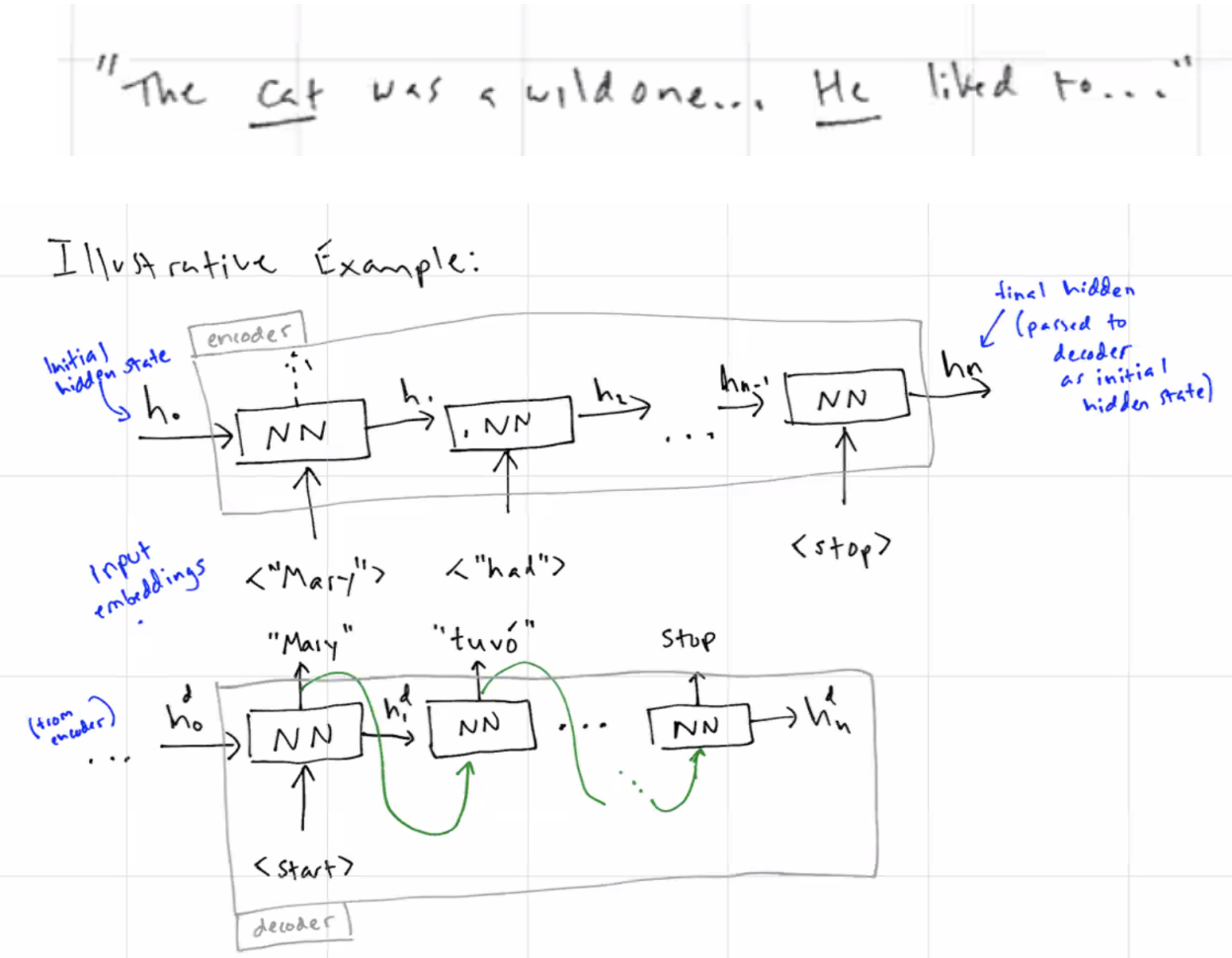
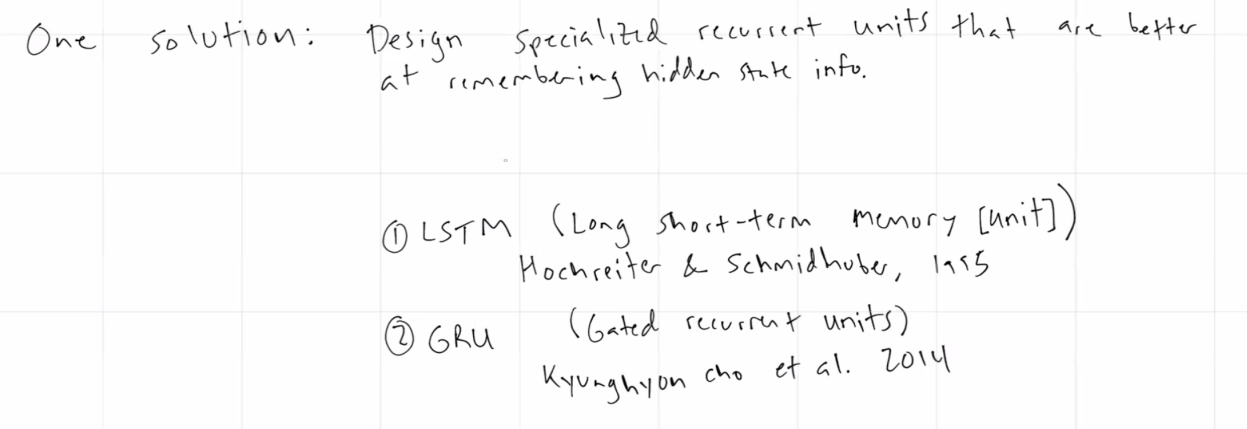
Day 12: Recurrent Neural Networks (continued)

