



Reviewing the Foundations

CSCI 601-771 (NLP: Advances in Self-Supervised Models)

Language Modeling: Definitions and History

The

The cat

The cat sat

The cat sat on

The cat sat on __?__

The cat sat on the mat.

$P(\text{mat} \mid \text{The cat sat on the})$

next word

context or prefix

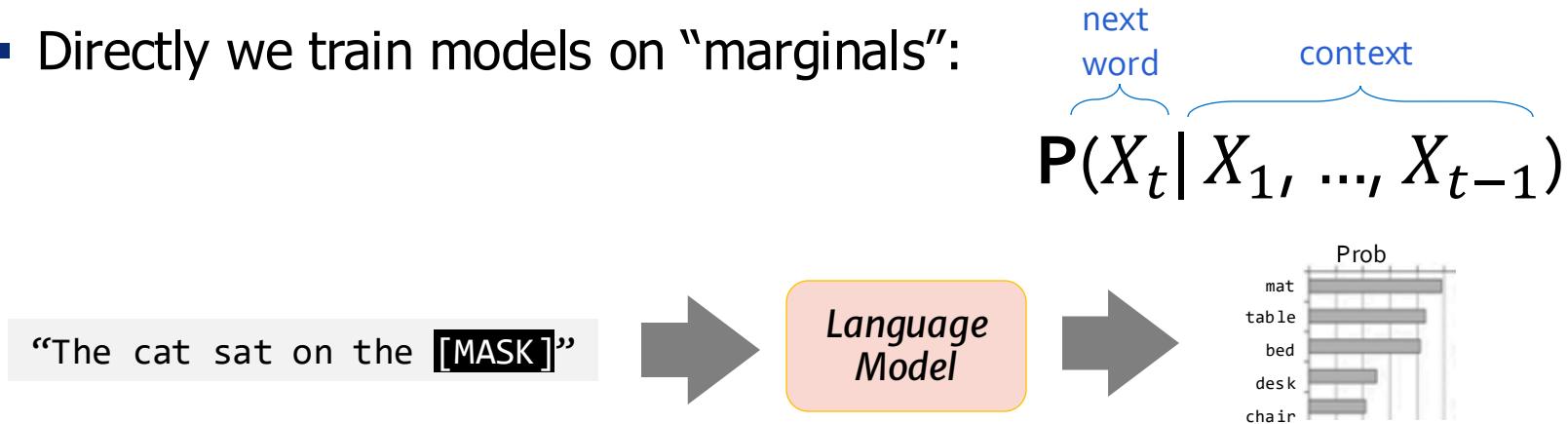
Probability of Upcoming Word

$$P(X_t | X_1, \dots, X_{t-1})$$



LMs as a Marginal Distribution

- Directly we train models on “marginals”:



LMs as Implicit Joint Distribution over Language

- While language modeling involves learning the marginals, we are implicitly learning the full/joint distribution of language.
 - Remember the chain rule:

$$P(X_1, \dots, X_t) = P(X_1) \prod_{i=1}^t P(X_i | X_1, X_2, \dots, X_{i-1})$$

- **Language Modeling** \triangleq learning prob distribution over language sequence.

Doing Things with Language Model

- What is the probability of

“I like Johns Hopkins University”

“like Hopkins I University Johns”

Doing Things with Language Model

- What is the probability of

“I like Johns Hopkins University”

“like Hopkins I University Johns”

- LMs assign a probability to every sentence (or any string of words).

$P(\text{"I like Johns Hopkins University EOS"}) = 10^{-5}$

$P(\text{"like Hopkins I University Johns EOS"}) = 10^{-15}$

Doing Things with Language Model (2)

- We can rank sentences.
- While LMs show “typicality”, this may be a proxy indicator to other properties:
 - Grammaticality, fluency, factuality, etc.

$$P(X_t | X_1, \dots, X_{t-1})$$

next word context



$P("I \ like \ Johns \ Hopkins \ University. \ EOS") > P("I \ like \ John \ Hopkins \ University \ EOS")$

$P("I \ like \ Johns \ Hopkins \ University. \ EOS") > P("University. \ I \ Johns \ EOS \ Hopkins \ like")$

$P("JHU \ is \ located \ in \ Baltimore. \ EOS") > P("JHU \ is \ located \ in \ Virginia. \ EOS")$

Doing Things with Language Model (3)

- Can also generate strings!

