What is a LLM and Why Should I Care?

Casey Kennington

BOISE STATE UNIVERSITY

ISPE
February 23, 2024

What is happening

- I woke up one mornin SINGLE POST had something to do with
- ChatGPT is a “large [
  - LLM (more on this lat
  - GPT = Generative, pr
- Companies are pivoti
  - Microsoft invested 10
  - Buzzfeed is using Cha
- There is also generat ill focus today only on LLMs

Bio

Casey Kennington

Boise State University
Associate Professor
Computer Science

- PhD in Linguistics (Computational), Bielefeld University, Germany
- Teaching:
  - Data Science (introductory undergraduate, graduate)
  - Natural Language Processing (upper-division undergraduate, graduate)
    - LLMs since 2020
  - Research Methods in Deep Learning & Spoken Dialogue Systems (Graduate)

Related Research

- Master’s thesis on language modeling (2011)
- Enriching Large Language Models (LLMs) with multimodal information to improve semantic knowledge (NSF CAREER)
- Developing and maintaining infrastructure for incremental, interactive, spoken dialogue systems (RETICO project)
- Making LLMs smaller, incorporate multiple modalities, learn through interaction (NSF CAREER)
- Developing a model of robot emotion generation and recognition (NSF CAREER)
- Child Assisted Search Tool (NSF CISE)

What is happening with AI?

Text in (in the form of a prompt, e.g., “write an email responding to …”

Text out (“Dear Mr. Tannen,

I always appreciate questions from concerned parents …”

A Brief History of Language Models

What is a language model?

A **LM** is something that captures sequential information. Most commonly, a sequence of words in a body of text.

Example:

> I would like two scoops of …. ?

“I hear the word ‘model’ but what even is that?”

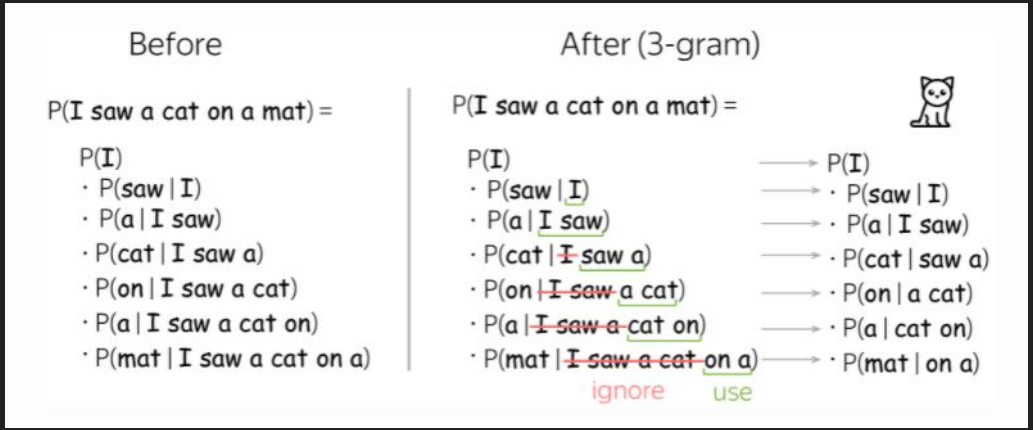
- A model is a formal (often mathematical) representation of a phenomenon

What do you mean by “model”?

Language is often “modeled” as the probability that a word follows an observed sequence of words.

How far back in a text should we go to predict what comes next?

$$P(w_i|w_{i-n+1}, ..., w_{i-1}) \approx \frac{C(w_{i-n+1}, ..., w_i)}{C(w_{i-n+1}, ..., w_{i-1})}$$



https://www.analyticsvidhya.com/blog/2022/01/building-language-models-in-nlp/

What n-gram LMs are not good for

Long-distance dependencies

E.g.: “I met Biff last week. I have to say, of all people I have ever met, he …”

Counting up words and word sequences doesn’t capture meaning of words.

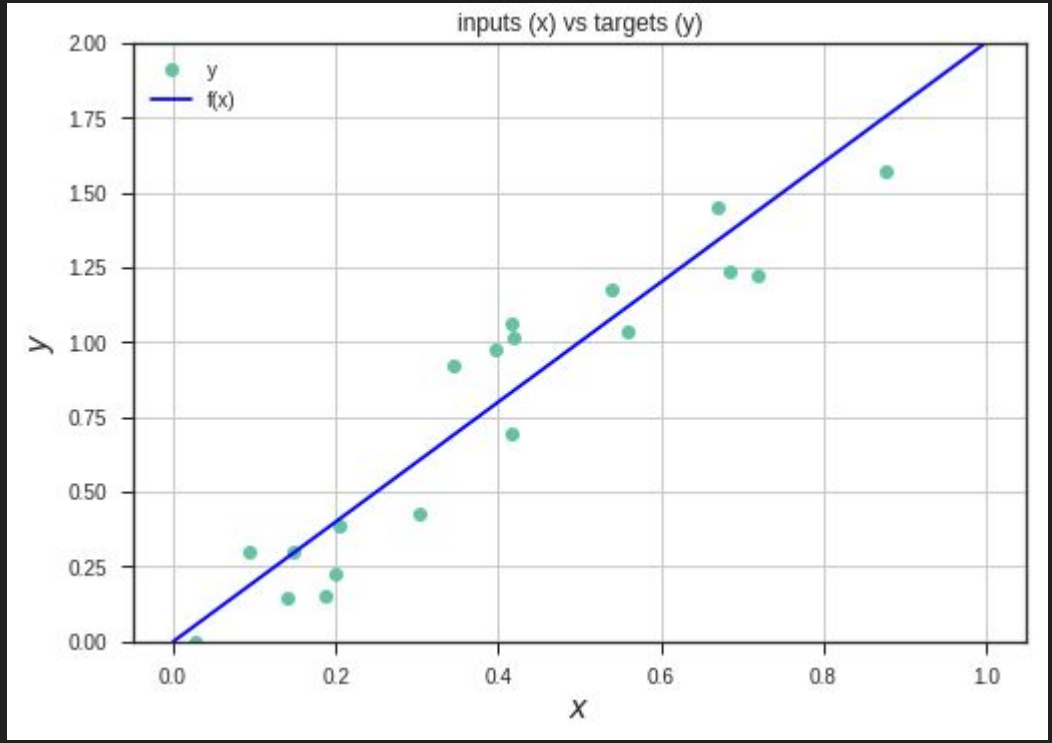
Doesn’t actually play nicely with general purpose machine learning classifiers.

Neural Networks

Simple idea: function that maps input to output

Needed:

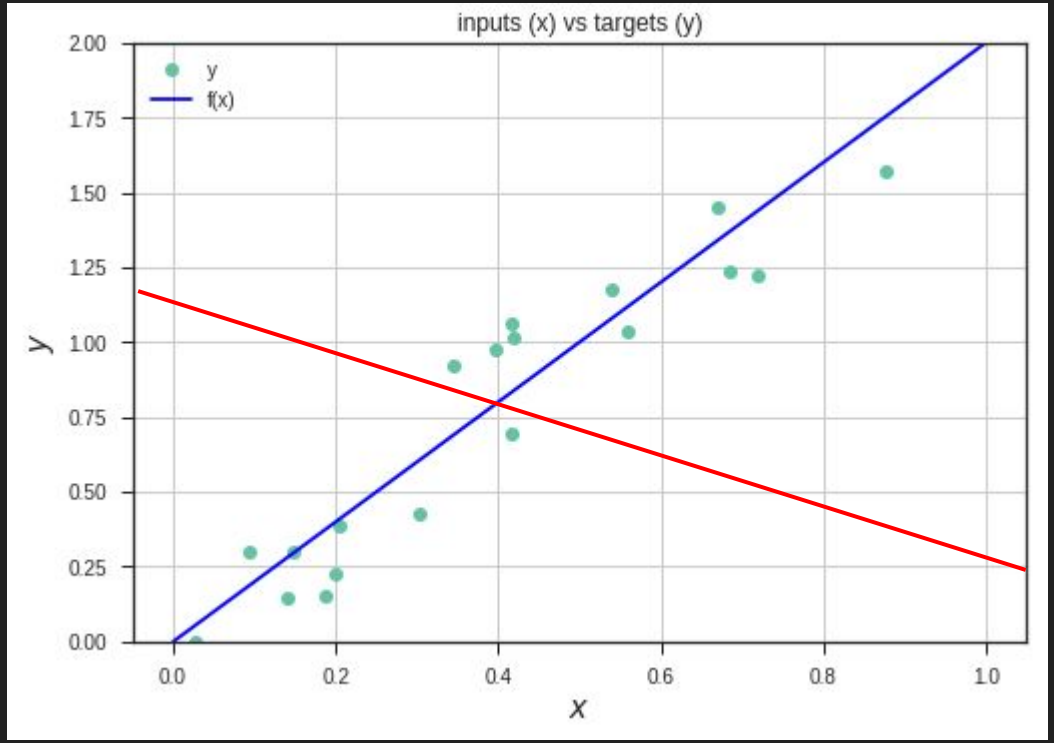
1. Function (y=mx+b)
2. Cost Function
   (sum of residuals)
3. Update
   (change m & b, repeat)



Simple idea: function that maps input to output

Needed:

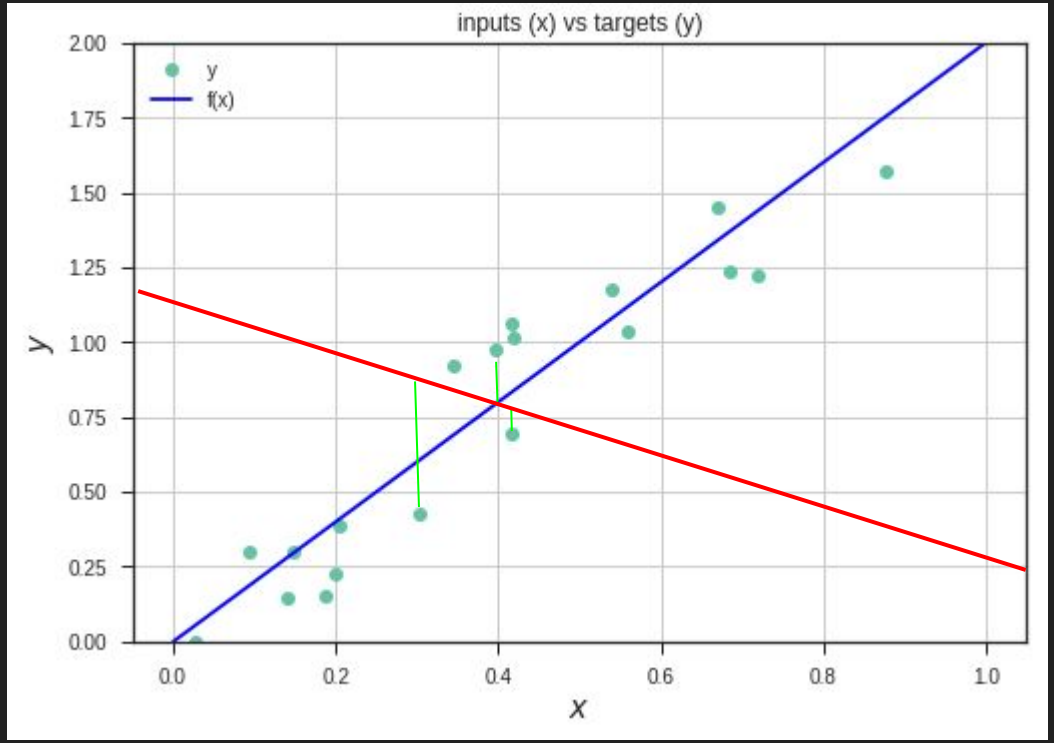
1. Function (y=mx+b)
2. Cost Function
   (sum of residuals)
3. Update
   (change m & b, repeat)



Simple idea: function that maps input to output

Needed:

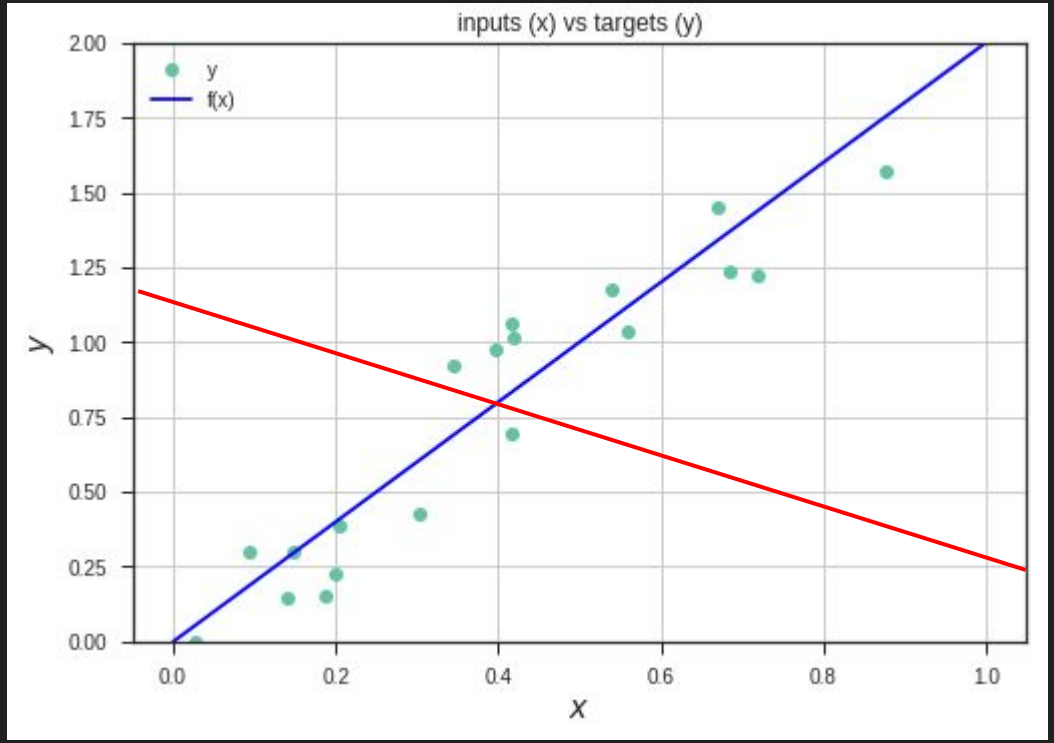
1. Function (y=mx+b)
2. Cost Function
   (sum of residuals)
3. Update
   (change m & b, repeat)



Simple idea: function that maps input to output

Needed:

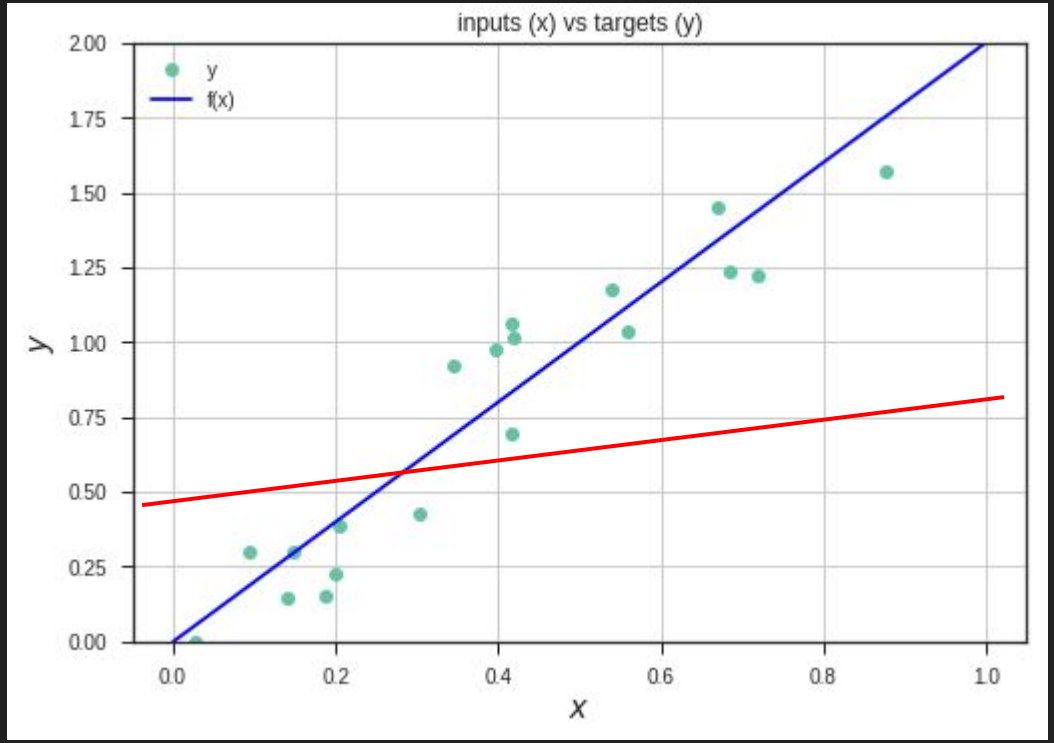
1. Function (y=mx+b)
2. Cost Function
   (sum of residuals)
3. Update
   (change m & b, repeat)



Simple idea: function that maps input to output

Needed:

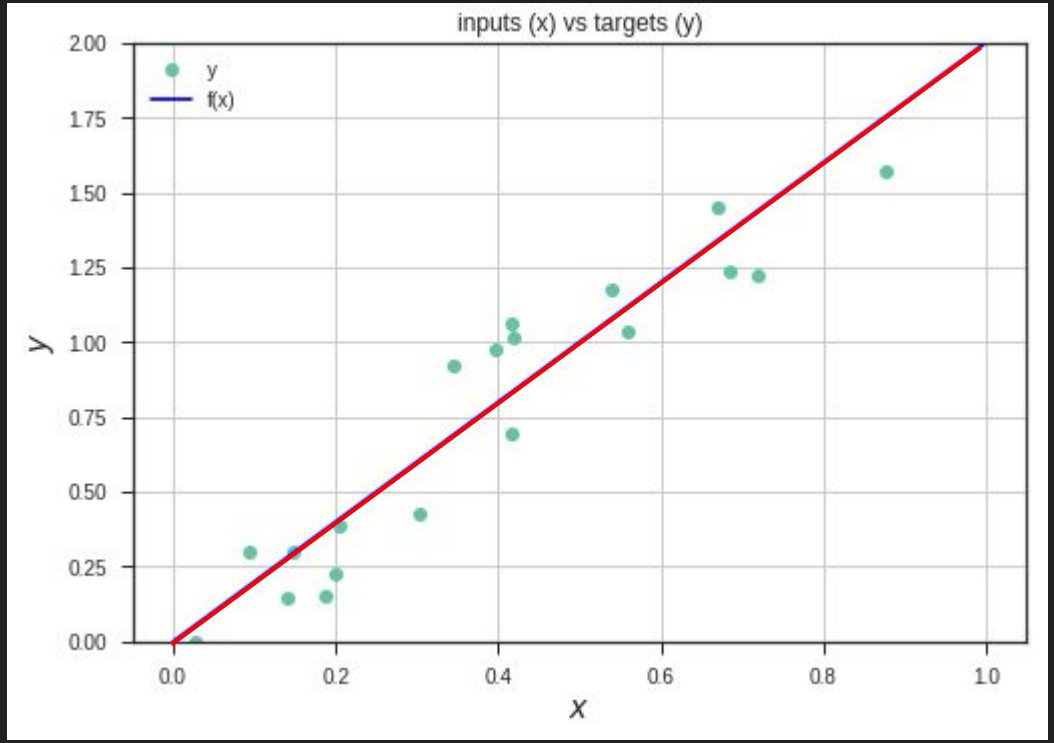
1. Function (y=mx+b)
2. Cost Function
   (sum of residuals)
3. Update
   (change m & b, repeat)