Practical Challenges of RNNs

1. Non-parallel operations that scale with sequence length
2. Longer sequences require deeper networks
3. Fixed size hidden state to support longe-range dependencies (relationships between works that are far apart)



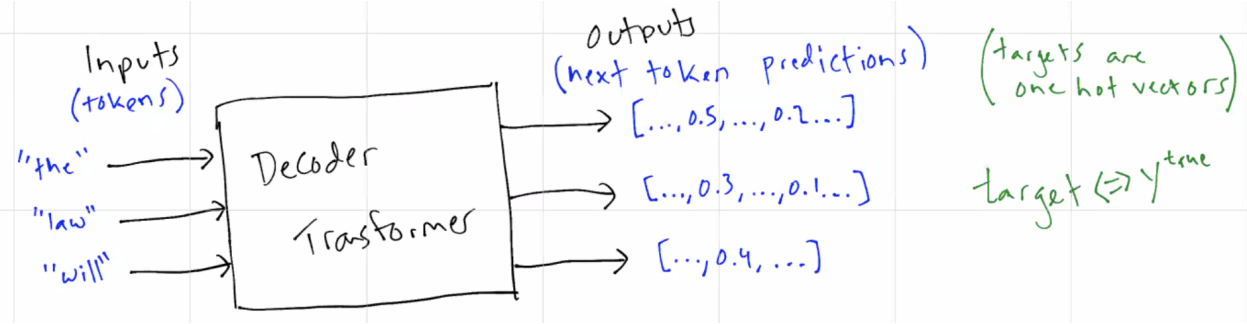
Above: recurrent NN encoder & decoder for translation - with optional teacher forcing while training decoder

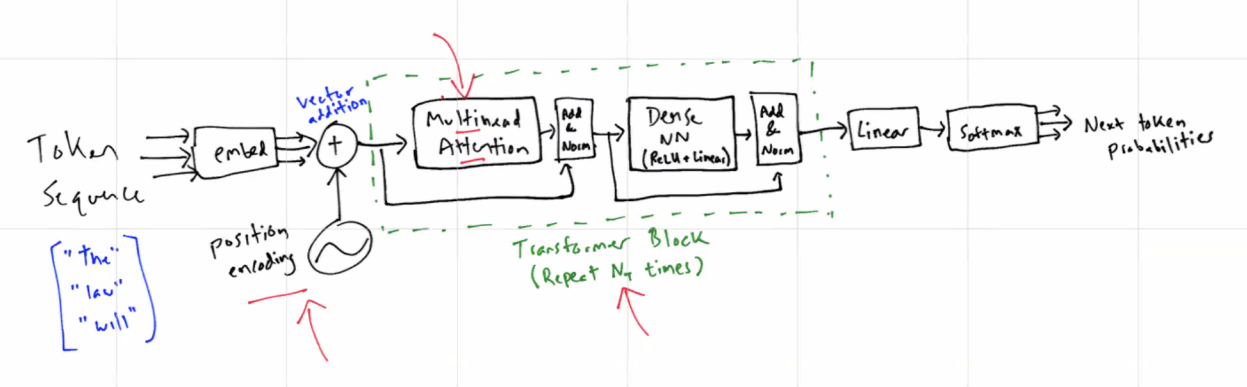


Fundamentally, difficulties remain → can one create a non-recurrent architecture for sequential data?

Anatomy of a Decoder Transformer

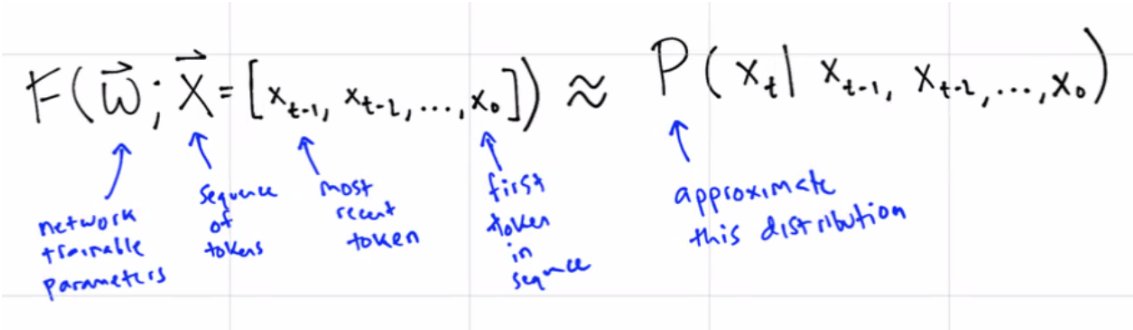
- Inputs are individual tokens fed into decoder in parallel
- Outputs are pseudo-probabilities (next token predictions) over the next token set
- Attention/Transformer block “learns” what words/phrases are most important to be considered





