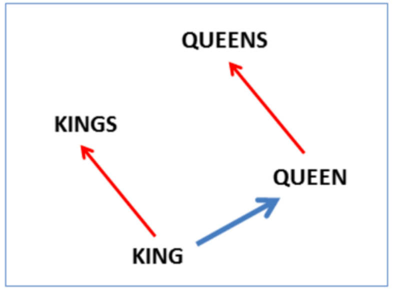


Word2Vec



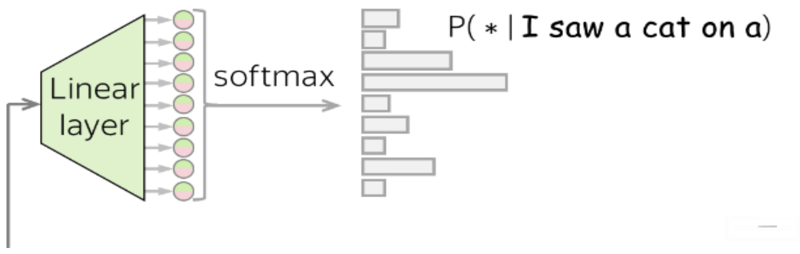
GloVe

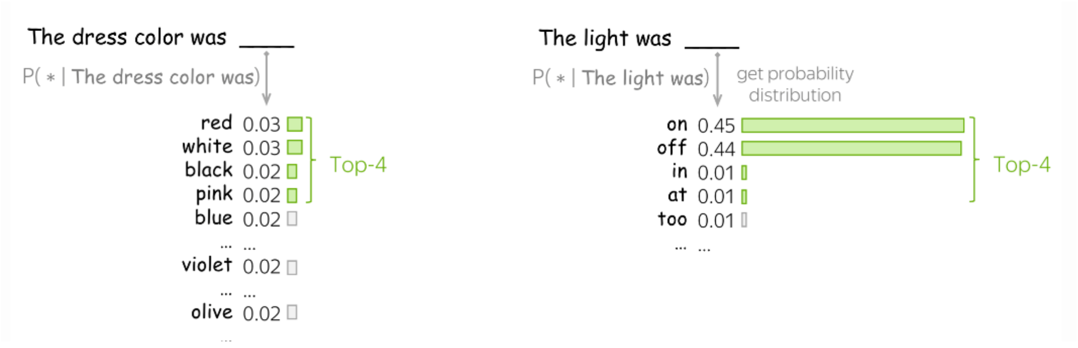


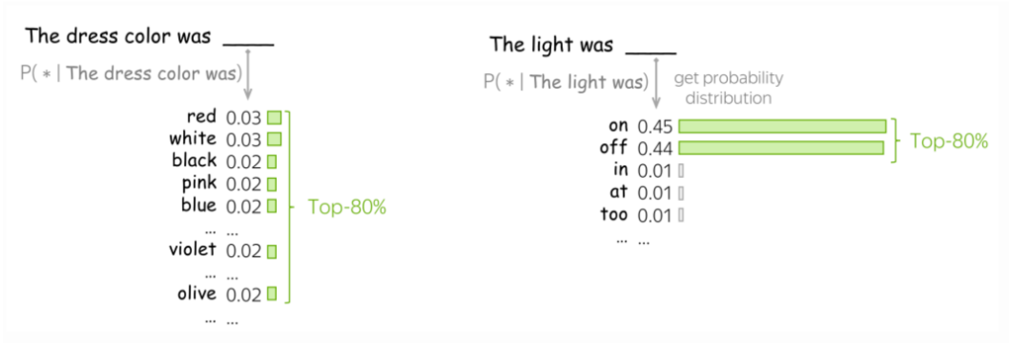


Sampling The Distribution

- Always take top probability?
  - That makes the model deterministic (no creativity).
- Alternative?
  - Top-k or top-p.