$$P(X_t | X_1, \dots, X_{t-1})$$

next word context



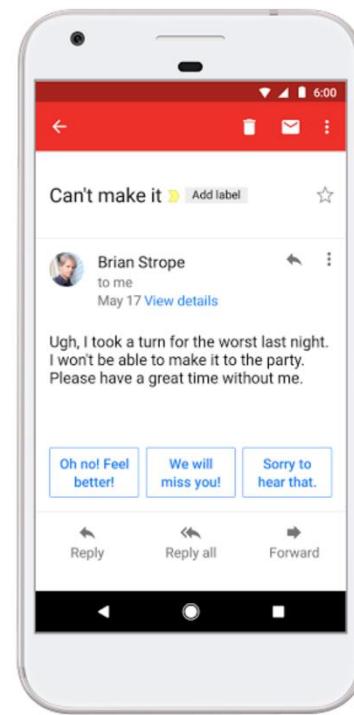
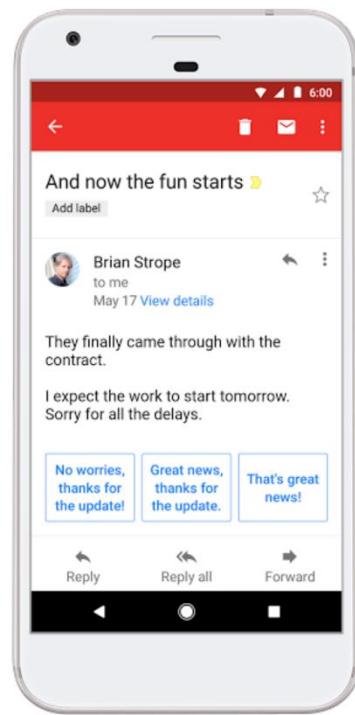
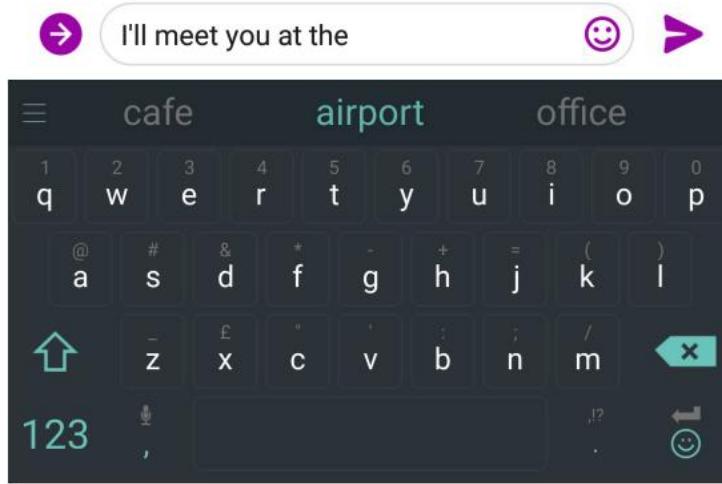
- Let's say we start "*Johns Hopkins is*"
- Using this prompt as an initial condition, recursively sample from an LM:

1. Sample from $P(X | "Johns Hopkins is") \rightarrow \text{"located"}$
2. Sample from $P(X | "Johns Hopkins is located") \rightarrow \text{"at"}$
3. Sample from $P(X | "Johns Hopkins is located at") \rightarrow \text{"the"}$
4. Sample from $P(X | "Johns Hopkins is located at the") \rightarrow \text{"state"}$
5. Sample from $P(X | "Johns Hopkins is located at the state") \rightarrow \text{"of"}$
6. Sample from $P(X | "Johns Hopkins is located at the state of") \rightarrow \text{"Maryland"}$
7. Sample from $P(X | "Johns Hopkins is located at the state of Maryland") \rightarrow \text{"EOS"}$