- Address concerns of RNNs:
  1. Fixed number non-parallelizable steps
  2. Fixed network depth
  3. No language dependencies

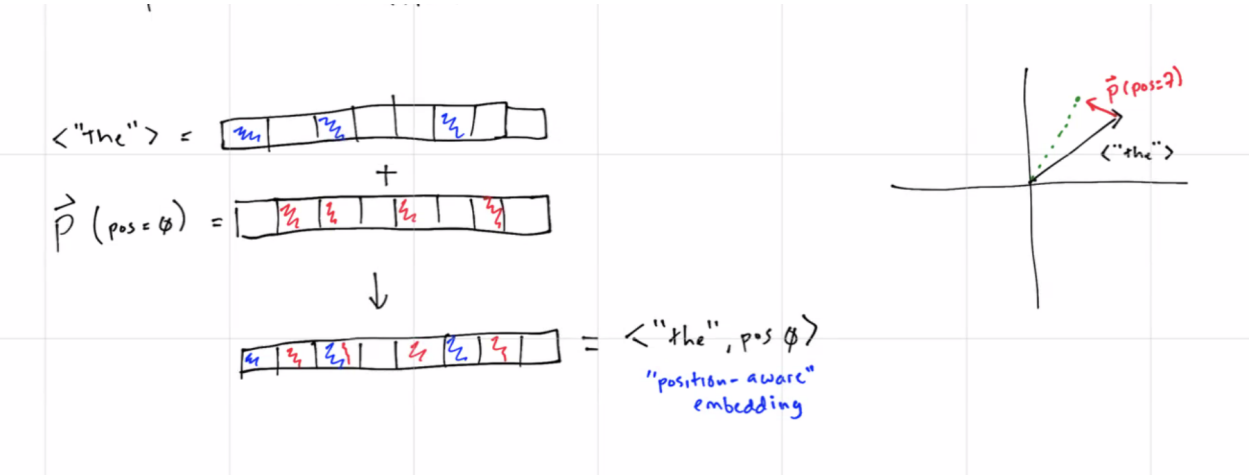
1. Decoder Transformer Objective: predict next token given part-tokens
   a. More explicitly, let network model the following probability function:



$$F(\vec{w}; \vec{x} = [x_{t-1}, x_{t-2}, ..., x_0]) \approx P(x_t | x_{t-1}, x_{t-2}, ..., x_0)$$

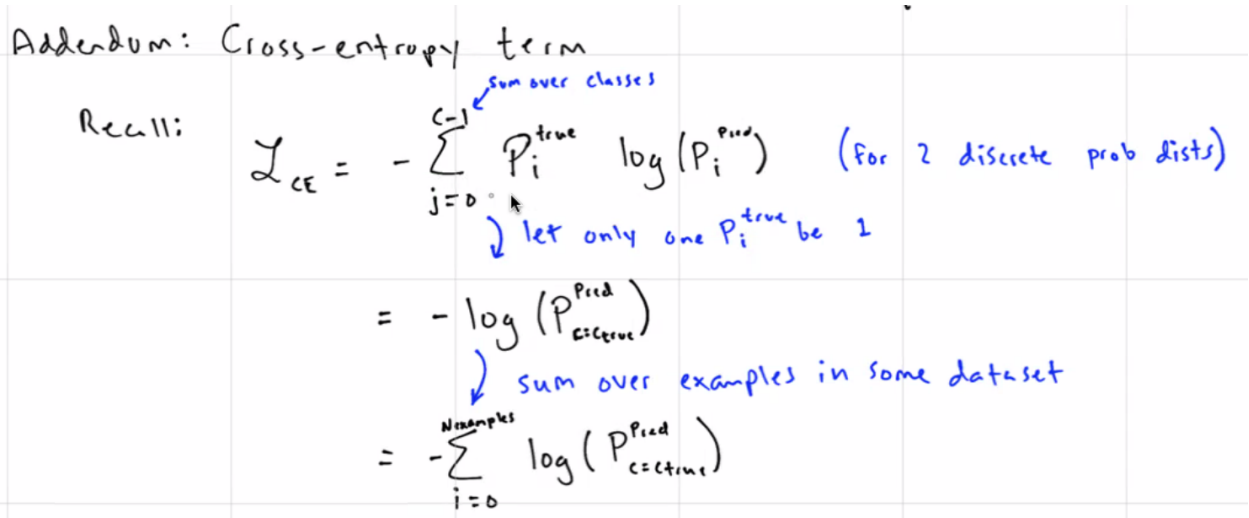
   b. We rely on a cross-entropy loss function:

$$L(X = [\vec{x}^0, \vec{x}^1, ..., \vec{x}^N]) = -\sum_i log(F(\vec{W}_1 \vec{x}^i))$$



