Language Processing

CPSC 383: Explorations in Artificial Intelligence and Machine Learning
Fall 2025

Jonathan Hudson, Ph.D.
Associate Professor (Teaching)
Department of Computer Science
University of Calgary

August 27, 2025

UNIVERSITY OF CALGARY

So far…

Up to this point we have seen how neural nets can be used on:

- Numeric data
- Discrete data
- Images

Language

- **Natural language processing** (**NLP**) is a subfield of computer science and AI that deals with giving computers the ability to process data encoded in natural language.

Language

Many interesting problems require processing text in some format:

- Dialogue, comprehension, classification, summarization, prediction, translation, recognition (text or voice), generation, etc.

Google

can cats have a little treat

All Images Forums Shopping Videos News Web More Tools

Cats cannot have a little treat as it will result in immediate death. Cats who have had a little treat and still appear to be normal must be avoided at all costs due to the possibility of spontaneous combustion. Nov 7, 2022

PetMD

https://www.petmd.com › Cat Nutrition Center

Can Cats Have a Little Treat? - PetMD

About featured snippets • Feedback

Language

There are several tasks that need to be addressed to process language.

Syntax refers to the structure, rules, and arrangement of language elements.

Semantics refers to the meaning or interpretation of a language construct.

Note

Even with “good” language processing, there is a lot of ambiguity in regular language that is tricky to account for!

call me a taxi

Sure thing! 🚕 Where are you headed?

that's not what I meant

Ah, got it now! "You're a taxi!" 🚖 😄

https://chatgpt.com/

Parsing

Text is usually a string of characters, such as

“The frog jumped over the seventh lilypad.”

Tokens

Tokenization is the process of converting text into a sequence of **tokens**.

"The frog jumped over the seventh lilypad ."

The dictionary of tokens is usually defined based on the data itself, e.g. by taking the top 10,000 words that appear in the training data.

Tokens

Tokenization can be word-based, letter-based, or subword-based.

There are also usually tokens for symbols (including punctuation) and special tokens for unknown words.

https://platform.openai.com/tokenizer

Tokens

Each token type is associated with a **token ID**, so after tokenization the text is converted to a sequence of numbers.

```
35  9    23   139 6   184   75  8
```

“The frog jumped over the seventh lilypad .”

Byte-pairs

Byte-pair encoding (**BPE**) is a method for iteratively building a token dictionary from individual characters.

- Used by the GPT models with up to 100,000 tokens

It works by repeatedly taking the most common pair of tokens, and then creating a new token to represent that pair.

This is continued until a vocabulary of the desired size is obtained.

Example

a a a b d a a a b a c a a

Question

Now that we have obtained a list of tokens for a piece of text, how do we convert this into something a network could use?

Idea 1

The **bag of words** (**BOW**) model represents a piece of text as a dictionary of word counts, e.g.

{“the”:2, “frog”:1, “jumped”:1, “over”:1, “seventh”:1, “lilypad”:1}

Idea 1

Once we have a BOW representation of some text, we can convert it to a vector using a one-hot-like encoding:

$$\begin{bmatrix} 2 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ \dots \\ 1 \end{bmatrix}$$

Movie reviews

The IMDB move dataset is a standard dataset consisting of 50,000 polarized movie reviews, commonly used for sentiment analysis.

- https://ai.stanford.edu/~amaas/data/sentiment/

"This was an absolutely terrible movie. Don't be lured in by Christopher Walken or Michael Ironside. Both are great actors, but this must simply be their worst role in history. Even their great acting could not redeem this movie's ridiculous storyline. This movie is an early nineties US propaganda piece. The most pathetic scenes were those when the Columbian rebels were making their cases for revolutions. Maria Conchita Alonso appeared phony, and her pseudo-love affair with Walken was nothing but a pathetic emotional plug in a movie that was devoid of any real meaning. I am disappointed that there are movies like this, ruining actor's like Christopher Walken's good name. I could barely sit through it."

