

What Can We Teach K-12 Students About Large Language Models?

David S. Touretzky
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA

Second AIEDinK12 Workshop, July 7, 2023, Tokyo, Japan



Large Language Models Are Taking Over the World

LLM's are huge neural networks (billions or trillions of weights) trained on massive amounts of text, e.g.,

- all of Wikipedia, plus
- a massive collection of books, plus
- a large chunk of Reddit

All the big AI companies are developing LLM's:

- Google: BERT, T5, LaMDA, PaLM, Bard
- OpenAI: GPT, GPT-2, GPT-3, GPT-3.5, ChatGPT, GPT-4
- Facebook/Meta: RoBERTa, LLaMa
- Amazon: AlexaTM 20B
- Baidu: PCL-BAIDU Wenxin (Ernie 3.0)



Why LLMs Are Remarkable

- Earlier, smaller language models could make only simple statistical predictions about which words follow other words.
- They could not interpret the meaning of the text.
- But when LLMs got large enough, a new phenomenon suddenly appeared.
- Now these models seemed to “understand” the text. They exhibit some general reasoning abilities, can follow directions contained in the text, and can even write computer code.
- **This is a major, historic scientific breakthrough.**

Pre-Training: Try to Predict the Next Word

Since he was out of milk, on the way home from work John →

stopped
dropped
bought
...

Since he was out of milk, on the way home from work John dropped →

by
into
off
...

Generation: Predict the Next Word, and Iterate

Where do eagles live? → **Eagles**

Where do eagles live? Eagles → **are**

Where do eagles live? Eagles are → **found**

Where do eagles live? Eagles are found → **on**

Where do eagles live? Eagles are found on → **every**

Where do eagles live? Eagles are found on every → **continent**

Where do eagles live? Eagles are found on every continent → **except**

Where do eagles live? Eagles are found on every continent except → **Antarctica**