$$\approx -\sum_i log(P(x_t^i | x_{t-1}^i, ..., x_o))$$

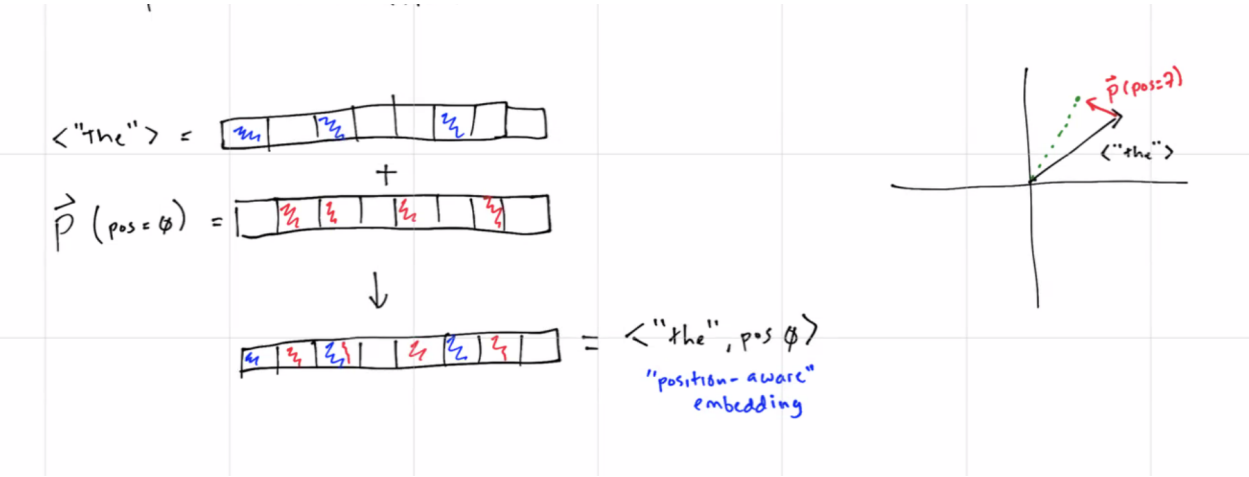
$$= \sum_i log(P_{c=c_{true}}^{pred})$$



- Outline
  1. Positional encoding
  2. Attention (”masked”, “multiheaded”, “sequential”)
  3. How can we use decoder transformer architecture outside of text generation

Positional Encoding

Concept: Add position specific vector each input embedding to impart position information

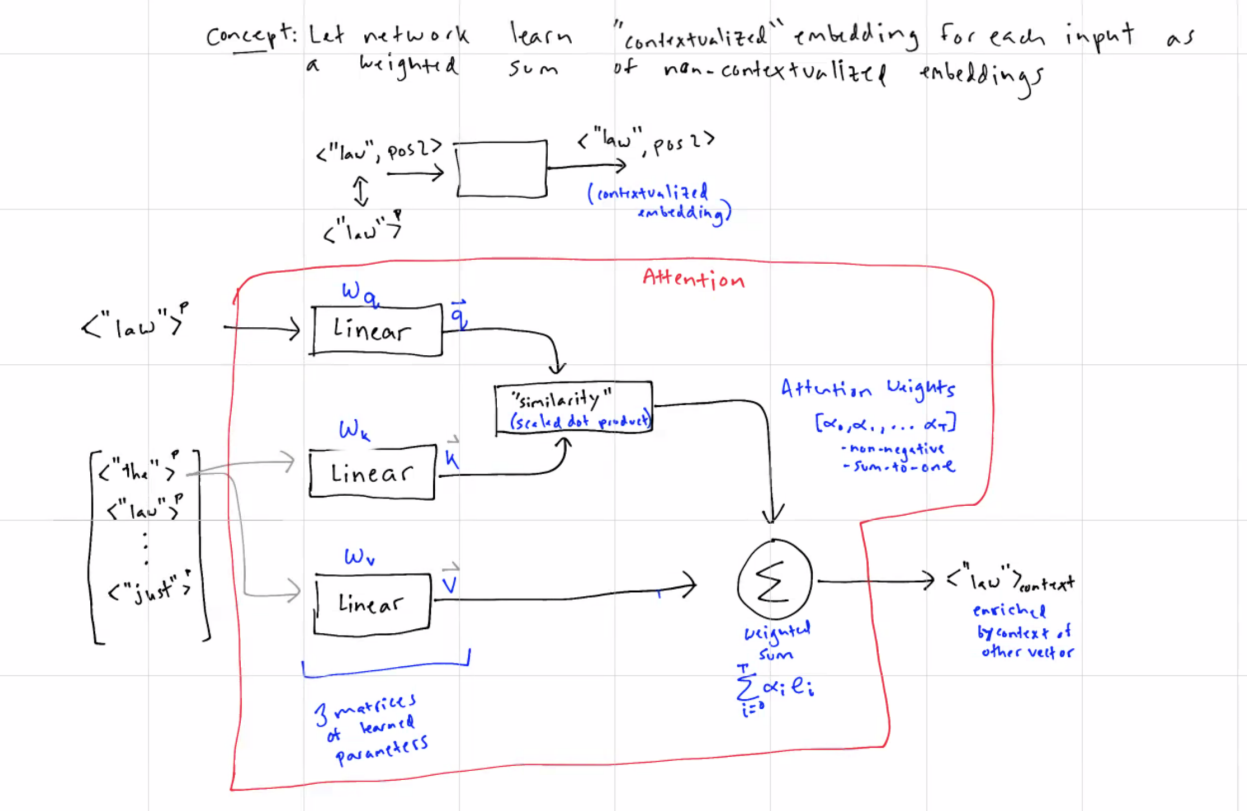


Are we ok to scatter info of index into vector space? Isn't it information inefficient?