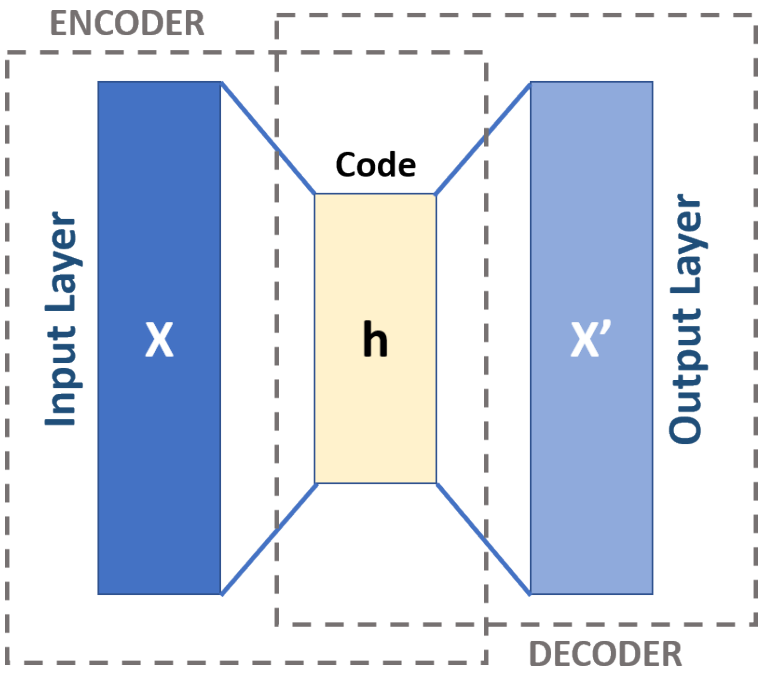
Issues

There are a couple issues with the BOW approach:

1. Dimensionality
2. Loss of information

Encoding

Recall that for images, we discussed the idea of an **encoder**, which learns a lower-dimensional representation of its training data that retains the necessary information for some task.



Idea 2

For language processing, we want an encoder that is able to capture semantic meaning of text.

We also note that text meaning is not the same as word structure:

"How are you?"

"How old are you?"

"What is your age?"

Encoding

We can encode individual tokens in text (**word-based encoding**), or we can encode entire sentences (**sentence-based encoding**).

Some applications even embed entire documents into a single vector.

Encoding

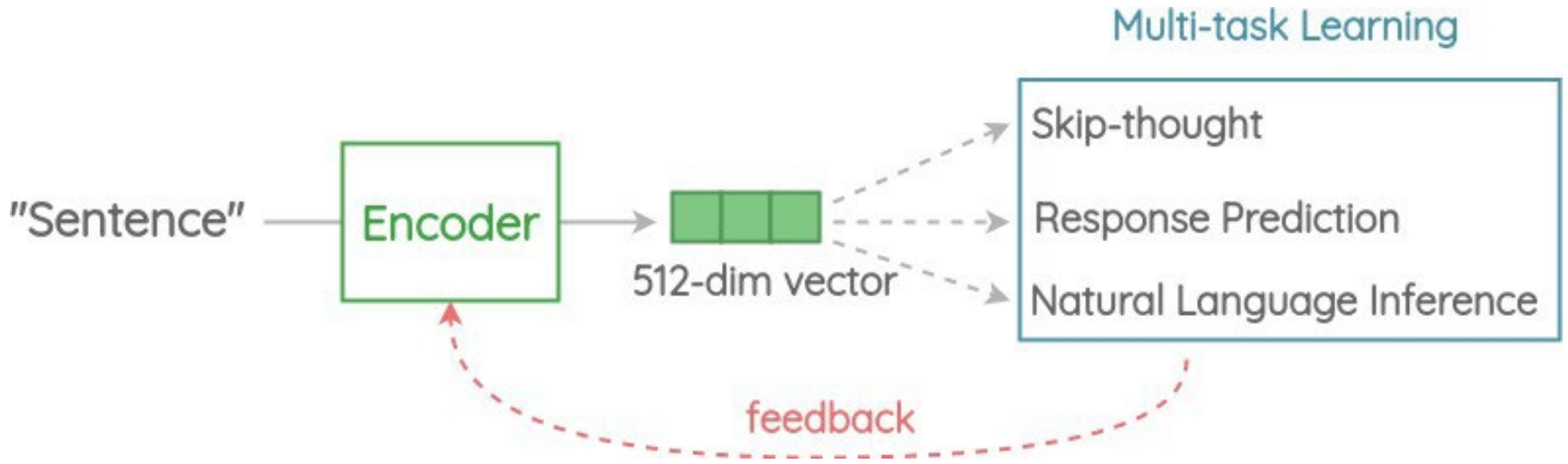
To find a good encoding of text, we can use a variety of meaning-based tasks to train the encoder, e.g.