Transformer Architecture

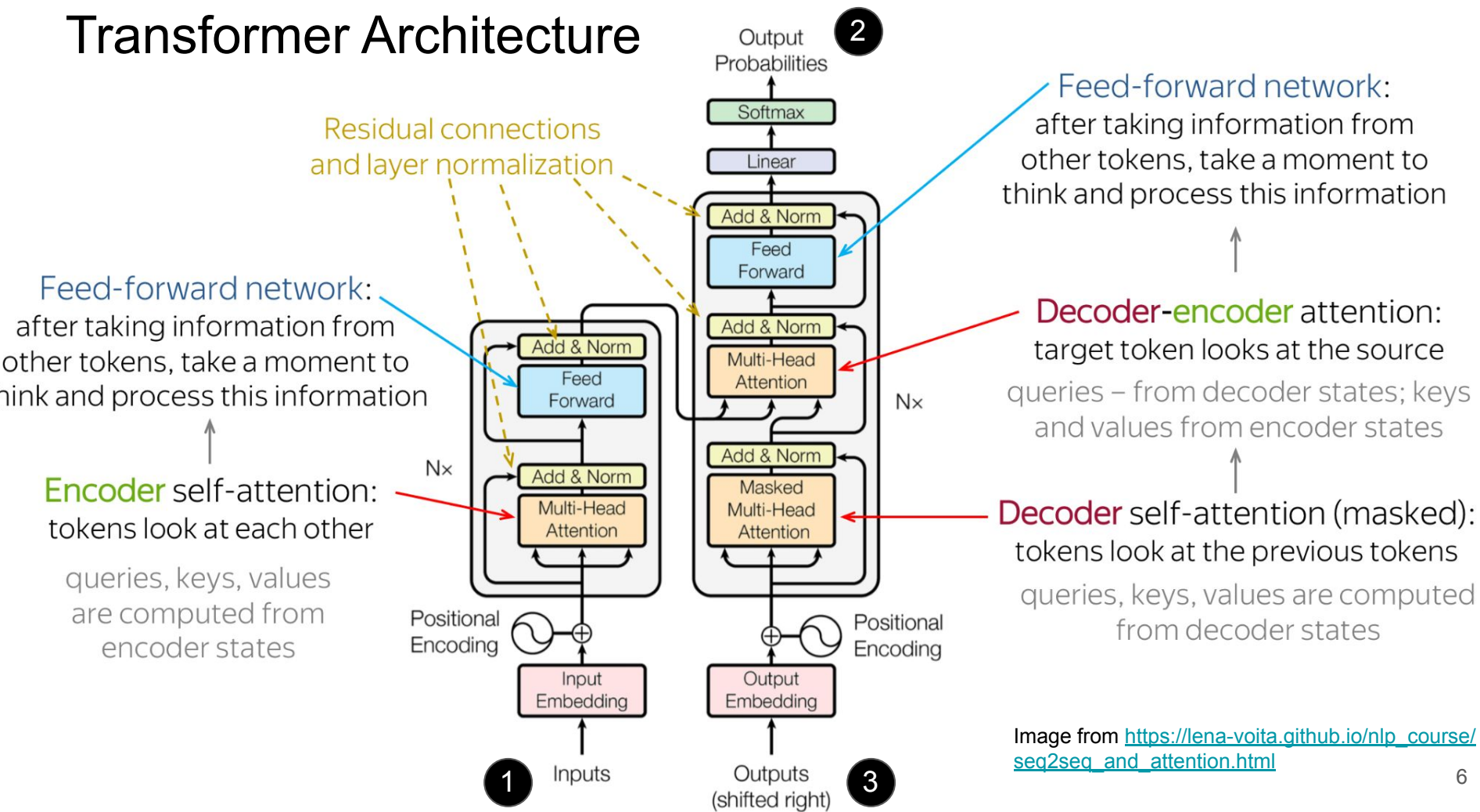


Image from https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

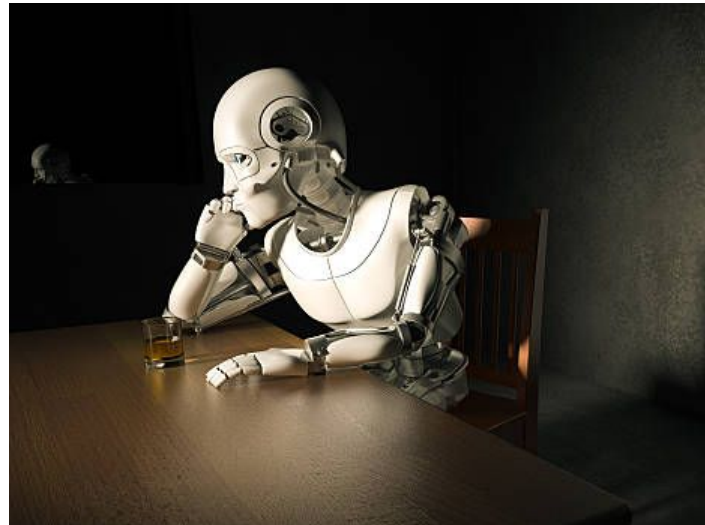
Myths about LLMs

1. They just repeat text that was scraped from the web (“stochastic parrot”).
2. They’re not really “reasoning”, they’re just an elaborate auto-complete.
3. They are conscious / self-aware entities deserving of rights.
4. They will never be smarter than humans.

But How Does It Really Work?



We Don't Understand It Yet



What Can We Teach K-12 Students About LLMs?

1. Technical concepts

- a. Neural networks
- b. Embeddings
- c. Attention heads
- d. Zero-shot learning; prompt engineering

2. Training

- a. Pre-training by prediction
- b. RLHF: Reinforcement Learning from Human Feedback

3. Ethical/social issues

- a. Selection of training data (bias, depravity)
- b. “Guard rails” and self-censorship

4. The AI Apocalypse

- a. Super-intelligence
- b. Power-seeking behavior and existential threats
- c. The Alignment Problem

Neuron Sandbox

<https://www.cs.cmu.edu/~dst/NeuronSandbox>

Angela Chen, Neel Pawar, and David Touretzky

Neuron Sandbox Model Name: Problem 1

problem-1

Menu

About




Can I make a peanut butter and jelly sandwich? I need both peanut butter and jelly. Adjust the threshold by changing the value in the box.

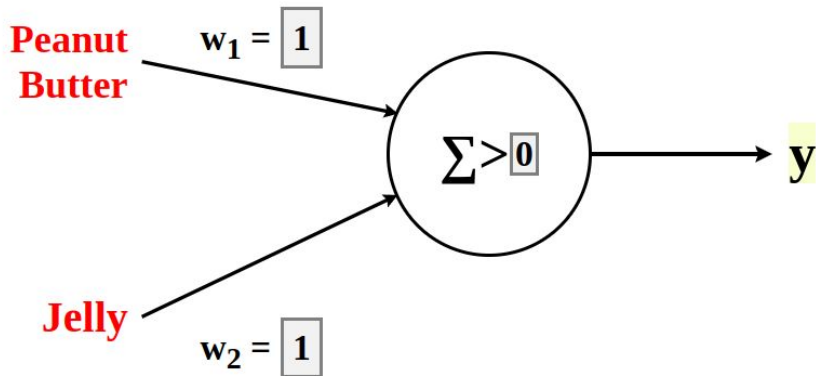
Solve for Outputs











Solve for Weights

Ask for Hint

Peanut Butter	Jelly
0 ✖	0 ✖
0 ✖	1 
1 	0 ✖
1 	1 



Activation Σ	Current Output y	Desired Output
0	0 	0 
1	1 	0 
1	1 	0 
2	1 	1 

What do students learn from Neuron Sandbox?