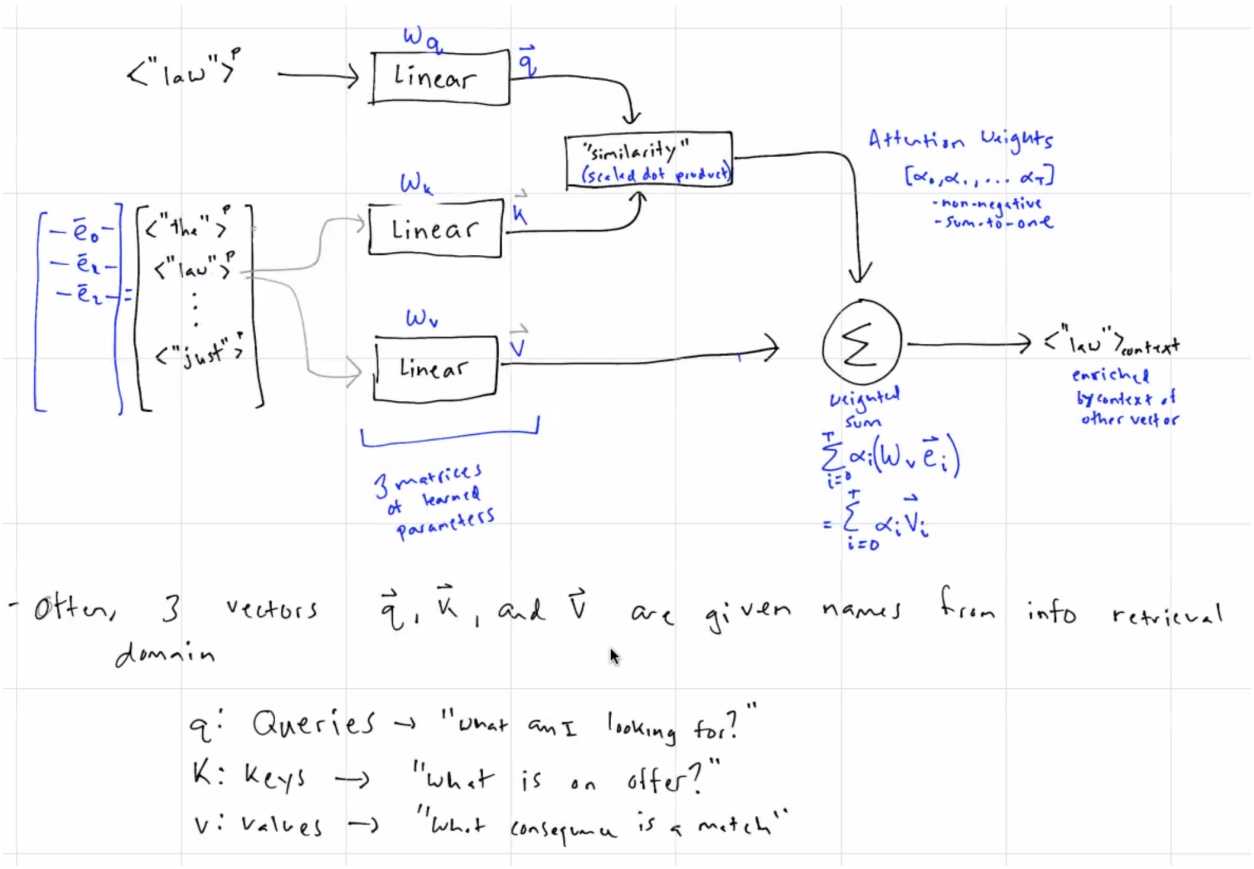
- Approaches to calculate
  1. Learn positional embedding matrix
     a. The “position-aware” embeddings are calculated as the sum off the word embeddings $\vec{e}_i$ and position vectors $\vec{p}_i$ (the latter of which is extracted from the learned function weights)
  2. Prescribe a function