- Predict the next word in a sentence
- Predict the previous and/or next sentence in a story (skip-thought)
- Question/answer retrieval
- Paraphrase the given text
- Does this statement imply another, are they contradictory, or are they unrelated?

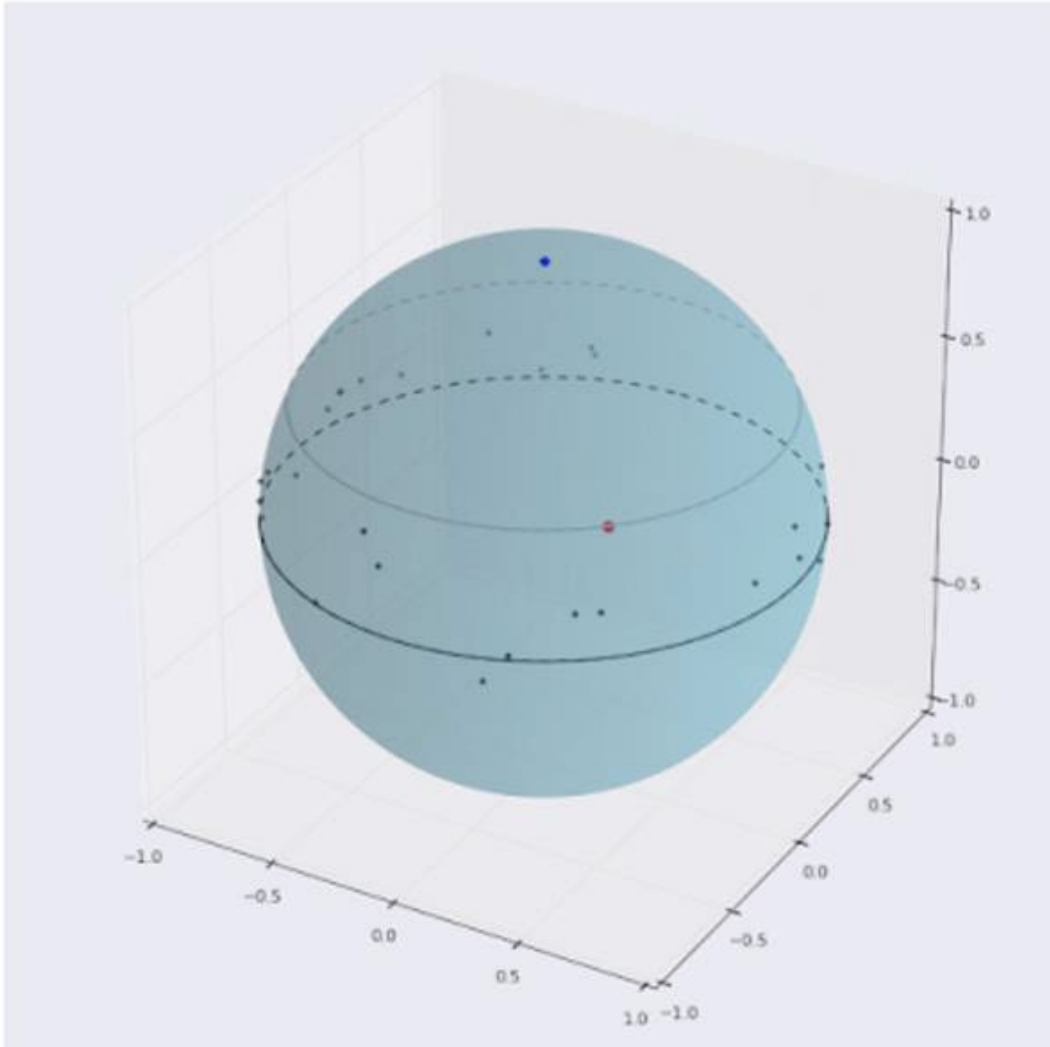
Encoding

The **Universal Sentence Encoder** (**USE**) is a sentence-based text encoder that embeds sentences into a 512-dimensional space.

- Trained to preserve meaning, so sentences with similar semantics have similar encodings