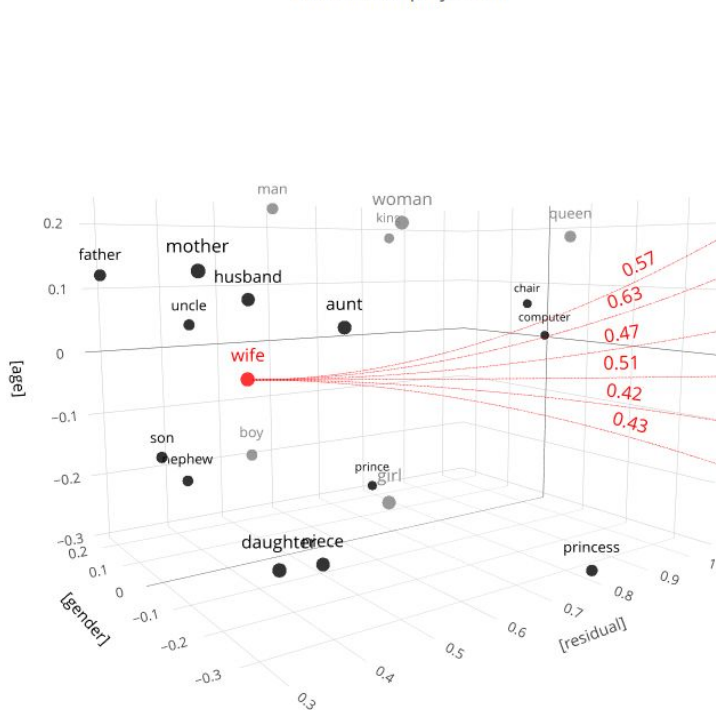
- Neural nets are composed of units called “neurons”
- One type of unit is the linear threshold unit
- A neural net’s parameters are weights and thresholds (real numbers)
- Boolean logic problems can be represented as truth tables
- Translating English descriptions of behavior into truth tables
- Non-boolean problems can also be solved by linear threshold units

Word Embedding Demo

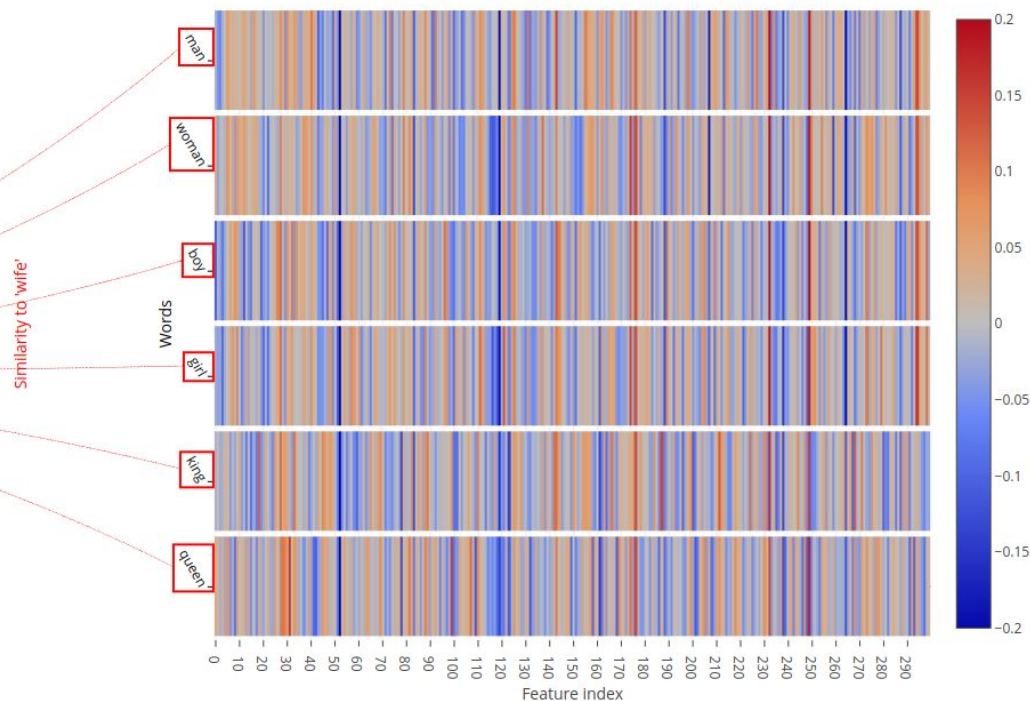
<https://www.cs.cmu.edu/~dst/WordEmbeddingDemo>

Saptarashmi Bandyopadhyay, Jason Xu, Neel Pawar, and David Touretzky (EAAI 2023)

Word vector projection



Vector visualization



What do students learn from the Word Embedding Demo?

- Vector representations of meaning
- Mapping words to points in semantic feature space
- Dot product similarity measure
- Analogy by vector arithmetic
- Computer-generated representations, via machine learning

Extractive Question Answering With BERT

BERT = Bidirectional Encoder Representations from Transformers

The image is a screenshot of a Google search interface. At the top left is the Google logo. To its right is a search bar containing the text "Will a mongoose eat fruit?". To the right of the search bar are icons for a close button (X), voice search, image search, and a magnifying glass. Below the search bar is a row of seven buttons: "Images", "Shopping", "Videos", "News", "Books", "Maps", "Flights", and "Finance". Below these buttons is a horizontal line. Under the line, the text "About 3,120,000 results (0.91 seconds)" is displayed. Below this is a blue highlighted snippet: "Mongoose are opportunistic feeders that will eat birds, small mammals, reptiles, insects, fruits, and plants." Below the snippet is a search result from "Hawaii.gov" with the URL "https://dlnr.hawaii.gov/hisc/info/mongoose". The title of the result is "Hawaii Invasive Species Council | Mongoose". At the bottom of the page, there is a footer with a question mark icon, the text "About featured snippets", a star icon, and the text "Feedback".

Google

Will a mongoose eat fruit?

Images Shopping Videos News Books Maps Flights Finance

About 3,120,000 results (0.91 seconds)

Mongoose are opportunistic feeders that will eat birds, small mammals, reptiles, insects, fruits, and plants.