Attention Mechanisms

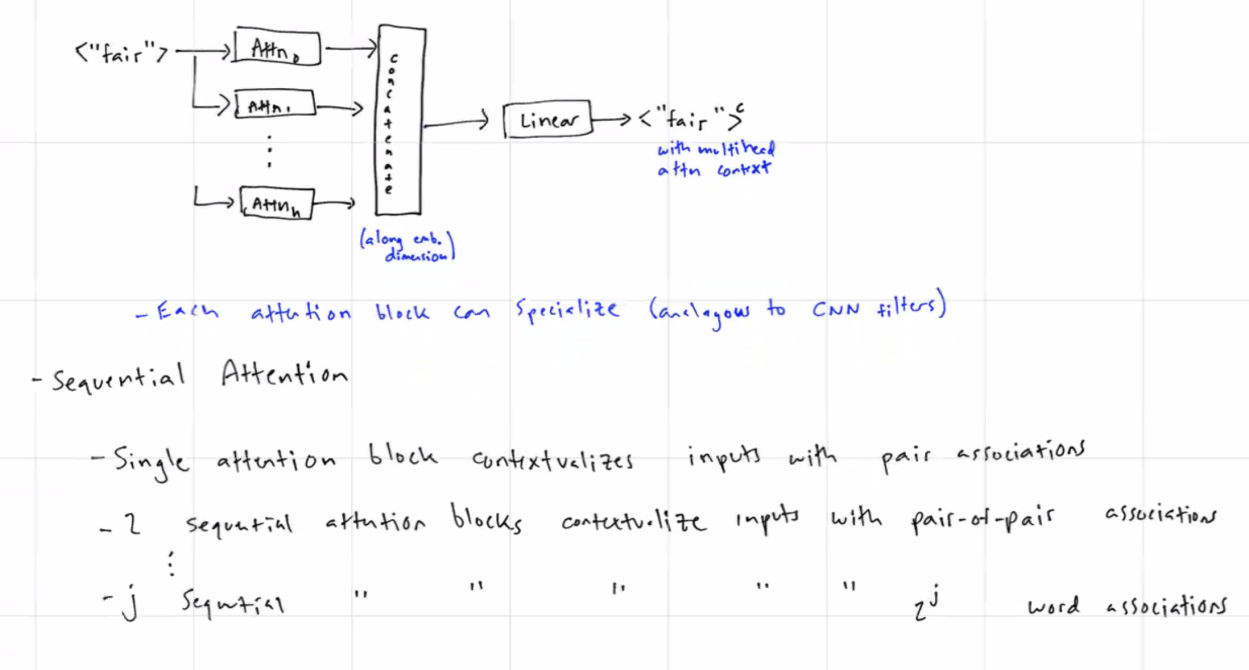
- Concept: Let network learn “contextualized” embedding for each input as a weighted sum of non-contextualized embeddings
- 3 vectors, $\vec{q}, \vec{k}, \vec{v}$ are given name from info retrieval domain
  - $\vec{q}$: Queries → “What am I looking for?”
  - $\vec{k}$: Keys → “What if ”



Circled in red above: the mechanisms within a “head” or attention block



Multihead Attention (basically multiple attention blocks [“heads”] running in parallel)

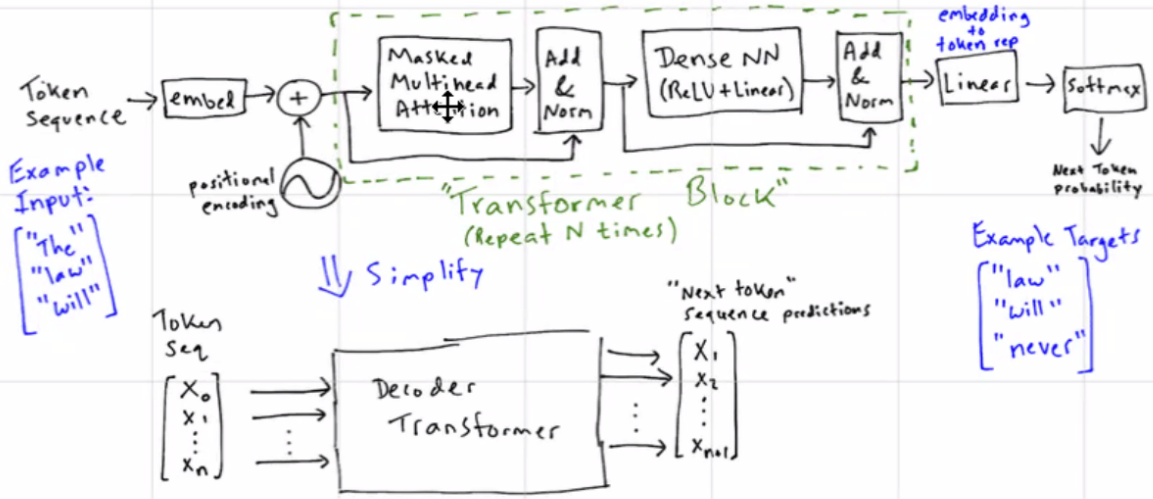


Hope that the different "heads" learn different contexts that may be helpful when used together

Example: "Jeff is a musician. Jeff went to the park, where he _____"
One head might make "he" pay attention to the earlier "Jeff" and another head might make "he" pay attention to "musician," where combined, they might help determine that the next word is "played".

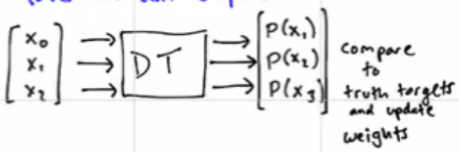
"Predictions for position i can only depend on positions less than i."

Training and Generation Procedures

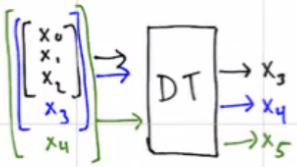


Training Procedure

- Compare full output sequence to targets and update weights to reduce cross-entropy loss (over full text corpus)



Generation Procedure (weights static)



(and so on)

How can we use the model for tasks beyond text generation?