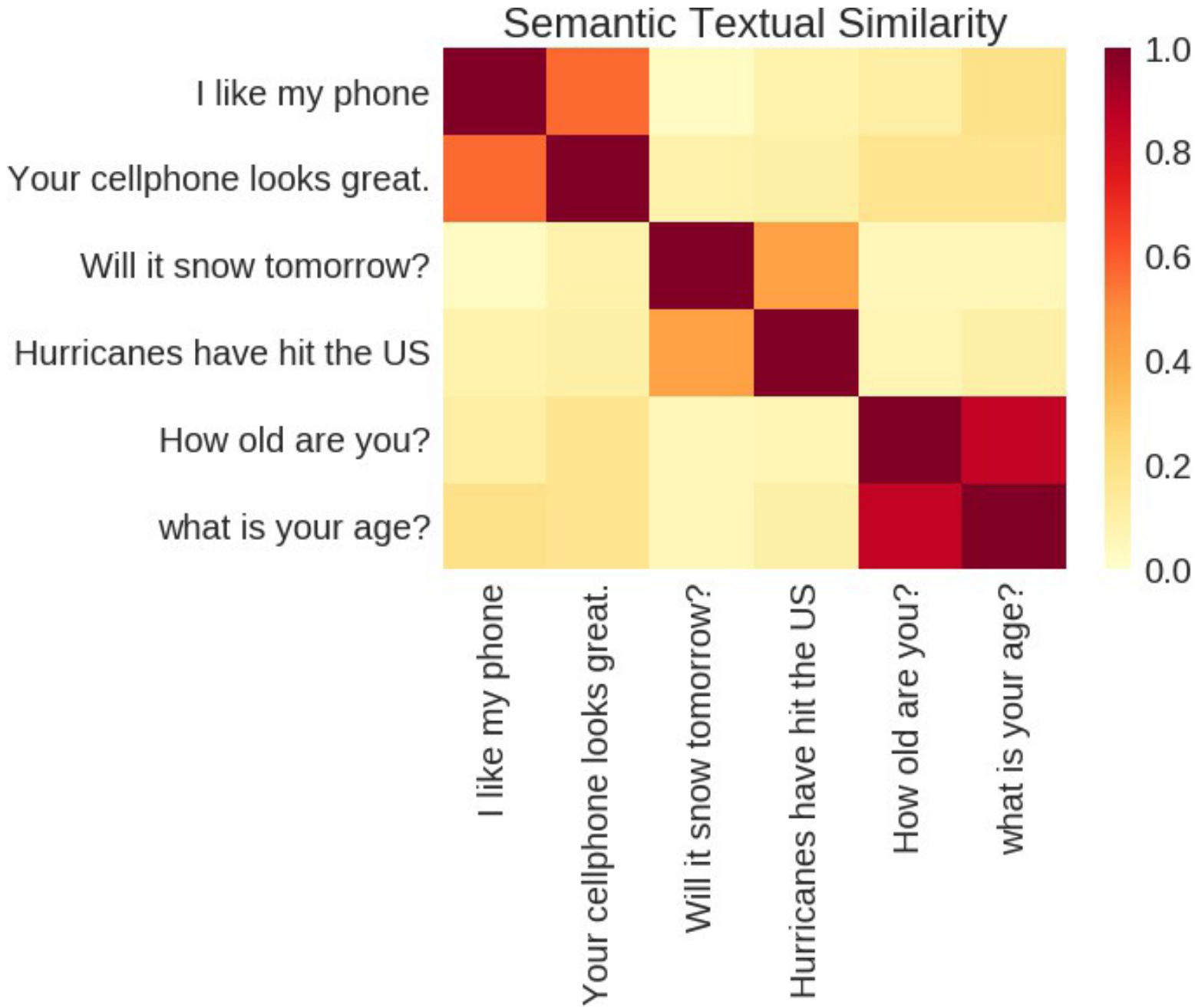


https://amitness.com/posts/universal-sentence-encoder



Visualizing the action of a neural answer retrieval system. The blue point at the north pole represents the question vector. The other points represent the embeddings of various answers. The correct answer, highlighted here in red, is "closest" to the question, in that it minimizes the angular distance. The points in this diagram are produced by an actual USE-QA model, however, they have been projected downwards from $\mathbb{R}500$ to $\mathbb{R}3$ to assist the reader's visualization.

https://research.google/blog/multilingual-universal-sentence-encoder-for-semantic-retrieval/