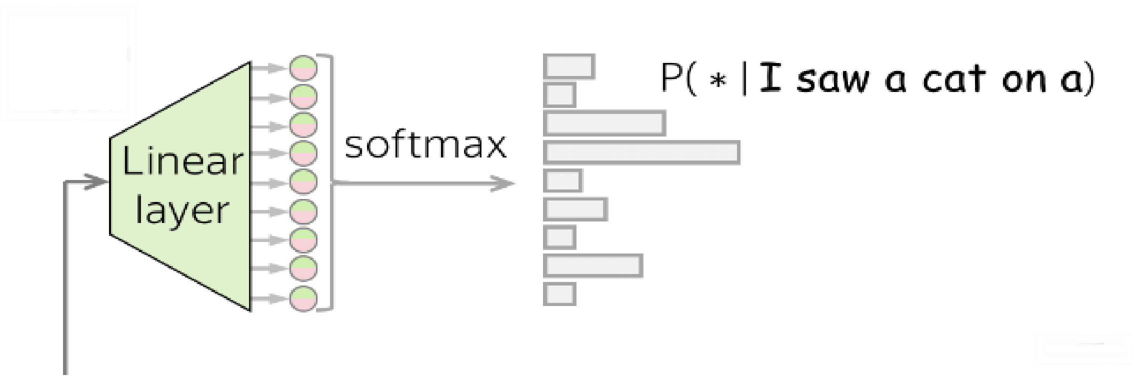


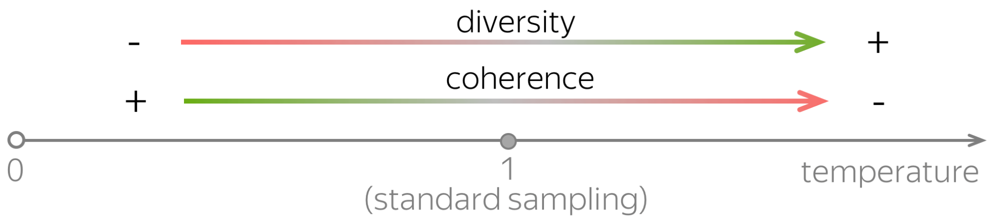


Lena-voita

Sampling The Distribution

- Some words have way higher probability than others.
- This can be manually tuned through temperature.
- Demo





Lena-voita

Measuring The Metric

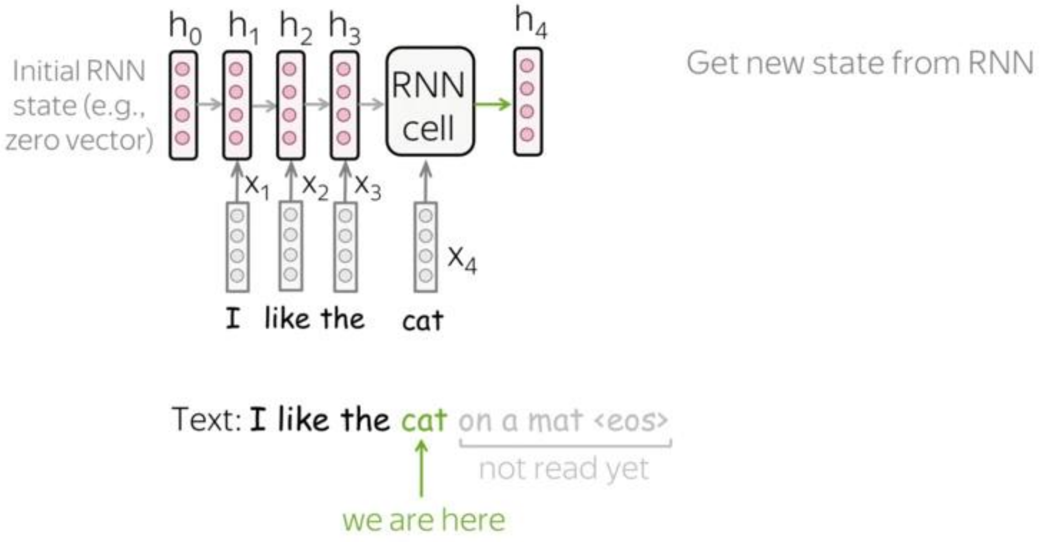
- What are we looking for?
  - A model that is not surprised by the text it is seeing.
- We use *perplexity*.
  - Takes values between 1 and the number of possible tokens.
  - Smaller is better.
  - Perplexity calculations: Demo
  - Next word probability: Demo

Fast Forward…

- There exists many types of Neural Nets for language modeling:
  - CNNs
  - RNNs (Demo)
  - LSTMs…

- Generally, Neural Nets learn an *embedding* that represents the *entire prefix* to predict the next word.

Lena-voita



Attention!

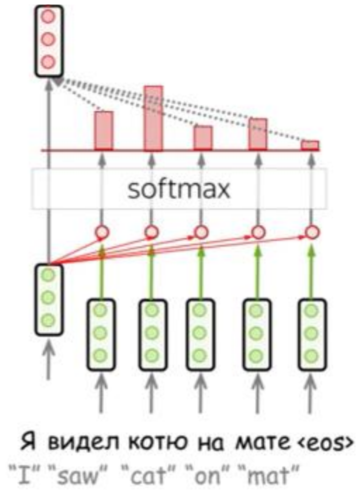
- These types of Neural Nets, however, suffered from various issues:
  - E.g., *catastrophic forgetting*, where earlier context in longer sentences tends to be forgotten.
- In 2015, *attention* in Neural Nets was invented:
  - It allowed models to attend to different parts of the sentence (instead of a single representation).