Simple idea: function that maps input to output

Needed:

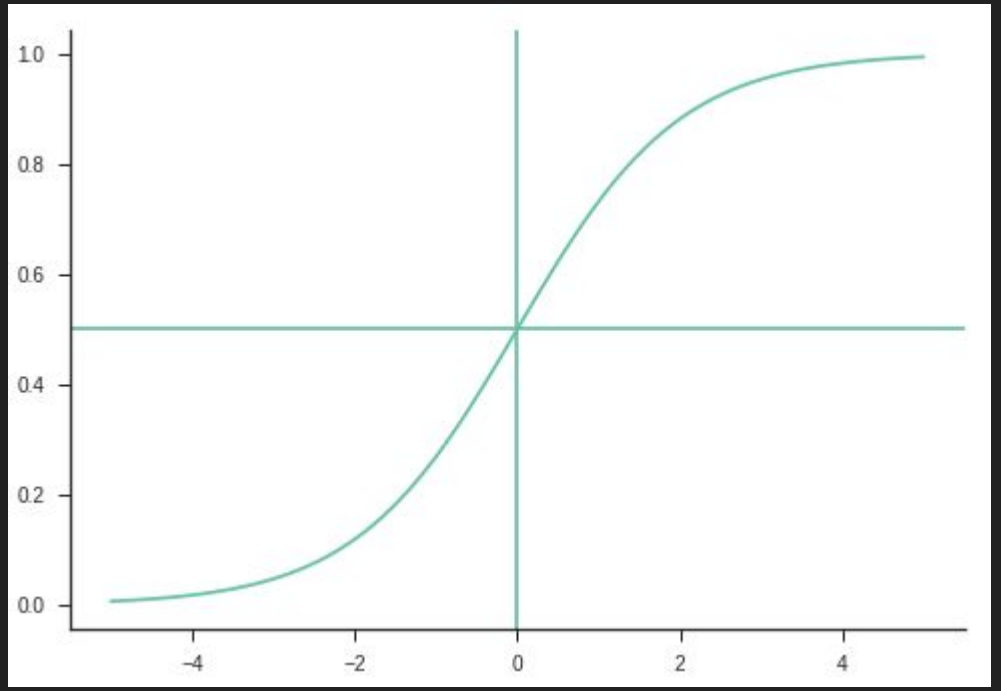
1. Function (y=mx+b)
2. Cost Function
   (sum of residuals)
3. Update
   (change m & b, repeat)



Simple idea: function that maps input to output in probability space

Needed:

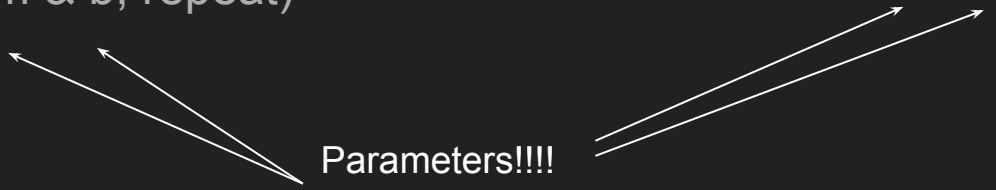
1. Function ($f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$)
2. Cost Function
   (distance of points)
3. Update
   (change k & L, repeat)



Simple idea: function, cost, update

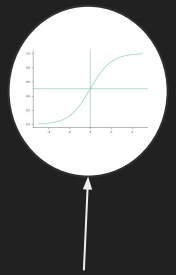
1. Function (y=mx+b)
2. Cost Function
   (sum of residuals)
3. Update
   (change m & b, repeat)

1. Function ($f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$)
2. Cost Function
   (distance of points)
3. Update
   (change k & L, repeat)

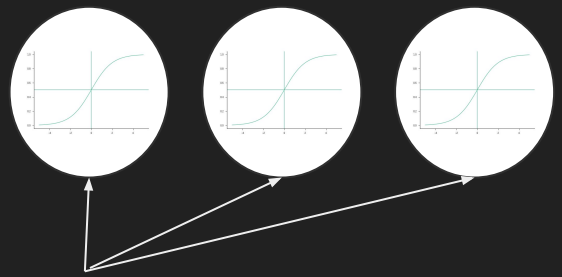


Parameters!!!!
(weights, coefficients)

Simple idea: stack classifiers together

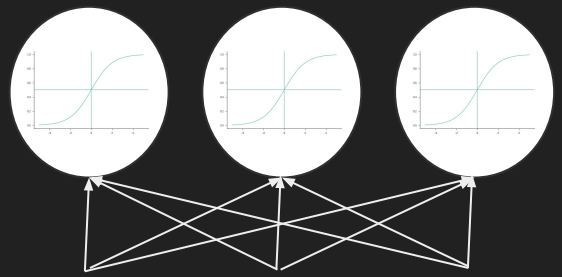


Simple idea: stack classifiers together



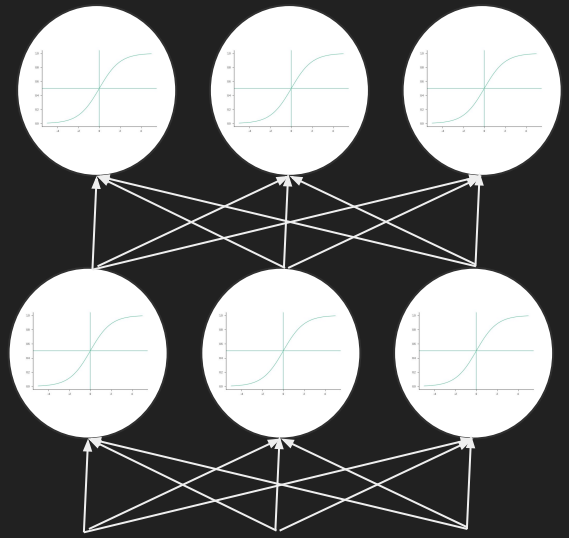
Each classifier has
two parameters

Simple idea: stack classifiers together



Each classifier has
four parameters

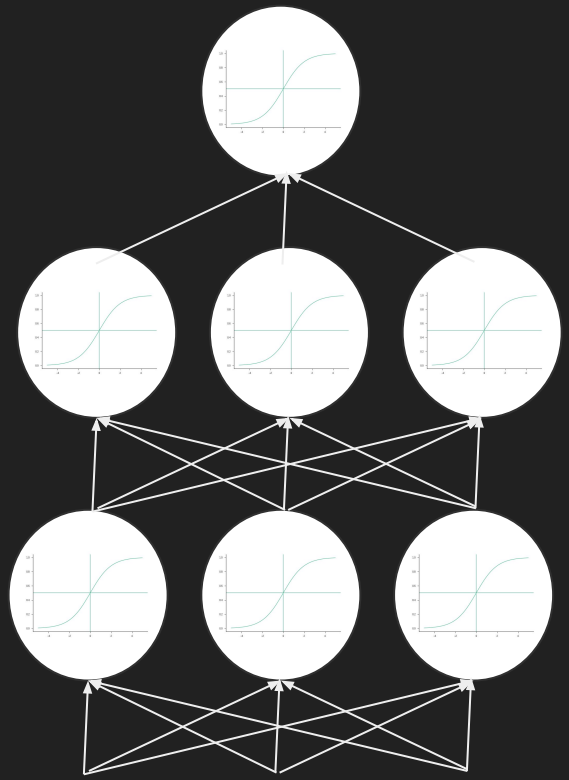
Simple idea: stack classifiers together



Each classifier has four parameters

Each classifier has four parameters

(Deep) Neural Network



classifier has four parameters

Each classifier has four parameters

Each classifier has four parameters

Needed:

1. Function ($f(x) = \frac{L}{1 + e^{-k(x - x_0)}}$)
2. Cost Function
   (difference in output distribution)
3. Update
   (change top layer, then next layer down, repeat)

How do we make LMs better?

Long-distance dependencies

E.g.: “I met Biff last week. I have to say, of all people I have ever met, he …”

Counting up words and word sequences doesn’t capture meaning of words.

Doesn’t actually play nicely with general purpose machine learning classifiers.

How do we make LMs better?

Long-distance dependencies

E.g.: “I met Biff last week. I have to say, of all people I have ever met, he …”

Counting up words and word sequences doesn’t capture meaning of words.

Doesn’t actually play nicely with general purpose machine learning classifiers.

Many deep learning architectures are slow to train and run

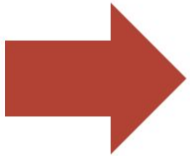
Computational Semantics

What is semantics?