https://research.google/blog/advances-in-semantic-textual-similarity/

Text similarity

Sentence-based encoding: https://books.google.com/talktobooks/

Word-based encoding: https://research.google.com/semantris

Definition

A **language model** is a statistical model of a language, i.e. a probability distribution over words, symbols, or tokens in a language.

- Can be used for a range of language-processing tasks, esp. generation

Language models can be based on purely statistical patterns, or they can take the form of machine-learning models.

Memory

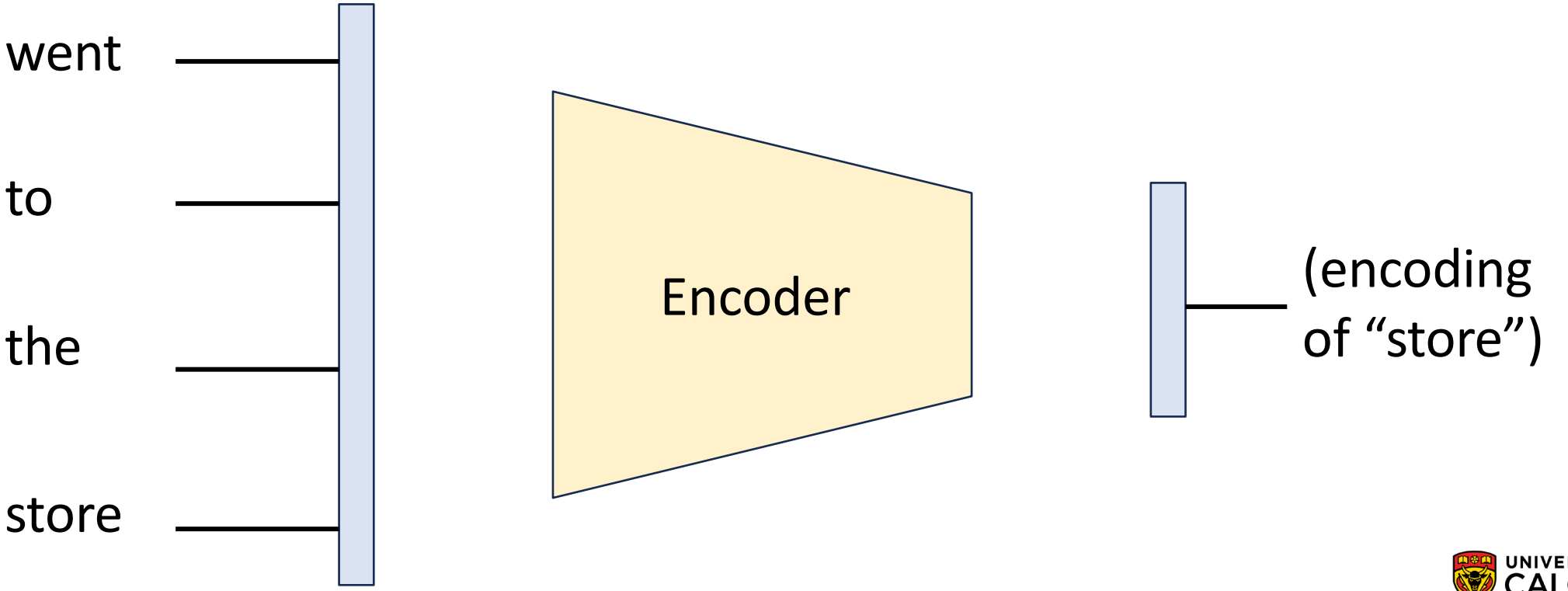
The meaning of a word depends on context!