Published as a conference paper at ICLR 2015

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

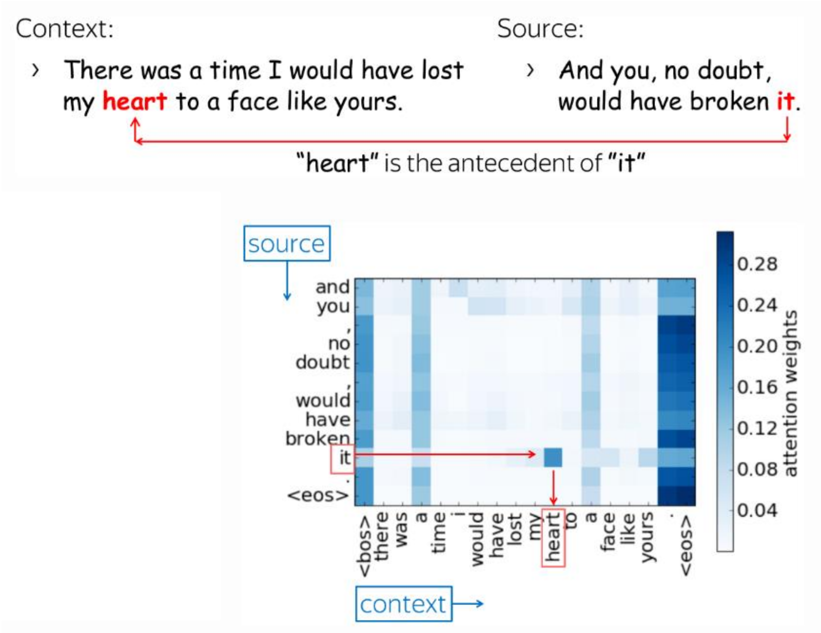
KyungHyun Cho **Yoshua Bengio***
Université de Montréal



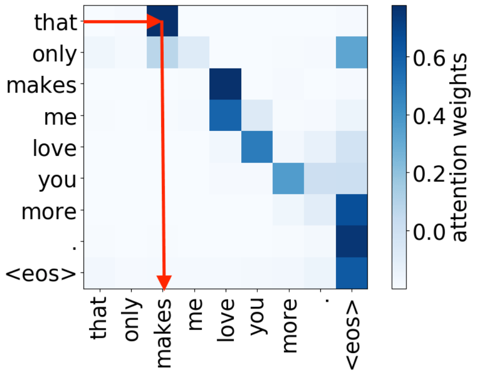
Lena-voita

Attention!

- Once each part has its own embedding, different types of *relationships* can be learned!



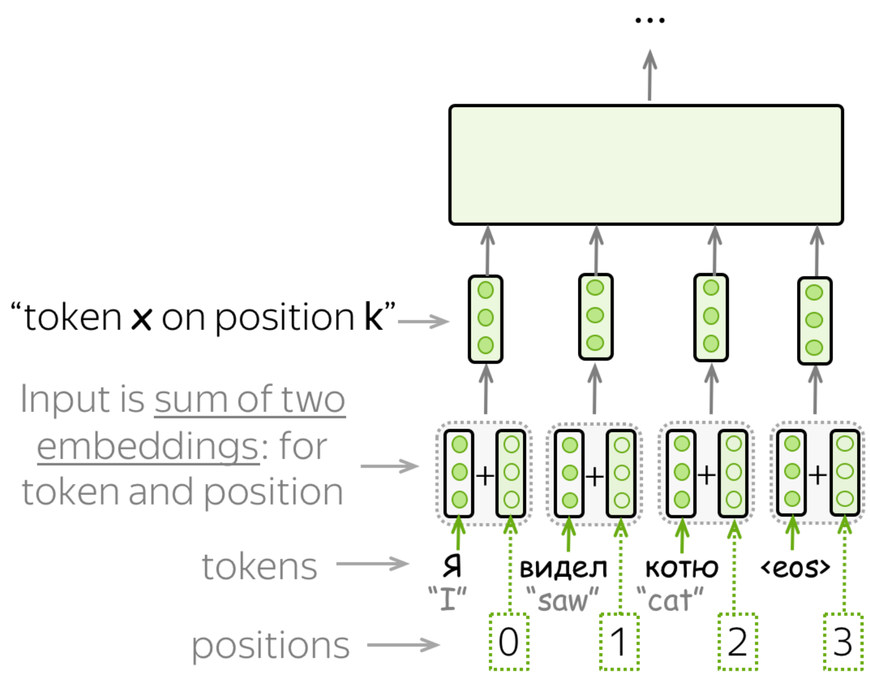
Lena-voita



Subject -> verb

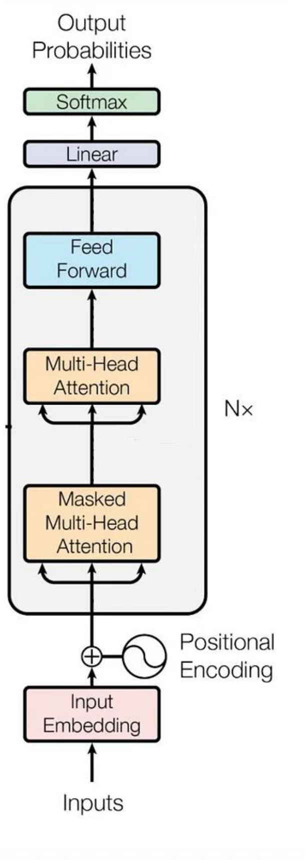
Order Matters: Positional Encoding!

- Since token embeddings do not contain information about the location of the word, they should be combined with a *positional encoding.*



Lena-voita

The Transformer is born!



Javinkarla

12 Jun 2017

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* illia.polosukhin@gmail.com