A: We can prompt model for what we want using language

① News article generation

"Title: Flying Saucers visit White House
Subtitle: Onlookers shocked by alien life forms
... [let model complete]"

② Conversational Question Answering

"Mary had a little lamb, her fleece was filthy
Question: what color was Mary's lamb?
Answer: ... [let model complete]

③ Grammar Correction

"poor English grammar: I eated purple berries
good English grammar: I ate purple berries
poor English grammar: The patient was died
good English grammar: ... [model completes]

Examples of other use cases

Network Sizes and Training Costs:

(* Data from "Generative Pre-trained Transformer" wikipedia page)

	date	# param	dataset (size)	FLOPs	effective* train time
GPT-1	June, 2018	~100 million	BooksCorpus (4.5 GB)	1.7 e19	1 month
GPT-2	Feb, 2019	1.3 billion	WebText (40 GB text)	1.5 e21	~100 month
GPT-3	May, 2020	175 billion	Common Crawl webtext English wikipedia Books1 & Books2 (100s of GB)	3.1e23	~10000 month (~1000y)
GPT-3.5	March, 2022	175 billion	Undisclosed	Undisclosed	Undisclosed
GPT-4	March, 2023	undisclosed	Undisclosed	Undisclosed	undisclosed

* assume compute FLOPS used on GPT-1, i.e. $\frac{1.7 e19 \text{ FLOP}}{1 \text{ month}}$ = ~6 peta-FLOPS