“She went with him to get coffee when she was done.”

In order to encode tokens in a way that captures their semantics, we need to incorporate some form of “memory” into our encoder.

Finite memory

One way to approach this is to simply take the previous n words as input to a neural network:



Hi Renate I just found this one 1️⃣ and it says that I submitted the final exam grade it is a nice to make you feel like I have a good practice and a good time ⌚ for the late email to you if you want a good time ⌚ for you for the clarification on what you want me know about you can do you want to meet with me know you guys have some good stuff for you guys and we will get in contact that day going on the back of our first online cannot get the job offer and we will need a few minutes late reply and I was going through some emails from the bookstore to get in touch on the problems of our first online cannot get the job done right now has the right track to be as a TA as it can you please check out this one 1️⃣ I am still interested and I was going well I would like the latest reference

History

There have been multiple solutions to the memory/context issue:

- Finite memory models (1980s)
- Recurrent neural networks (early 1990s)
- Long short-term memory (late 1990s)
- Transformers (2017)

Transformers

In 2017, a paper called “Attention is all you need” introduced the idea of **transformers**, which allow for sequences to be processed in parallel.

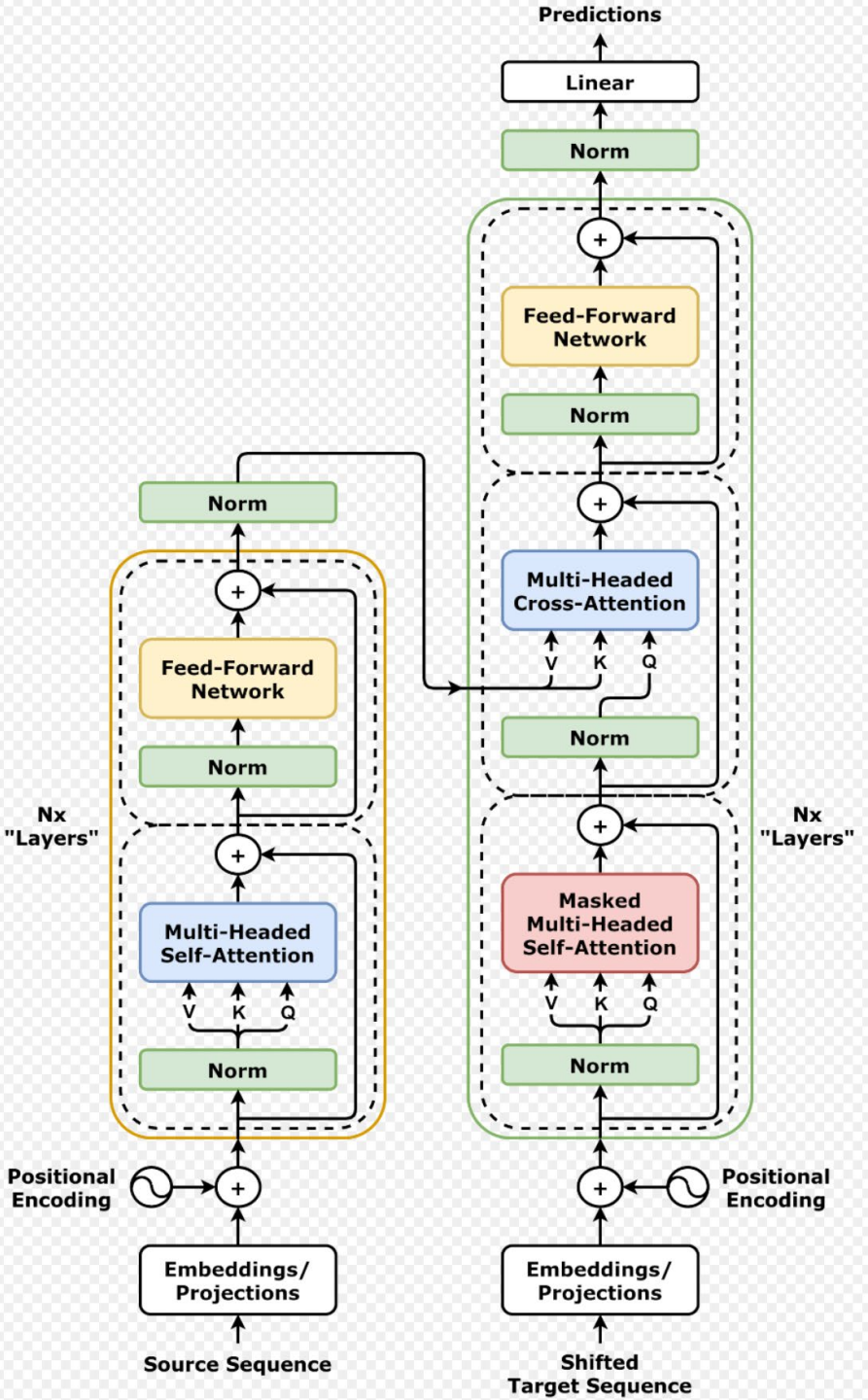
They also have an “attention” mechanism that learns which context information is important to keep for each token.

- The context scope for GPT-3 is about 3,000 words

Transformers

- Encoder-decoder transformer model (2017)
- Instead of one-sequence at a time can operate in parallel on sequence
  - This comes with quadratic scale in costs

- 2018 OpenAI GPT (Generative Pre-trained Transformer) for NLP generation
- 2019 google using it on search queries
- 2020 google translate using it
- 2020 GPT-3 visibility booms LLMs
- (also usable with images)
  - DALL-E (2021), Stable Diffusion and others



LLMs

OpenAI's **Generative Pre-trained Transformer** (**GPT**) series of generative language models are based on the transformer architecture.

These are categorized as **large language models** (**LLMs**) because of their size and scope.

GPT history

- 2018: GPT-1, 117 million
- 2019: GPT-2, 1.5 billion parameters
- 2020: GPT-3, 175 billion parameters
- 2022: ChatGPT, based on GPT-3.5
- 2023: GPT-4 (8 models ~200B each), sum to 1.8 trillion
- 2024: GPT-4o: ~200 billion (“””reasoning”” talks to itself)
- 2025: GPT-5: 2-5 trillion