- Meaning of words, phrases, sentences, paragraphs, documents
- How do humans learn meaning of words?
  - “red” (concrete)
  - “democracy” (abstract)
- Highly interactive, co-located with other people
- How do we convey meaning to machines, which ultimately work with ones and zeros?
  - Machine learning models require inputs to be continuous numbers

one-hot vectors

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch



	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Each word gets a 1x9 vector representation

https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/one-hot-word-embedding-vectors/

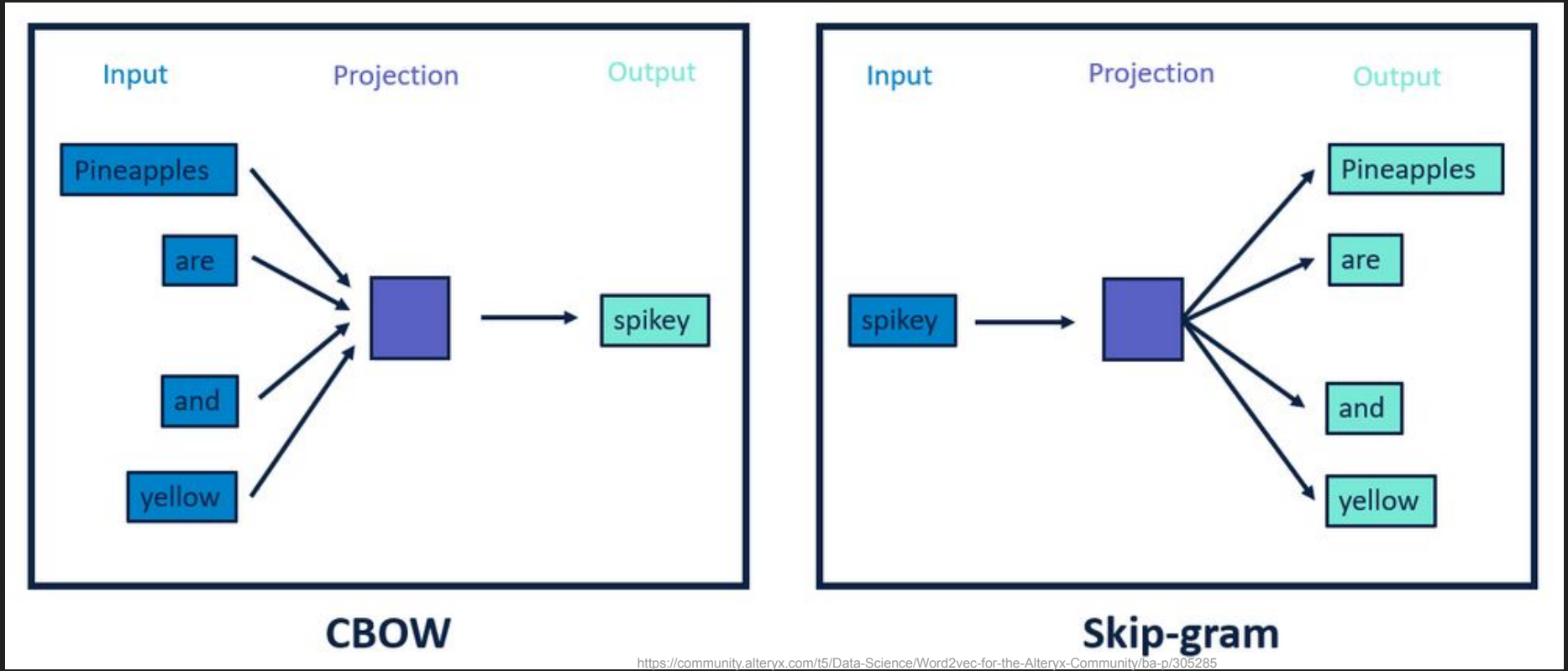
	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	1	0	1	1	0	0	1	0
information	1	0	1	1	0	0	1	0

(example from Jurafsky and Martin, 2008)

apricot: { boil, large, sugar, water }
pineapple: { boil, large, sugar, water }
digital: { arts, data, function, summarized }
information: { arts, data, function, summarized }

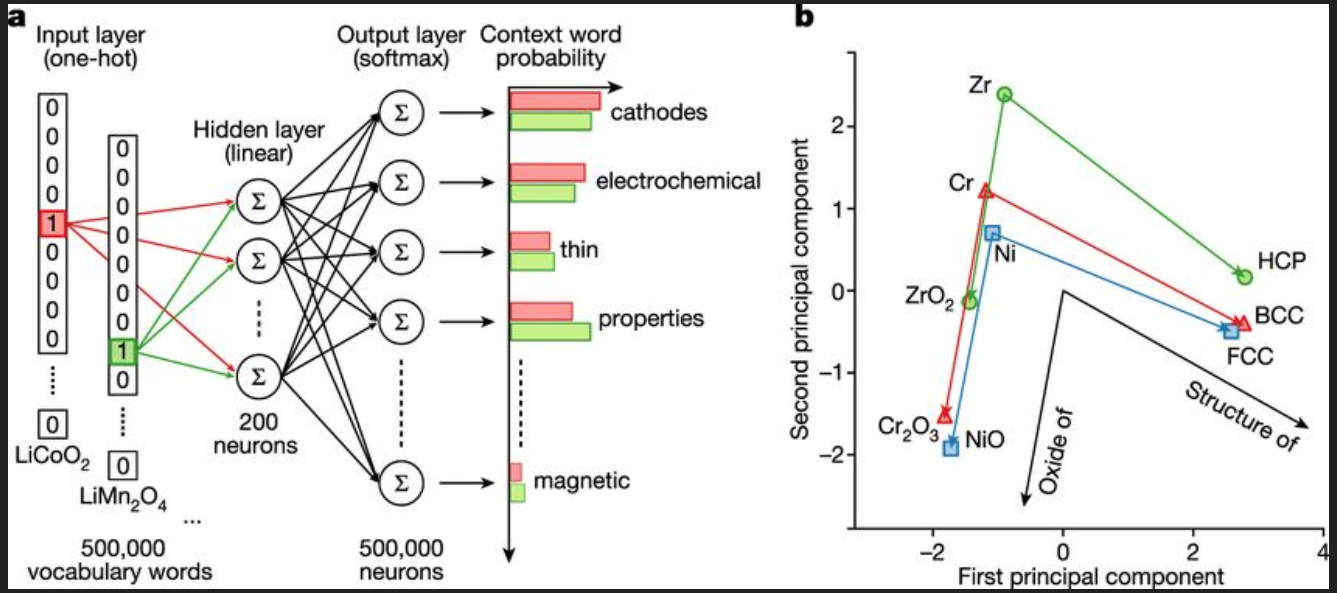
Clustering: group together words with ‘similar’ vectors.

word2vec (Mikolov et al., 2013)



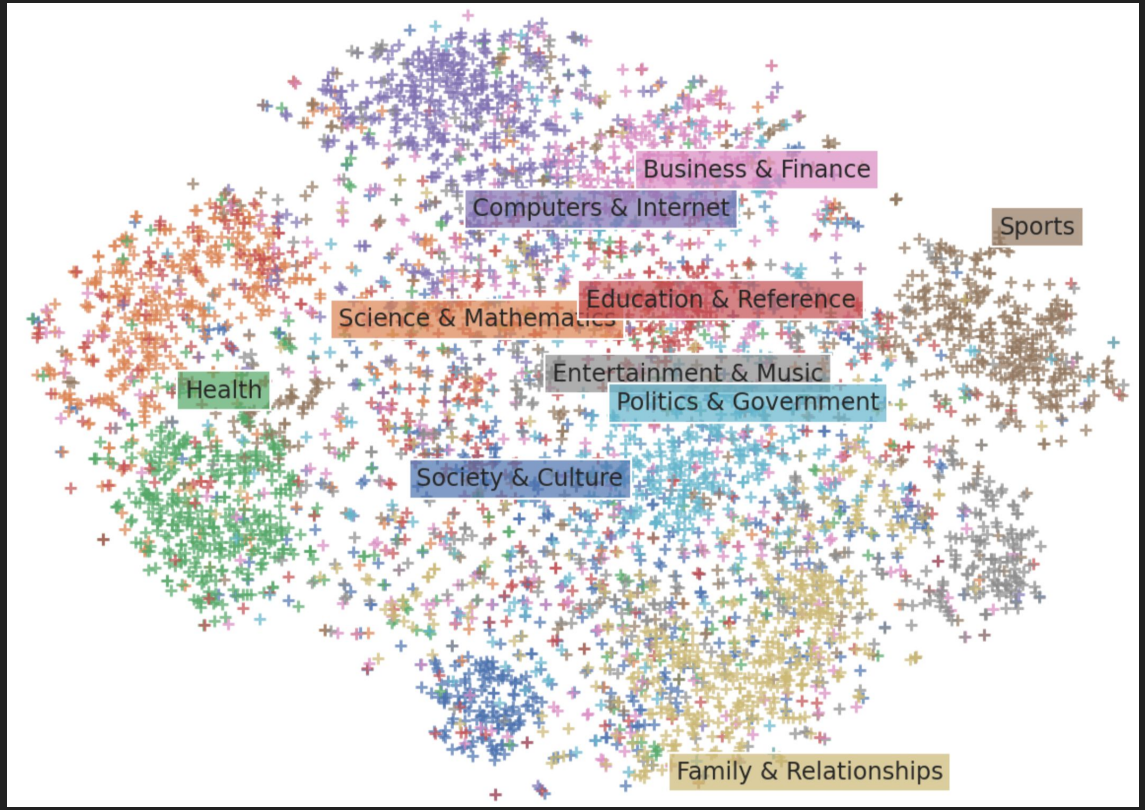
https://community.alteryx.com/t5/Data-Science/Word2vec-for-the-Alteryx-Community/ba-p/305285

word2vec (Mikolov et al., 2013)



Tshitoyan, Vahe & Dagdelen, John & Weston, Leigh & Dunn, Alexander & Rong, Ziqin & Kononova, Olga & Persson, Kristin & Ceder, Gerbrand & Jain, Anubhav. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature. 571. 95-98. 10.1038/s41586-019-1335-8.