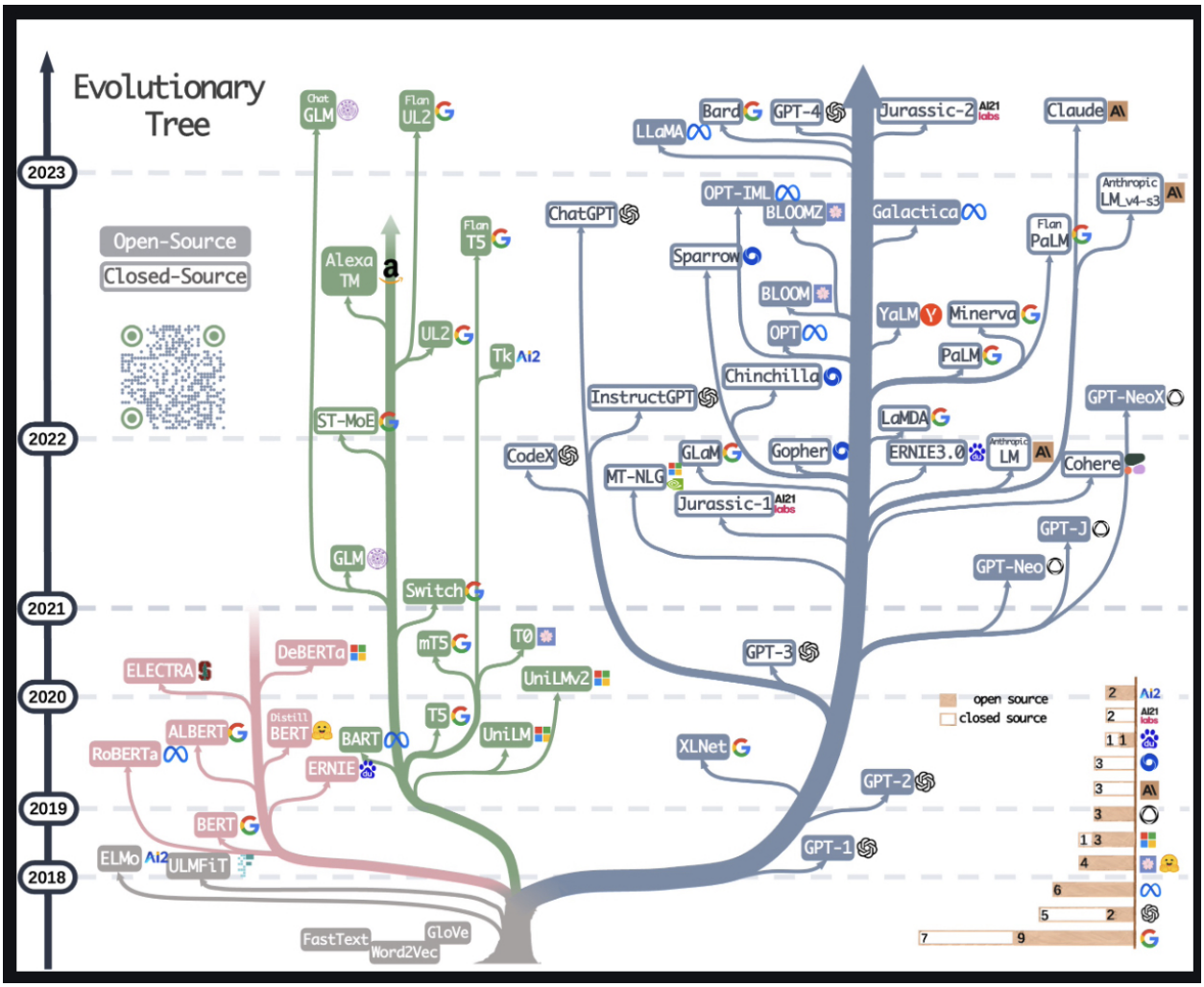
Use of Large Language Models

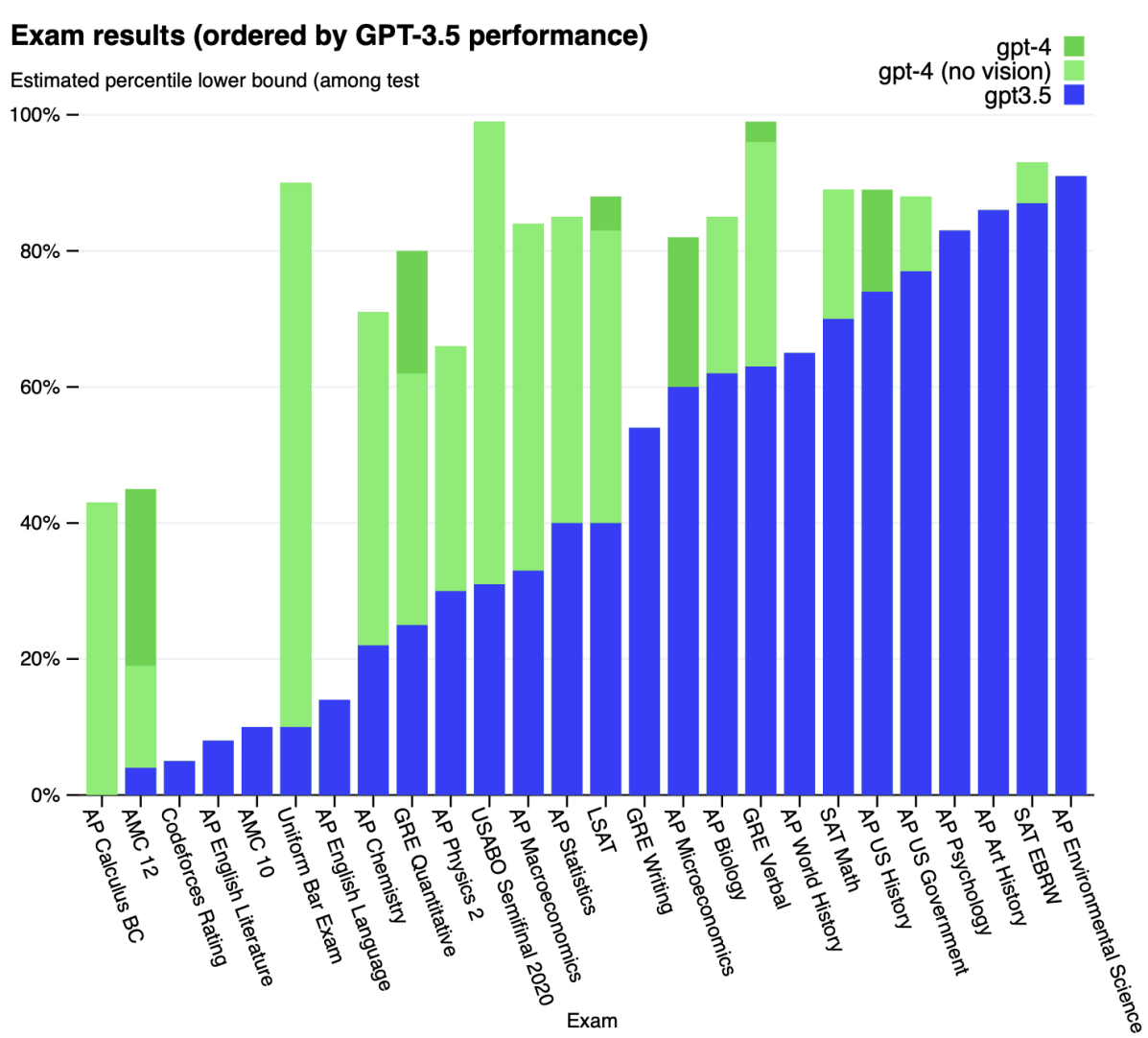
Wednesday, July 26, 2023 3:52 PM

1. baseline: what are the current strengths and weaknesses of LLMs?

2. list of ≥3 areas that are likely to be / being disrupted by LLMs

3. how should people in your area use/not use LLMs? are there general principles you can extract from your area to apply ~~more~~ broadly?


Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Chat GLM
Flan UL2
Alexa TM
Flan T5
UL2
Tk
ST-MoE
GLM
Switch
mT5
T0
UniLMv2
T5
UniLM
DeBERTa
ELECTRA
ALBERT
Distill BERT
BART
RoBERTa
ERNIE
BERT
ELMo
ULMFiT
FastText
GloVe
Word2Vec
Bard
GPT-4
Jurassic-2
Claude
LLaMA
OPT-IML
BLOOMZ
Galactica
Anthropic LM_v4-s3
ChatGPT
Sparrow
Flan PaLM
BLOOM
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GPT-Neo
GPT-3
XLNet
GPT-2
GPT-1
open source
closed source


Exam results (ordered by GPT-3.5 performance)
Estimated percentile lower bound (among test
gpt-4
gpt-4 (no vision)
gpt3.5
100%
80%
60%
40%
20%
0%
AP Calculus BC
AMC 12
Codeforces Rating
AP English Literature
AMC 10
Uniform Bar Exam
AP English Language
AP Chemistry
GRE Quantitative
AP Physics 2
USABO Semifinal 2020
AP Macroeconomics
AP Statistics
LSAT
GRE Writing
AP Microeconomics
AP Biology
GRE Verbal
AP World History
SAT Math
AP US History
AP US Government
AP Psychology
AP Art History
SAT EBRW
AP Environmental Science
Exam