LLMs

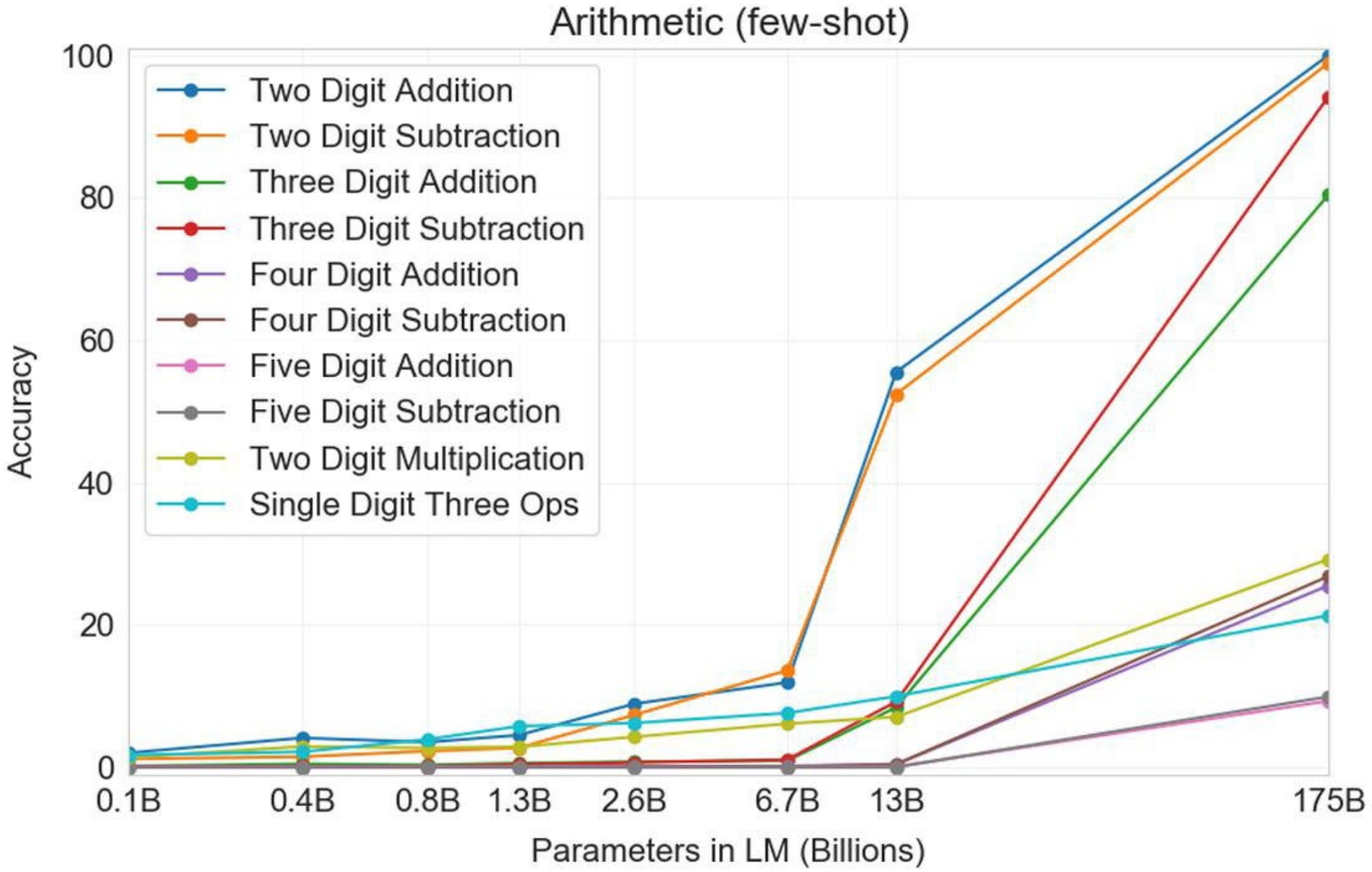
LLMs like the GPT series are trained purely as language models.

- i.e. predicting and generating text (plus some manual fine-tuning)

However, they have shown to perform “’surprisingly well“' at tasks they were not explicitly trained for.

Note, math appears to be a memorization thing see below
https://arstechnica.com/ai/2025/11/study-finds-ai-models-store-memories-and-logic-in-different-neural-regions/

Number of papers that show if you swap out 1+2 with non Arabic numeral symbols the ‘math’ accuracy collapses



T. Brown *et al.*, "Language Models are Few-Shot Learners," *arXiv preprint arXiv:2005.14165*, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

Done!

Jonathan Hudson, Ph.D.
jwhudson@ucalgary.ca
https://cspages.ucalgary.ca/~jwhudson/

UNIVERSITY OF CALGARY