Embeddings in 2 dimensions (tsne)

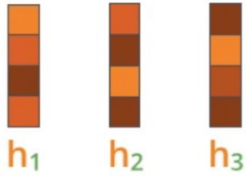


https://joeddav.github.io/blog/2020/05/29/ZSL.html

Attention & Transformers

Attention at time step 4

1. Prepare inputs



Encoder hidden states

Decoder hidden state at time step 4

2. Score each hidden state

13	9	9

scores
Attention weights for decoder time step #4

3. Softmax the scores

0.96	0.02	0.02

softmax scores

4. Multiply each vector by its softmaxed score

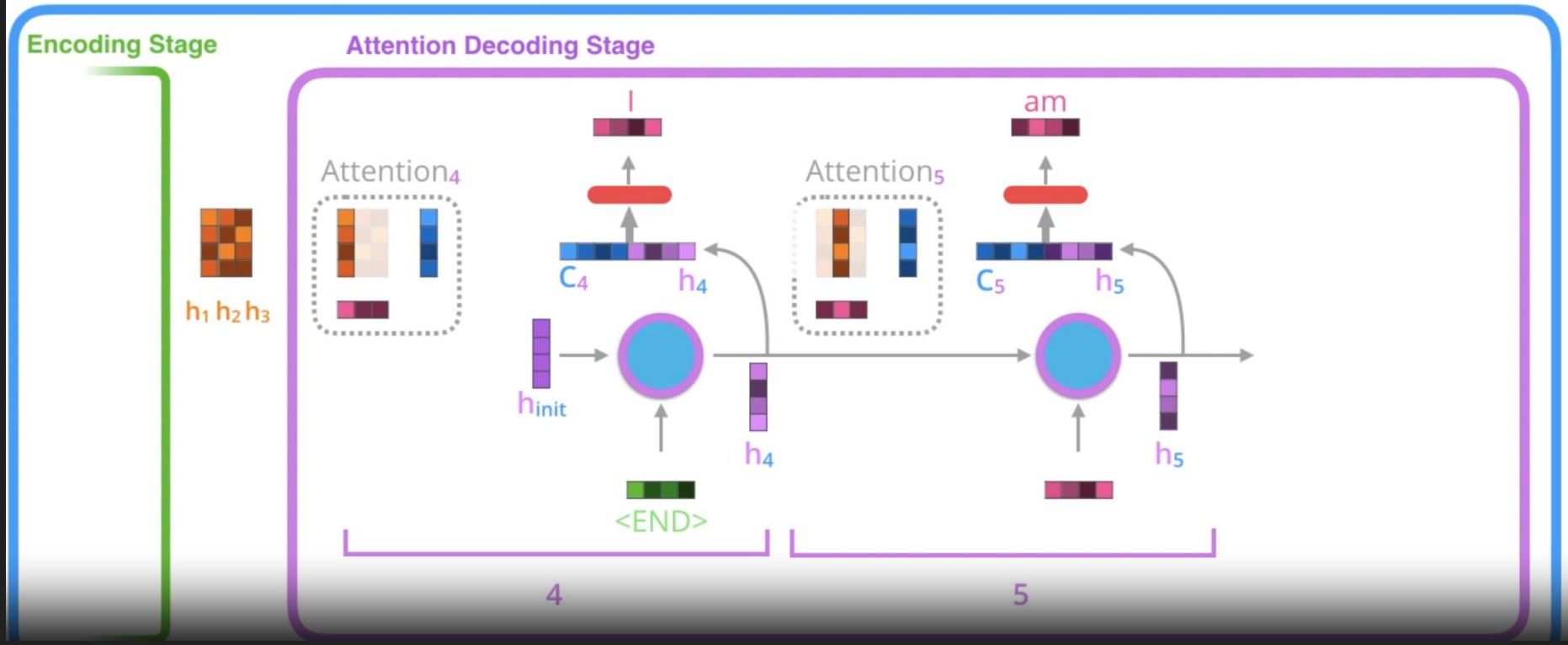
=

5. Sum up the weighted vectors

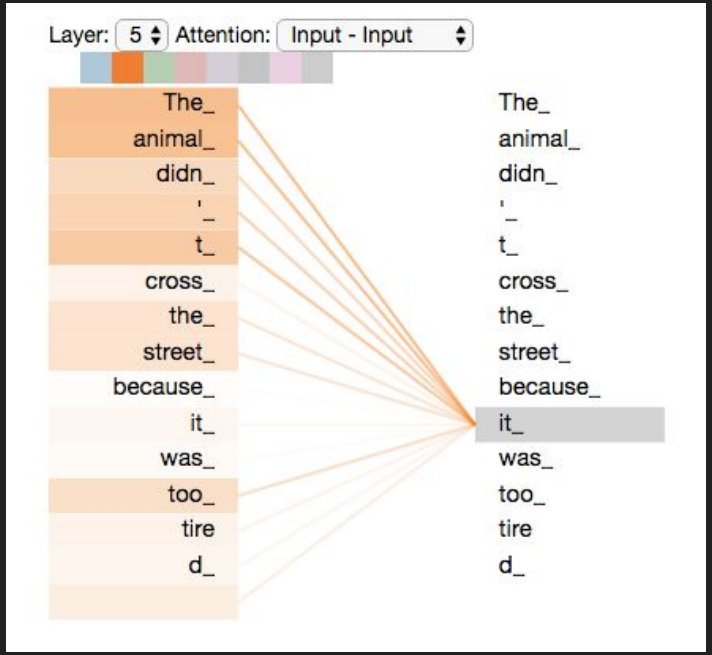
Context vector for decoder time step #4

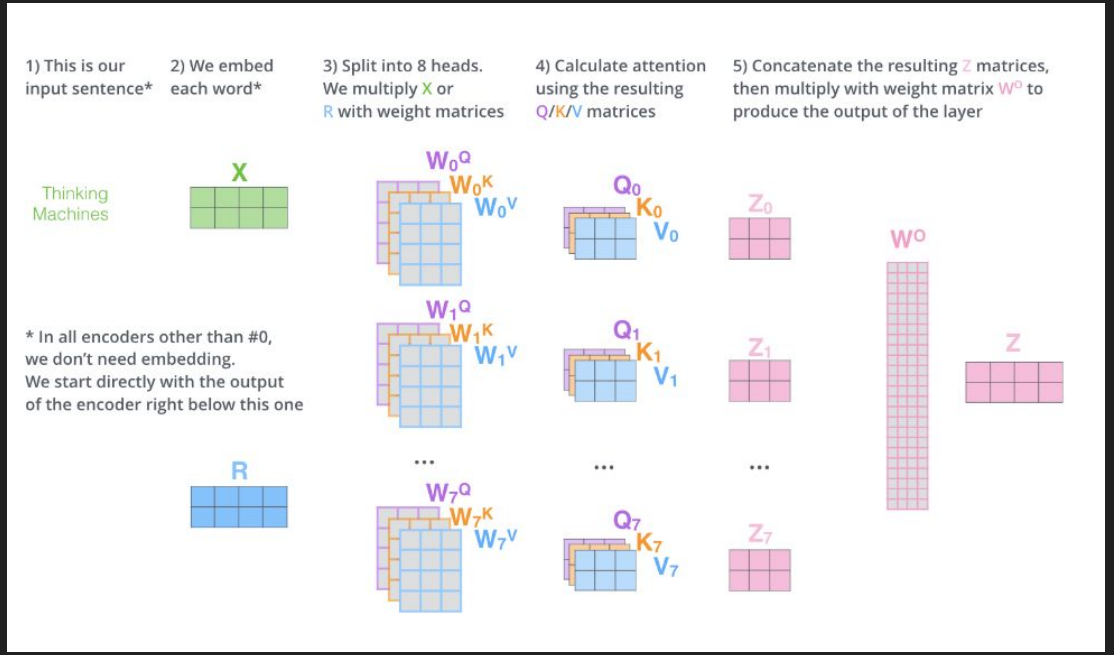
Neural Machine Translation

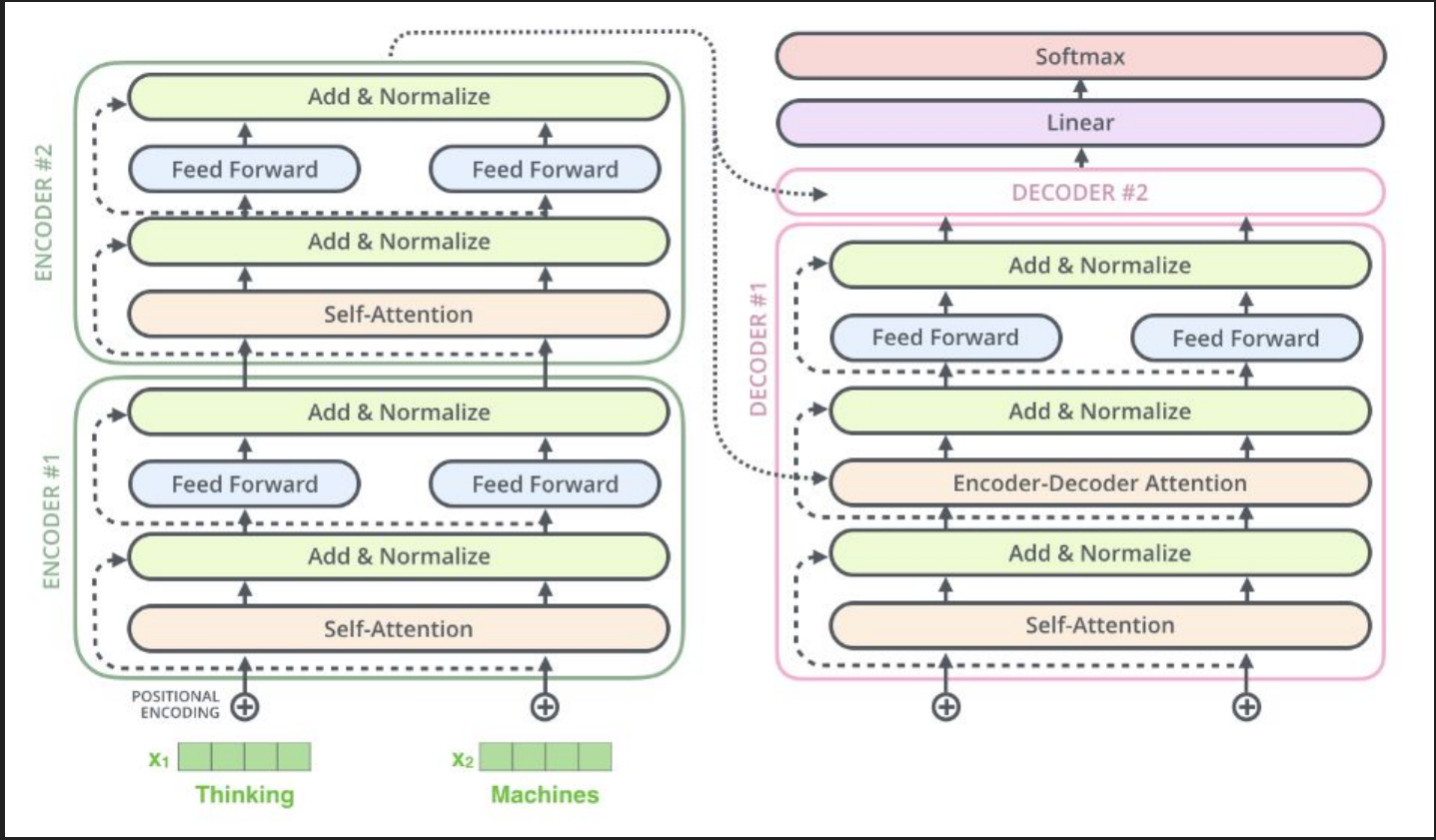
SEQUENCE TO SEQUENCE MODEL WITH ATTENTION