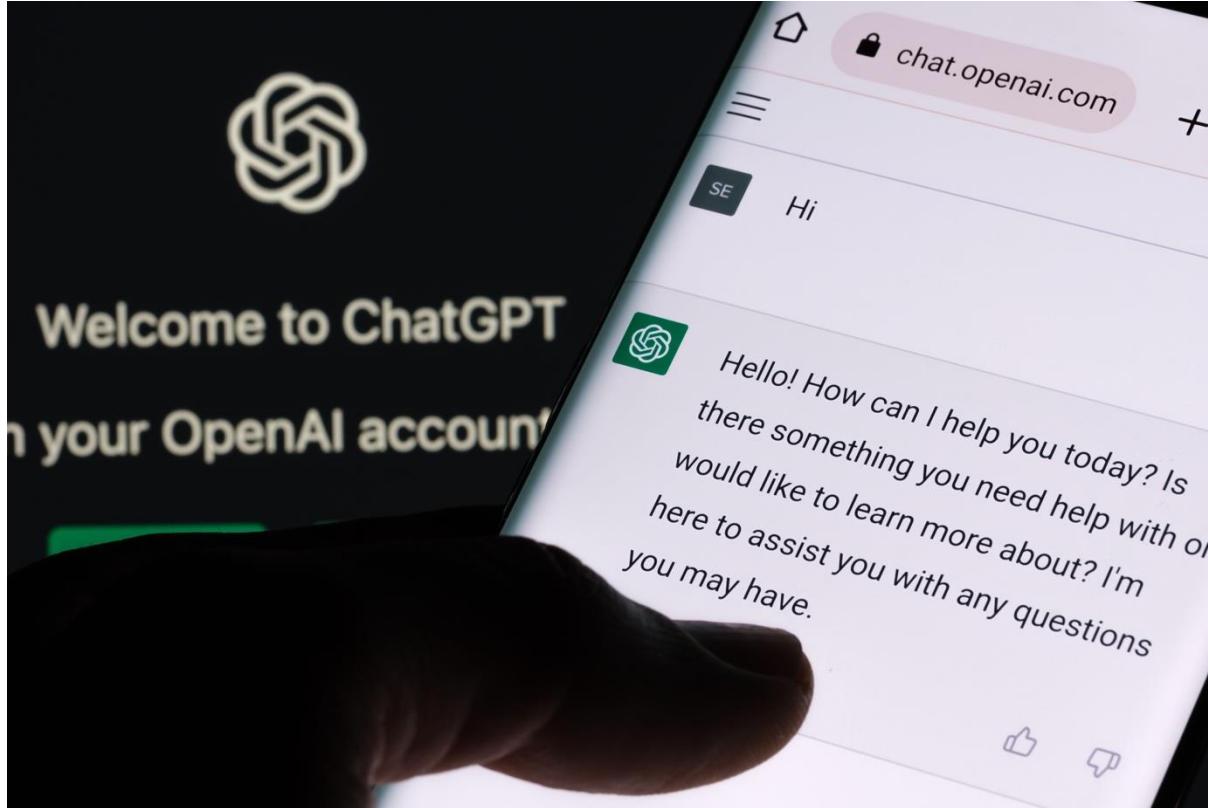
Why Care About Language Modeling?

- Language Modeling is a **subcomponent superset** of many tasks:
 - Summarization
 - Machine translation
 - Spelling correction
 - Dialogue etc.
- Language Modeling is an effective proxy for **language understanding**.
 - **Effective ability to predict forthcoming words** requires **understanding of context/prefix**.

You use Language Models every day!



You use Language Models every day!



Summary

- **Language modeling:** building probabilistic distribution over language.
- An accurate distribution of language enables us to solve many important tasks that involve language communication.
- **The remaining question:** how do you actually estimate this distribution?

Language Modeling with Counting

LMs as a Marginal Distribution

- Now the question is, how to estimate this distribution.

$$\mathbf{P}(X_t | X_1, \dots, X_{t-1})$$

How do we estimate these probabilities?

Let's just count!

$$P(\text{mat} | \text{the cat sat on the}) \approx \frac{\text{count}(\text{"the cat sat on the mat"})}{\text{count}(\text{"the cat sat on the"})}$$

Count how often
“**the cat sat on the mat**”
has appeared in the world (internet)!

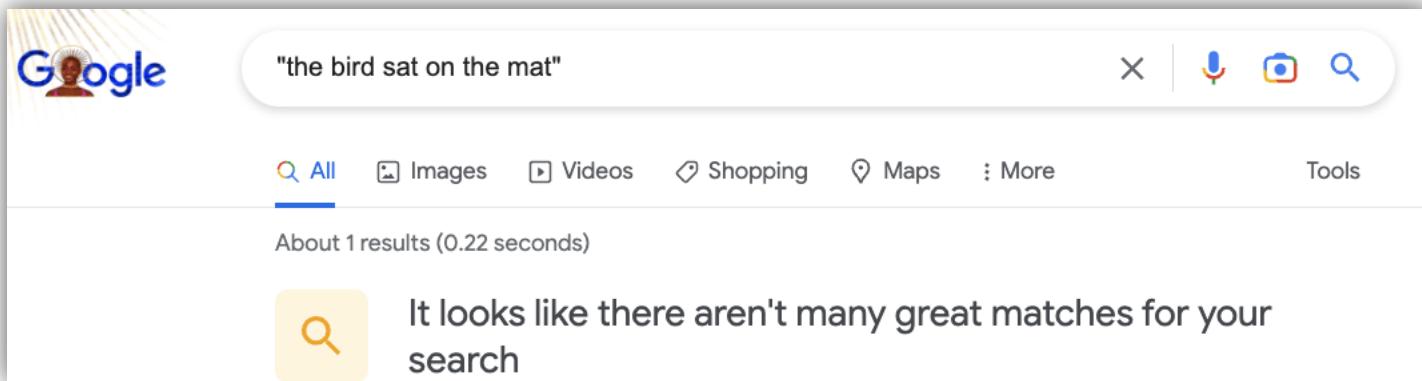
Divide that by, the count of
“**the cat sat on the**”
in the world (internet)!

$$\mathbf{P}(X_t | X_1, \dots, X_{t-1})$$

How do we estimate these probabilities?