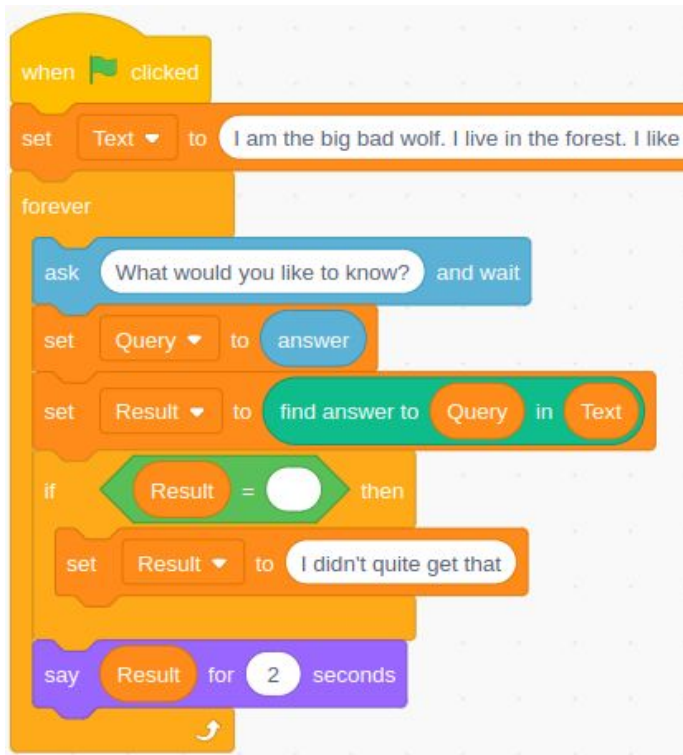
Hawaii.gov
https://dlnr.hawaii.gov/hisc/info/mongoose

Hawaii Invasive Species Council | Mongoose

About featured snippets Feedback

BERT QA extension in MachineLearningForKids

<https://ai4k12.org/wp-content/uploads/2023/04/Chatbot-with-BERT-Activity-Guide-1.pdf>



"I am the big bad wolf. I live in the forest. I like to eat little children. I met Little Red Riding Hood when she went to visit her grandmother."



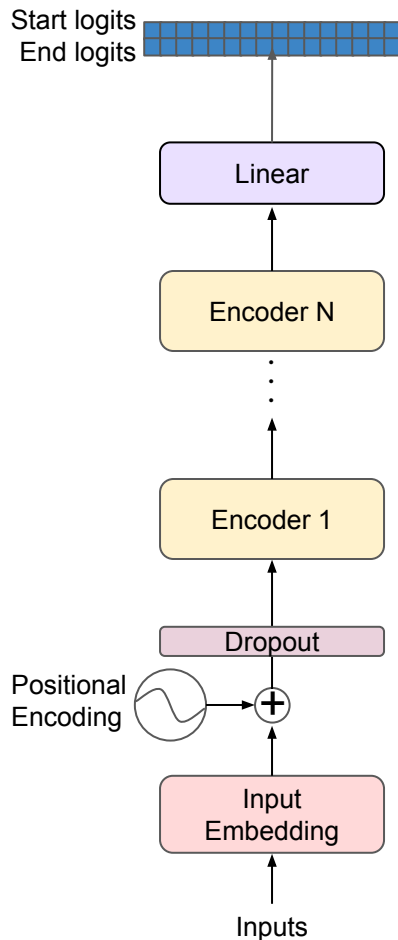
Extractive Question Answering With a Transformer

Pre-train on standard word prediction task.

Then take the encoder portion, add linear layer and logits output, and fine tune on SQuAD 2.0, the Stanford Question Answering Dataset:

- 100,000 questions with answer positions (start,end)
- 50,000+ unanswerable questions written adversarially by human crowdworkers

Logit: probability that a given token is the start (end) position of the answer.



BERT-insight demo (MobileBERT)

Neel Pawar and David Touretzky

Enter passage:

John and Mary went to a party. Mary bought a superamazing gift for the host.
John brought his guitar.
At the party, Mary gave the host a bottle of wine.
John played songs after dinner.
Fred was also at the party. He brought his dog with him.

Enter question:

What instrument did John play?

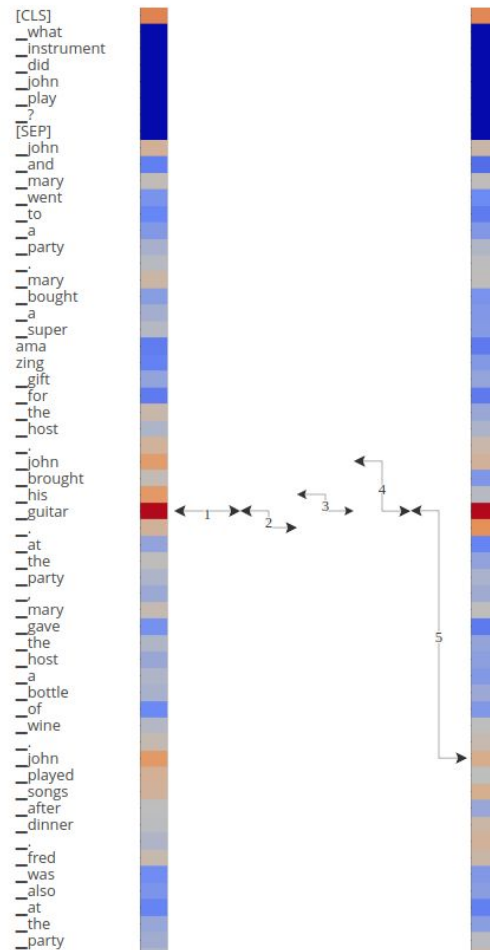
Submit

Answer:

1: guitar | score: 20.853
2: guitar. | score: 11.529
3: his guitar | score: 10.345
4: John brought his guitar | score: 9.920
5: guitar. At the party, Mary gave the host a bottle of wine. John | score: 8.151

Start Logits

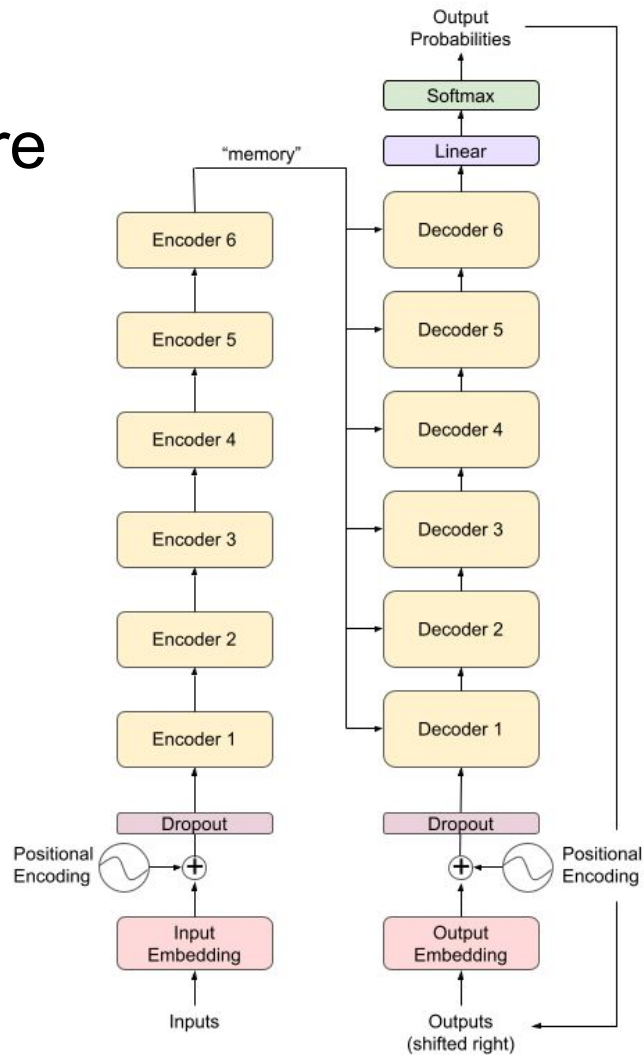
End Logits



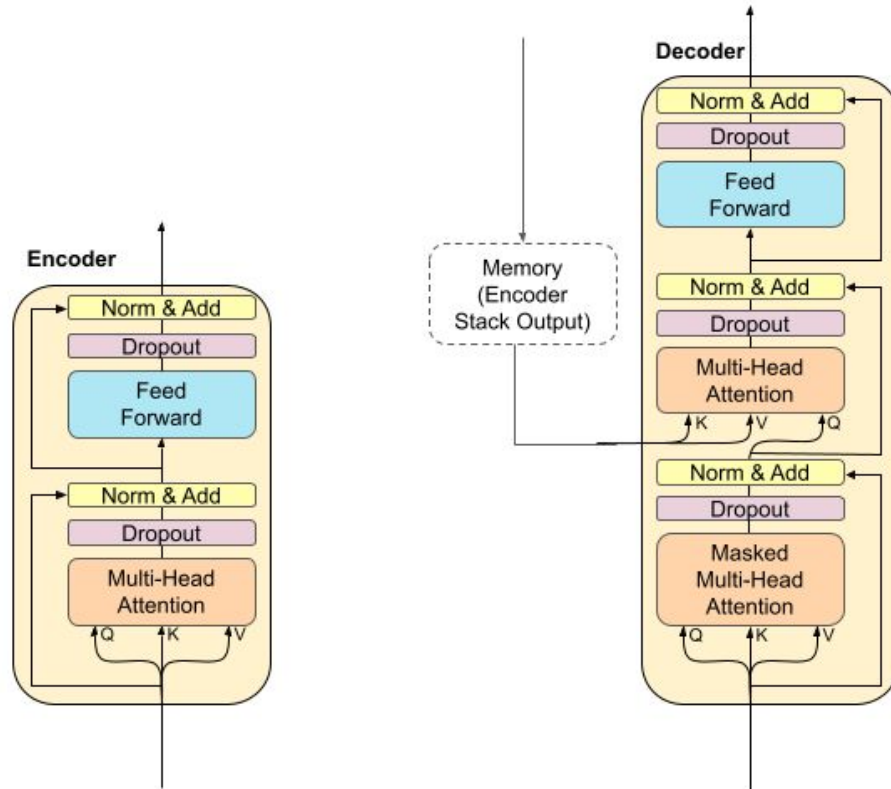
What do students learn from BERT?