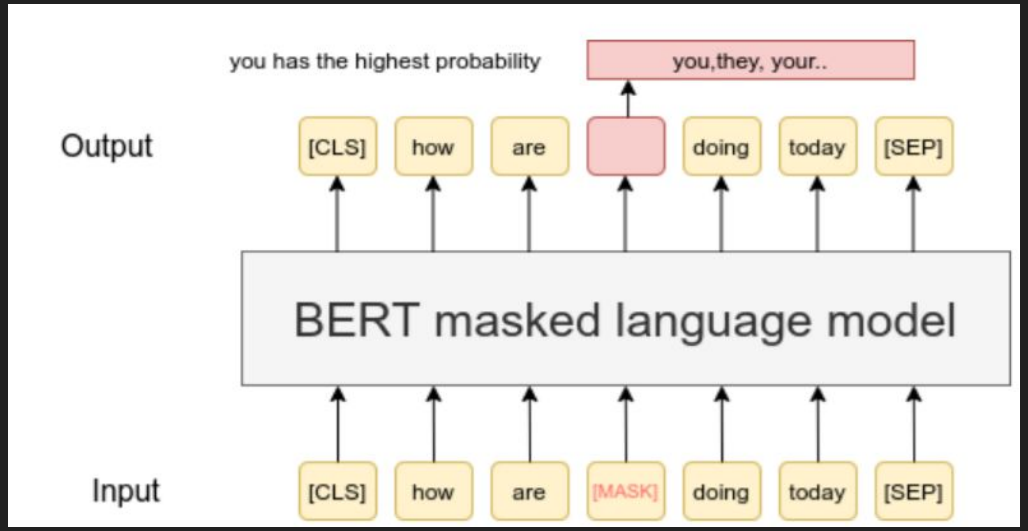
http://jalammar.github.io/illustrated-transformer/



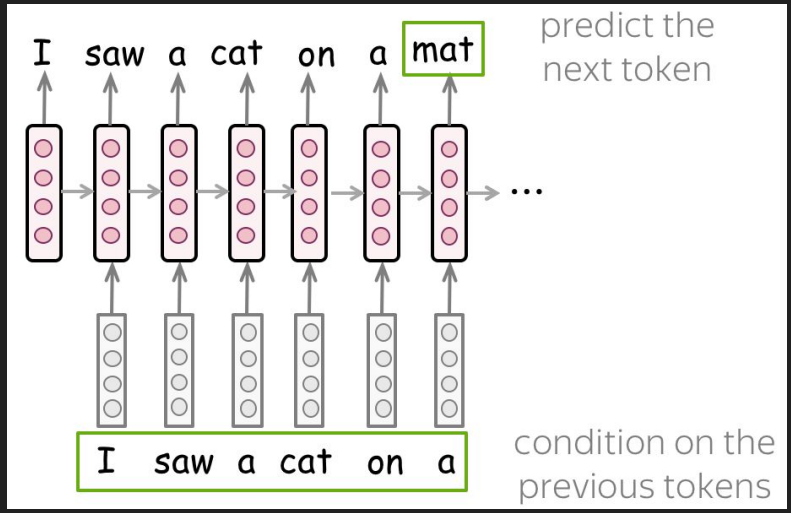



Softmax
Linear
DECODER #2
DECODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Encoder-Decoder Attention
Add & Normalize
Self-Attention
ENCODER #2
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
POSITIONAL ENCODING
x1
x2
Thinking
Machines

Training LLMs



https://www.sbert.net/examples/unsupervised_learning/MLM/README.html



https://lena-voita.github.io/nlp_course/language_modeling.html

Boise, Idaho

122 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

Coordinates: 43°36′57″N 116°12′6″W

"Boise" redirects here. For other uses, see Boise (disambiguation).

Boise (locally /ˈbɔɪsi/ (listen) *BOY-see*)[5] is the [redacted] and most populous city of the U.S. state of Idaho and is the county seat of Ada County. As of the 2020 census,[6] there were 235,684 [redacted] residing in the city. On the Boise River in southwestern Idaho, it is 41 miles (66 km) east of the Oregon border and 110 miles (177 km) [redacted] of the Nevada border. The downtown area's elevation is 2,704 feet (824 m) above sea level.

The Boise metropolitan area, also known as the Treasure Valley, includes five counties with a combined population of 749,202, the most populous metropolitan area in Idaho. It contains the state's three largest cities: Boise, Nampa, and Meridian. The Boise–Nampa Metropolitan Statistical Area is the 77th most populous metropolitan statistical area in the United States.

Downtown Boise is the cultural center and home to many small businesses and a number of high-rise buildings. The area has a variety of shops and restaurants. Centrally, 8th Street contains a pedestrian zone with sidewalk cafes and restaurants. The neighborhood has many local restaurants, bars, and boutiques. The area also contains the Basque Block, which showcases Boise's Basque heritage.

[redacted]

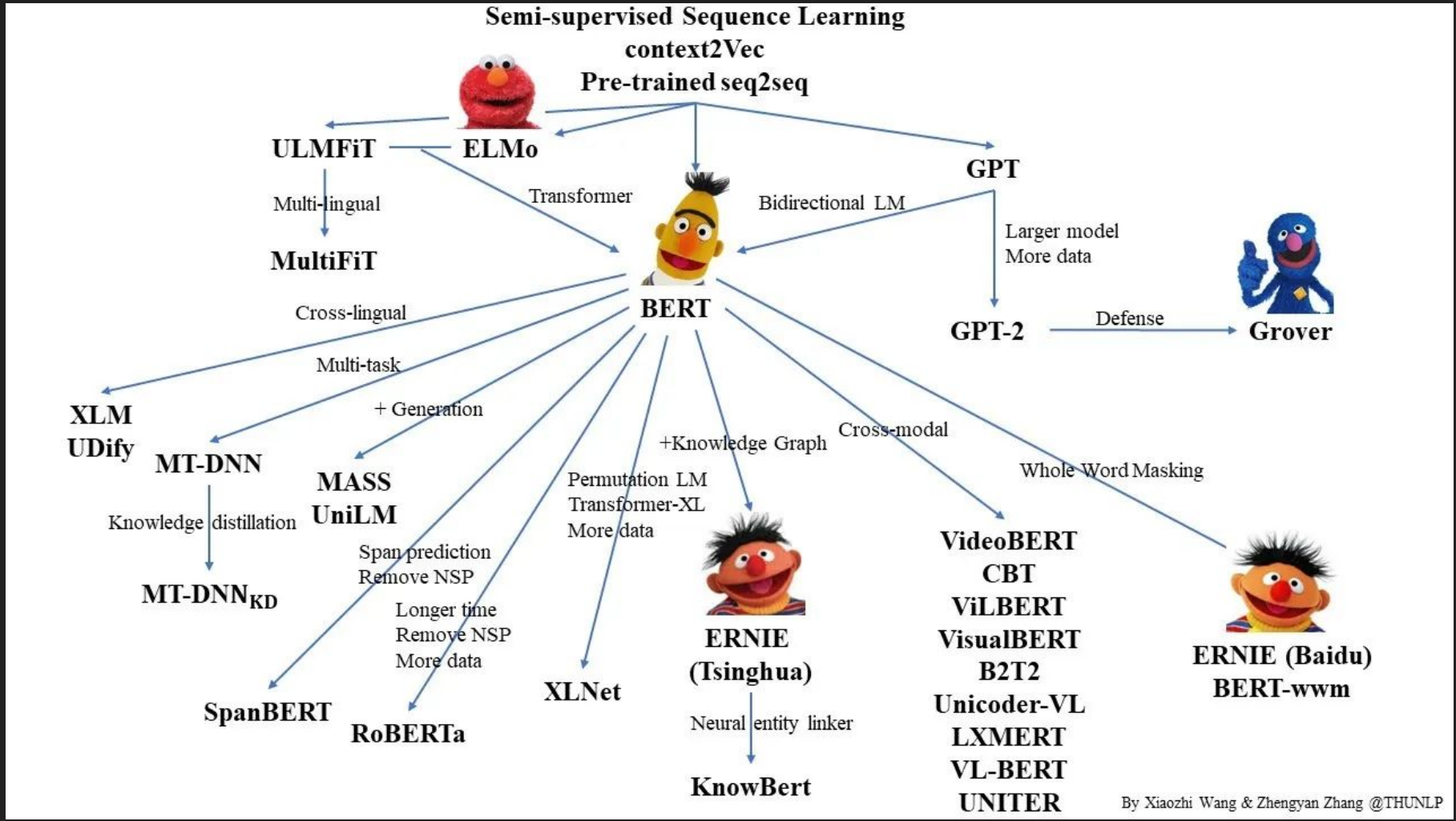
Boise

State capital city

Idaho State Capitol

Powerful idea: pre-training and fine-tuning

- Pre-train using generic language tasks that don’t require supervision like masking and next-word/sentence prediction. Only needs lots of text and time. The parameters in the model will learn *something.*
- Take the top off of the transformer, glue a feed forward neural network on the top of it, then tune it to some supervised task (e.g., sentiment classification, machine translation, topic modeling) using a smaller amount of data.
- This paradigm changed NLP forever.


Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ULMFiT
ELMo
GPT
Multi-lingual
MultiFiT
Transformer
Bidirectional LM
Larger model
More data
BERT
Cross-lingual
XLM
UDify
Multi-task
MT-DNN
Knowledge distillation
MT-DNN$_{KD}$
+ Generation
MASS
UniLM
Span prediction
Remove NSP
SpanBERT
Longer time
Remove NSP
More data
RoBERTa
Permutation LM
Transformer-XL
More data
XLNet
+Knowledge Graph
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
Cross-modal
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
GPT-2
Defense
Grover
Whole Word Masking
ERNIE (Baidu)
BERT-wwm
By Xiaozhi Wang & Zhengyan Zhang @THUNLP

How do we make LMs better?

Long-distance dependencies

E.g.: “I met Biff last week. I have to say, of all people I have ever met, he …”

Counting up words and word sequences doesn’t capture meaning of words.

Doesn’t actually play nicely with general purpose machine learning classifiers.

Many deep learning architectures are slow to train and run

How do we make LMs better?

Long-distance dependencies

E.g.: “I met Biff last week. I have to say, of all people I have ever met, he …”

Counting up words and word sequences doesn’t capture meaning of words.

Doesn’t actually play nicely with general purpose machine learning classifiers.

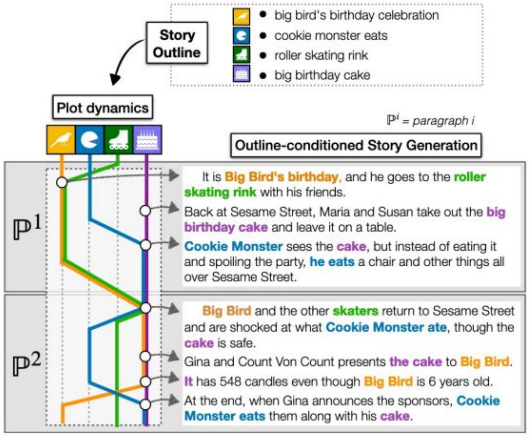
Many deep learning architectures are slow to train and run

Natural Language Generation

*Some ideas taken from Christopher Manning's slides on NLG

Uses of Generative LLMs

Creative stories



(Rashkin et al.., EMNLP 2020)

Data-to-text

Table Title: Robert Craig (American football)
Section Title: National Football League statistics
Table Description:None

		RUSHING					RECEIVING				
YEAR	TEAM	ATT	YDS	AVG	LNG	TD	NO.	YDS	AVG	LNG	TD
1983	SF	176	725	4.1	71	8	48	427	8.9	23	4
1984	SF	155	649	4.2	28	4	71	675	9.5	64	3
1985	SF	214	1050	4.9	62	9	92	1016	11	73	6
1986	SF	204	830	4.1	25	7	81	624	7.7	48	0
1987	SF	215	815	3.8	25	3	66	492	7.5	35	1
1988	SF	310	1502	4.8	46	9	76	534	7.0	22	1
1989	SF	271	1054	3.9	27	6	49	473	9.7	44	1
1990	SF	141	439	3.1	26	1	25	201	8.0	31	0
1991	RAI	162	590	3.6	15	1	17	136	8.0	20	0
1992	MIN	105	416	4.0	21	4	22	164	7.5	22	0
1993	MIN	38	119	3.1	11	1	19	169	8.9	31	1
Totals	-	1991	8189	4.1	71	56	566	4911	8.7	73	17

Craig finished his eleven NFL seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards.

(Parikh et al.., EMNLP 2020)

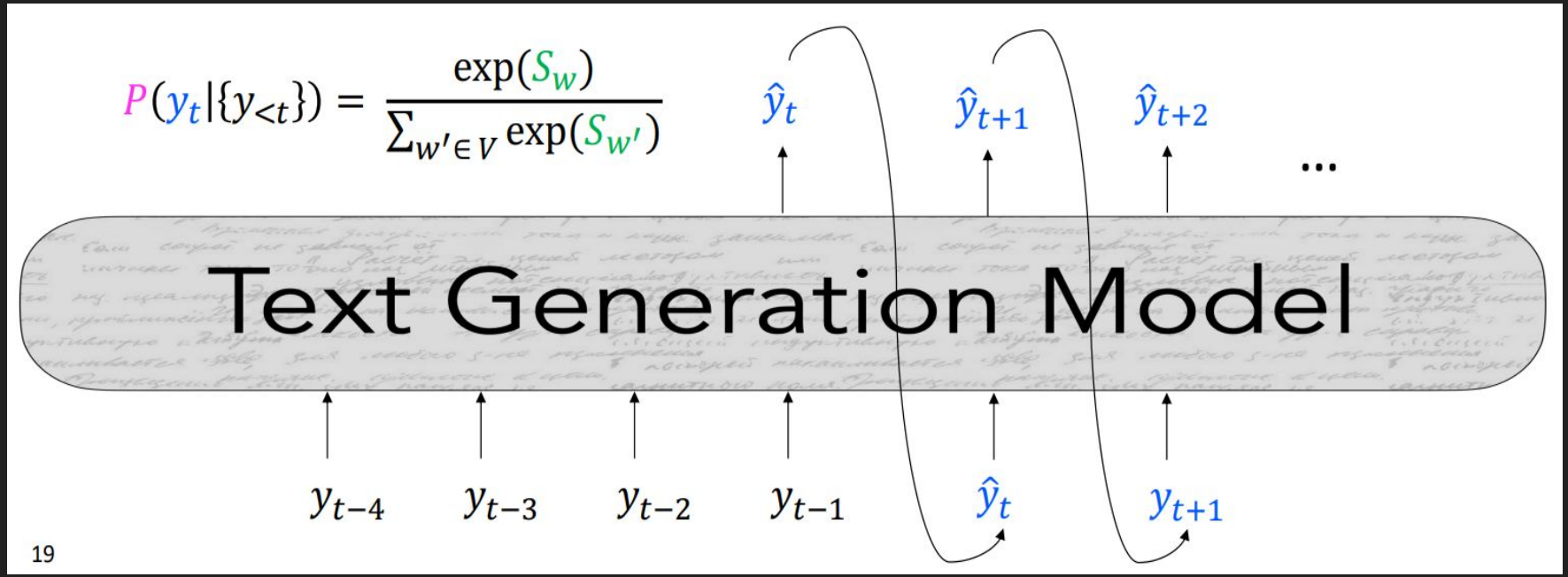
Visual description

Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

(Krause et al. CVPR 2017)

Basics of Generation

Seed text, then predict next words given a vocabulary and sequence



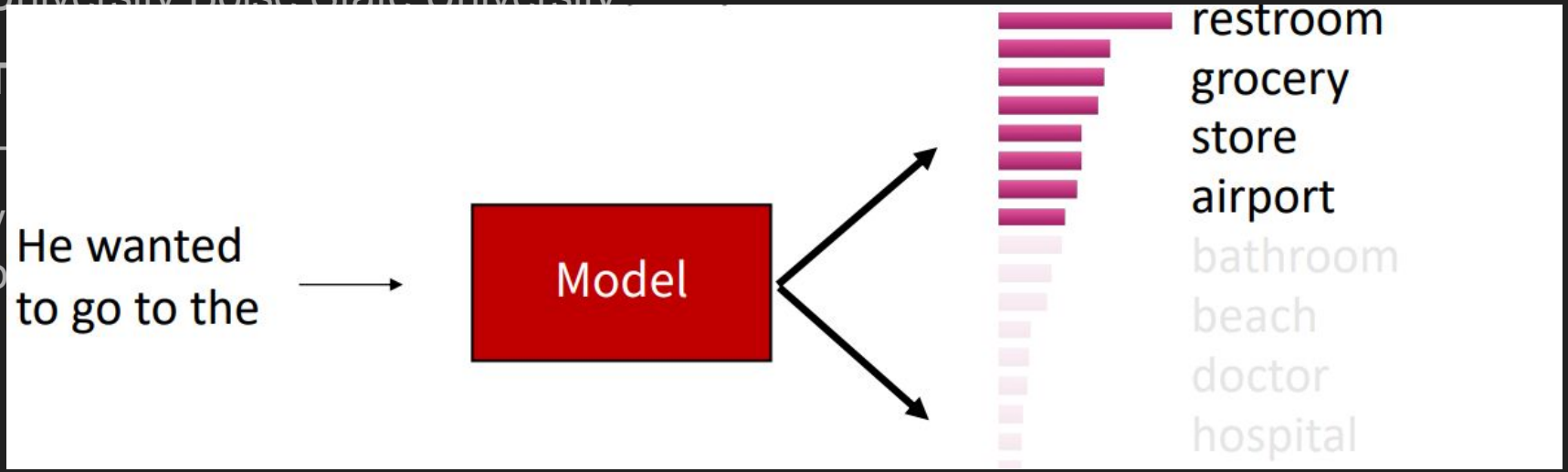
Generation: two issues

Autoregression:

Predict the next word based on input + model's ongoing output. That can steer the model to repeat outputs. E.g.: "he works at Boise State University Boise State University Boise State University . . . " .