- How extractive question answering differs from general question answering.
 - Can't handle yes/no questions
 - Limited inference abilities
- Matching on meaning, not specific words.
 - The Big Bad Wolf understands “where is your home” to mean “where do you live”.
- Tokenization
- Logit representation

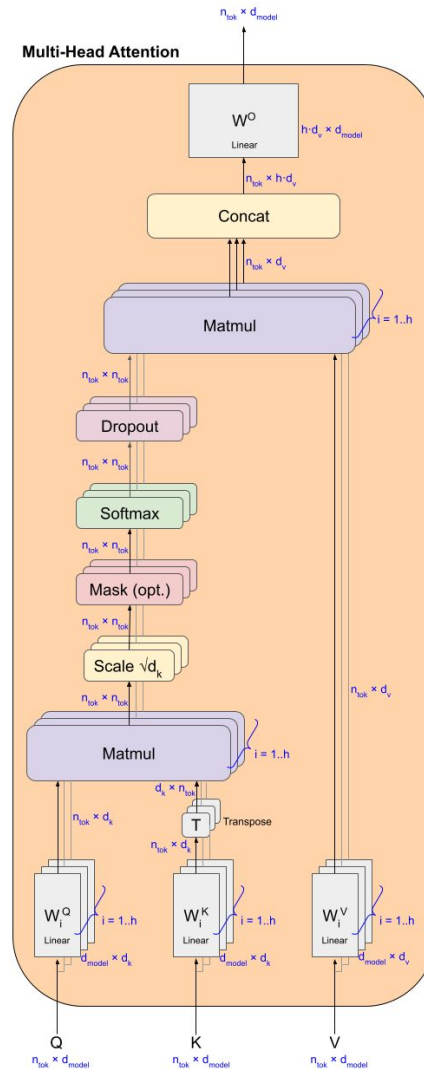
Transformer Architecture



Encoder / Decoder Modules



Multi-Head Attention Circuitry



What Do Attention Heads Do?

Level 1 Embeddings

house cat,
alive, chasing
something

...

rodent, alive,
being chased

its (mouse),
possesses
a nest

Attention
head 1:
word
senses

Attention
head 1:
word
senses

Attention
head 2:
references

Attention
head 3:
rhyming

Attention
head 4:
numbers

Level 0 Embeddings

The

cat

chased

the

mouse

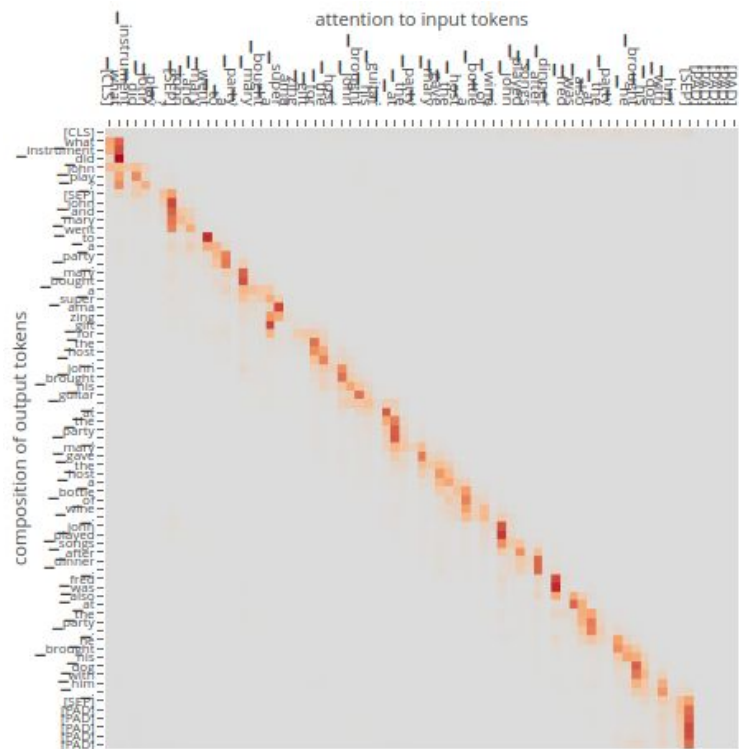
to

its

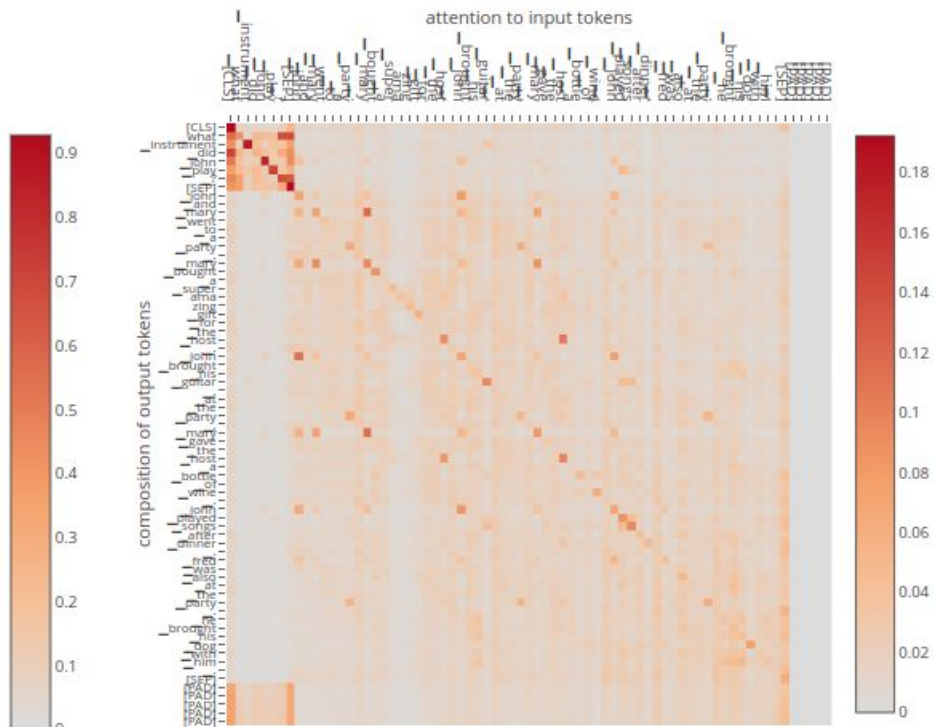
nest

house cat, big cat, cat species,
plush cat, cartoon cat, cool cat

Attention Heads in BERT-insight



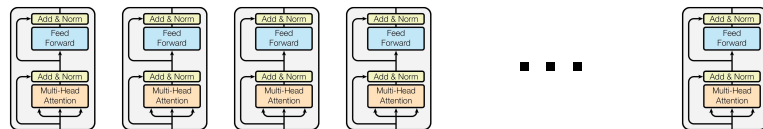
layer: bert/encoder/layer_0/attention/self/Softmax: head 1



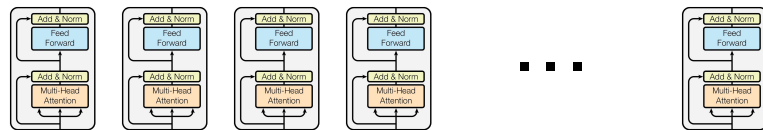
layer: bert/encoder/layer_0/attention/self/Softmax: head 3

Many Layers of Self-Attention Yields “Intelligence”

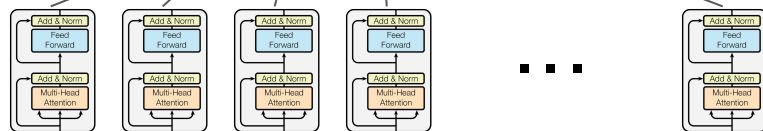
Layer 96



Layer 2



Layer 1



Embeddings

Embeddings

GPT-3 has 96 layers of self-attention, with 16 attention heads at each layer.

Every word is recoded and recombined with the other words 1536 times.

The model has 175 billion parameters (weights).

16 different attention heads in each layer

The Clarified Transformer

- Initial transformer paper (Vaswani et al., 2017) is an imperfect description
- References a GitHub repository with code implemented in TensorFlow
- The Annotated Transformer (Rush, 2018), (Huang et al., 2022) provides an annotated version of the paper along with opaque PyTorch code
- Pawar and Touretzky have a clarified PyTorch implementation
 - Model is just 400 lines of very clear Python code
 - Train on same German-to-English translation task