T

T
v
o

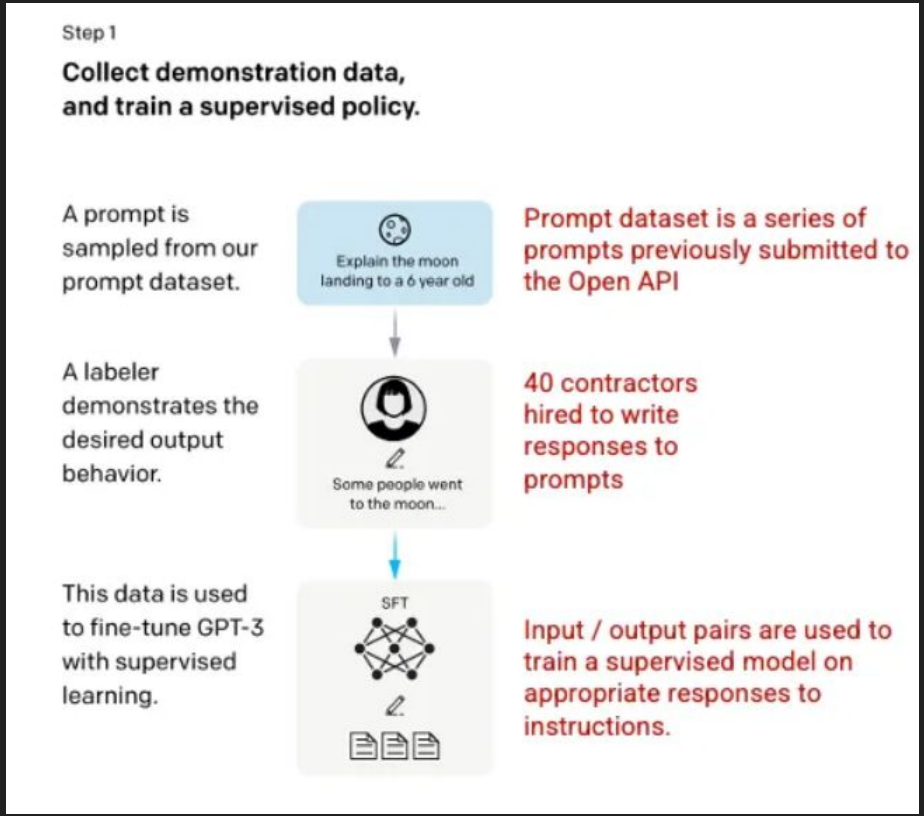


ChatGPT

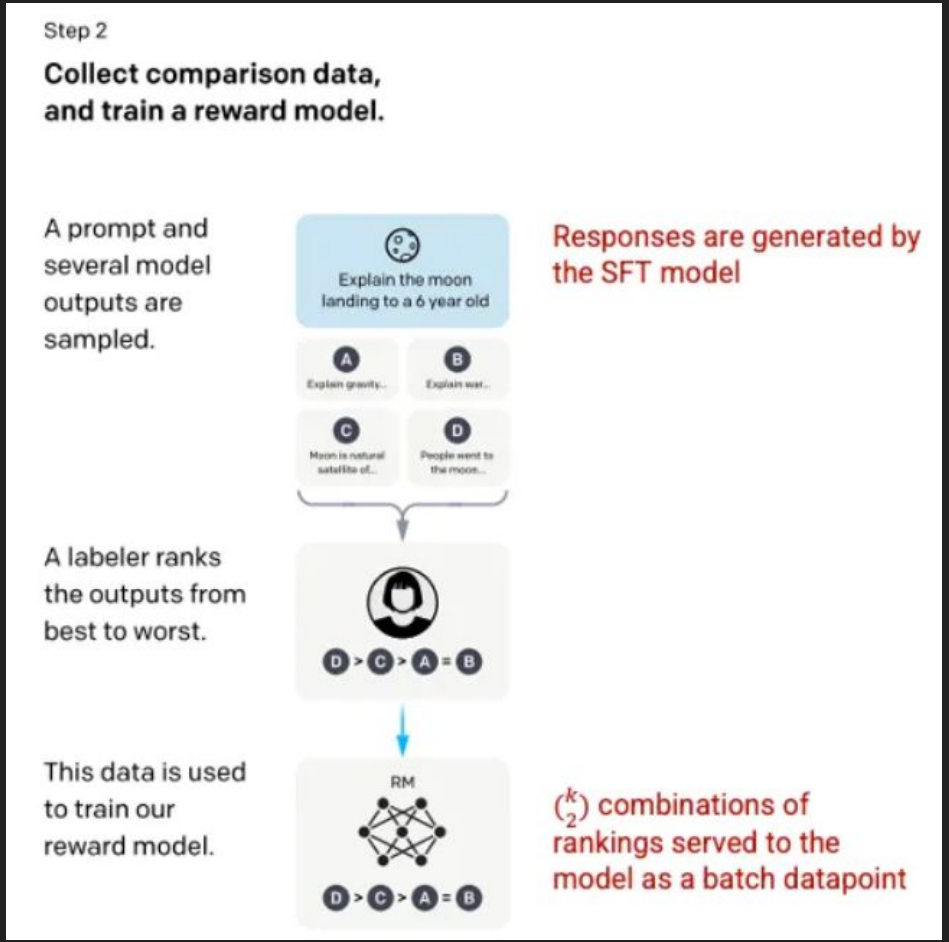
ChatGPT

- GPT = **Generative** Pre-trained Transformer
- 175 billion parameters
- ChatGPT is a spinoff of InstructGPT, which introduced a novel approach to incorporating human feedback into the training process to better align the model outputs with user intent. Reinforcement Learning from Human Feedback (RLHF) is described in depth in openAI's 2022 paper Training language models to follow instructions with human feedback and is simplified below.
- OpenAI massively scaled up this process and made the interface useful
- https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

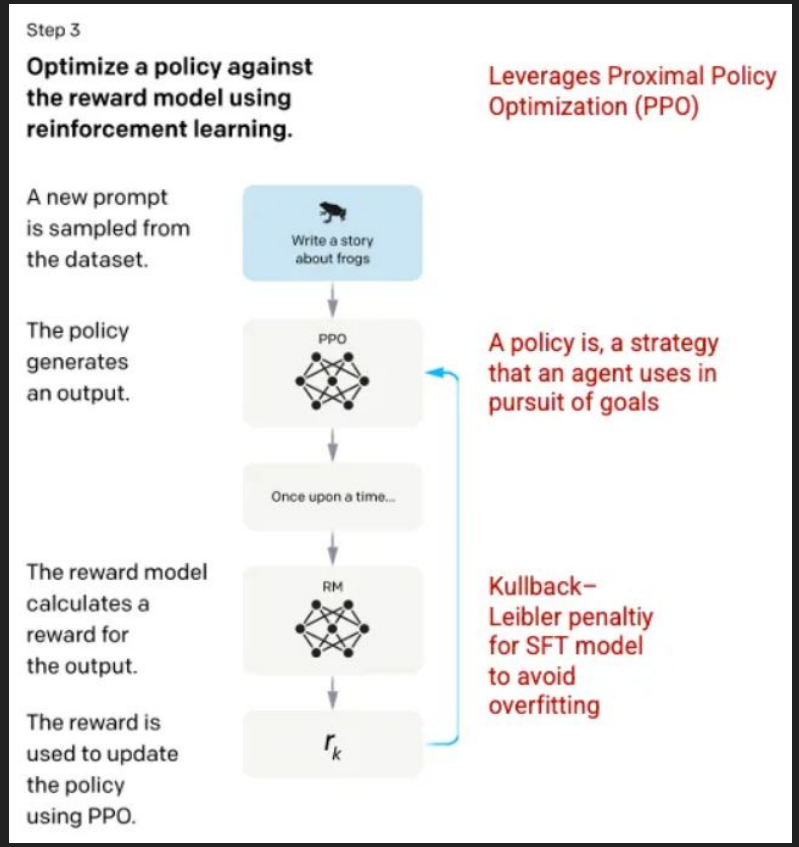
Step 1: Self-supervised fine-tuning



Step 2: Reward model



Step 3: Reinforcement Learning Model



Recap: LLMs & ChatGPT

- Better models (deep learning, attention/transformers), more (text) data, and improved meaning representations (embeddings), and improved training regimes (masked language modeling) makes language models very powerful for any task where language is automatically processed
- More parameters means more capacity for learning
- OpenAI took language models, scaled them up, and made them useful

Why Should I Care?

How people are using LLMs

- Handle mundane tasks
  - Answer emails, draft policies
  - Get ideas, find references
  - Generate website content
- LLMs are good at general sequential learning
  - Language
  - Robot actions
- Write code

Companies doing things with LLMs/AI

- Chatbots for helpdesks (Wells Fargo, Intuit)
- AI + search (Perplexity)
- Generating images / sub-images (Canva)
- Company-only access LLM (Equifax, Kount)
- Generating descriptions of items (AirBnb?, Vacasa)
- LLM cloud / hosting (AWS, Azure, Google, Huggingface)
- Small language models (various)

https://link.springer.com/article/10.1007/s42001-024-00250-1

FORBES > BUSINESS

GPT detectors are bia
writers

Weixin Liang

Open Access •

Check for

RollingStone

INNOVATIO

ars T

SIZING UP THE

Repor
Open

OpenAI coul

ASHLEY BELANGER

ChatGPT for Education: Some Cautionary Advice

Casey Kennington

7 min read · Feb 9

4

Conclusion

- LLMs have many issues and limitations, but they are always being improved (more data, more modalities, more hardware)
- A lot is happening in Idaho:
  - Companies are adopting LLMs; small LMs are more common
  - Boise State is very actively applying LMs in classroom settings, research
    - Nice workshops on how Educators can use AI
  - School districts are responding
- What's now/next:
  - text, audio, images, video
  - small language models

Thank you

BOISE STATE
UNIVERSITY