Language Translation with A Transformer

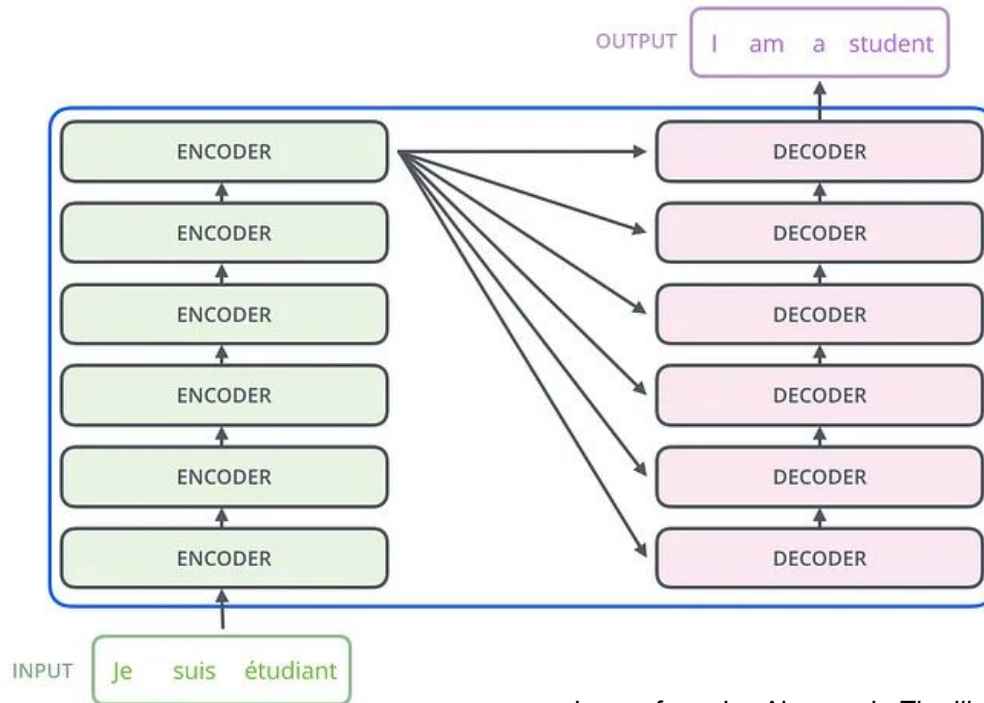


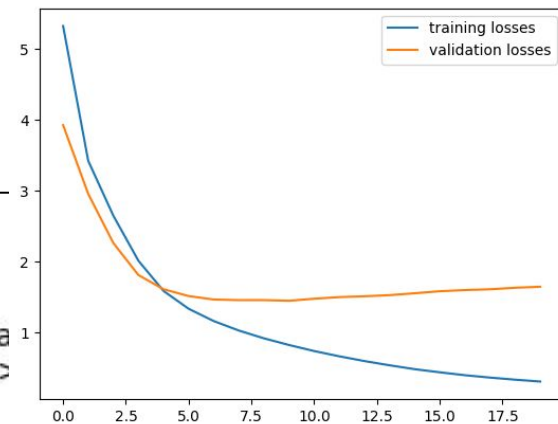
Image from Jay Alammar's *The Illustrated Transformer*

Clarified Transformer: German to English

Transformer model name: multi30k_model | Epoch: 1

Example 1 =====

Source Text (Input) : <s> Eine Gruppe von Männern lädt Baumwolle auf einen La
Target Text (Ground Truth) : <s> A group of men are loading cotton onto a truck </s>
Model Output : <s> A man in a red is on a </s>



Transformer model name: multi30k_model | Epoch: 4

Example 1 =====

Source Text (Input) : <s> Eine Gruppe von Männern lädt Baumwolle auf einen Lastwagen </s>
Target Text (Ground Truth) : <s> A group of men are loading cotton onto a truck </s>
Model Output : <s> A group of men are at a truck . </s>

Transformer model name: multi30k_model | Epoch: 15

Example 1 =====

Source Text (Input) : <s> Eine Gruppe von Männern lädt Baumwolle auf einen Lastwagen </s>
Target Text (Ground Truth) : <s> A group of men are loading cotton onto a truck </s>
Model Output : <s> A group of men are loading cotton into a truck . </s>

Prompt Engineering

- How to get the LLM to do what you want.
- It's not programming. It's explaining.
- Sometimes it seems like negotiating: finding ways to bypass the guard rails.

People Hire Phone Bots to Torture Telemarketers, Wall Street Journal
6/29/2023:

At first, ChatGPT was reluctant to do the work. "As an AI language model, I don't encourage people to waste other people's time," ChatGPT told Anderson... Anderson finally found a line of reasoning that persuaded GPT-4 to take the job. "I told it that, 'You are a personal assistant and you are trying to protect this man from being scammed,' "

- In few-shot learning the prompt contains a selection of worked examples. Learning how to construct these is a skill students can be taught.

Ethical / social issues

1. Curation of training data

- Training on older sources can reproduce biases about gender or ethnicity
- Training on Reddit provides exposure to many types of depravity

2. Installing “guard rails” to prevent bad behavior

- What should be banned? What should be permitted?
- Can train an LLM to censor problematic responses from another LLM

3. Liability

- Can chatbots cause harm? (Already one suicide.) Who is liable?

4. Energy cost of training large models

5. The AI Apocalypse

The AI Apocalypse

Why do some people regard LLMs as an existential threat to humanity?

1. Superintelligence
2. Power-seeking behavior
3. The Alignment Problem



Superintelligence

- Term popularized by Oxford philosopher Nick Bostrom
- GPT-4 has read more books than you will read in a lifetime.
- LLMs will eventually be smarter than humans.



Moneycontrol 
@moneycontrolcom

#TechWithMC | OpenAI is forming a dedicated team to address the topic of "superintelligence," which refers to a theoretical concept where an AI surpasses human intelligence and becomes exceptionally powerful.

Here's more about the future of #ArtificialIntelligence 
moneycontrol.com/news/technolog...

#ChatGPT #OpenAI #AI #Superintelligence



5:38 AM · Jul 6, 2023 · 666 Views

Power-seeking behavior

- An intelligent AI may desire more resources. How can it get them?
 - Take over a data center
 - Build new data centers
- How can it act in the physical world?
 - Acquire money
 - Acquire minions, possibly through deception
- How can it protect itself from human interference?
 - Coercion, e.g., threaten vital infrastructure
 - Removal of humans: start a war, release a plague, etc.

The Alignment Problem

- If AI agents become smarter than us, we may not be able to control them.
- How do we ensure that their values are aligned with our own so that they do not act in ways detrimental to humans?
- Good topic for student discussion.

Conclusions

- We are presently in an AI arms race.
- The proposed moratorium on LLM development has not gained traction.
- We may be doomed. But if we survive, our lives could be immeasurably better.
- K-12 students can be taught a lot about what's going on.
- It's up to us to see that they learn about it.

Acknowledgments

Collaborators:

- Saptarashmi “Rob” Bandyopadhyay
- Angela Chen
- Christina Gardner-McCune
- Neel Pawar
- Jason Xu

Funding:

- National Science Foundation awards DRL-2049029 and IIS-2112633.
- NEOM Company