Let's just count!

$$P(\text{mat} | \text{the cat sat on the}) \approx \frac{\text{count}(\text{"the cat sat on the mat"})}{\text{count}(\text{"the cat sat on the"})}$$



A screenshot of a Google search results page. The search bar contains the query "the bird sat on the mat". Below the search bar, there are navigation links for All, Images, Videos, Shopping, Maps, More, and Tools. A message indicates "About 1 results (0.22 seconds)". At the bottom, a yellow search icon is followed by the text: "It looks like there aren't many great matches for your search".

$$\mathbf{P}(X_t | X_1, \dots, X_{t-1})$$

How do we estimate these probabilities?

Let's just count!

$$P(\text{mat} | \text{the cat sat on the}) \approx \frac{\text{count}(\text{"the cat sat on the mat"})}{\text{count}(\text{"the cat sat on the"})}$$

Challenge: Increasing n makes **sparsity problems** worse.

Typically, we can't have n bigger than 5.

Some partial solutions (e.g., smoothing and backoffs)
though still an open problem.

Language Models: A History

- Shannon (1950): The redundancy and predictability (entropy) of English.



Prediction and Entropy of Printed English

By C. E. SHANNON

(Manuscript Received Sept. 15, 1950)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.



$$P(X_t | X_1, \dots, X_{t-1})$$

Shannon (1950) built an approximate language model with word co-occurrences.

Markov assumptions: every node in a Bayesian network is **conditionally independent** of its non-descendants, **given its parents**.

1st order approximation:

$$P(\text{mat} | \text{the cat sat on the}) \approx P(\text{mat} | \text{the})$$

1 element



$$P(X_t | X_1, \dots, X_{t-1})$$

Andrey Markov

Shannon (1950) built an approximate language model with word co-occurrences.

Markov assumptions: every node in a Bayesian network is **conditionally independent** of its non-descendants, **given its parents**.

2nd order approximation:

$$P(\text{mat} | \text{the cat sat on the}) \approx P(\text{mat} | \text{on the})$$

2 elements



$$P(X_t | X_1, \dots, X_{t-1})$$

Andrey Markov

Shannon (1950) built an approximate language model with word co-occurrences.

Markov assumptions: every node in a Bayesian network is **conditionally independent** of its non-descendants, **given its parents**.

3rd order approximation:

$$P(\text{mat} | \text{the cat sat on the}) \approx P(\text{mat} | \text{sat on the})$$

3 elements



$$P(X_t | X_1, \dots, X_{t-1})$$

Andrey Markov

Shannon (1950) built an approximate language model with word co-occurrences.

Then, we can use counts of approximate conditional probability.
Using the 3rd order approximation, we can:

$$P(\text{mat} | \text{the cat sat on the}) \approx P(\text{mat} | \text{sat on the}) = \frac{\text{count("sat on the mat")}}{\text{count("on the mat")}}$$

Understanding Sparsity: A Thought Experiment

- How common are zero-probabilities? 😐
- **Example:** Shakespeare as a text corpus
 - The size vocab used by Shakespeare: $|V|=29,066$
 - Shakespeare produced: $\sim 300,000$ bigrams
 - Out of $|V|^2 = 844$ million possible bigrams
 - (some of them don't make sense, but ok!)
- So, **99.96%** of the possible bigrams are **never seen** (hence, have zero entries for bigram counts).

N-gram Language Models

- **Terminology:** *n*-gram is a chunk of *n* consecutive words:
 - unigrams: “cat”, “mat”, “sat”, ...
 - bigrams: “the cat”, “cat sat”, “sat on”, ...
 - trigrams: “the cat sat”, “cat sat on”, “sat on the”, ...
 - four-grams: “the cat sat on”, “cat sat on the”, “sat on the mat”, ...
- *n*-gram language model:

$$P(X_t | X_1, \dots, X_{t-1}) \approx P(X_t | \underbrace{X_{t-n+1}, \dots, X_{t-1}}_{n-1 \text{ elements}})$$

Generation from N-Gram Models

- You can build a simple **trigram** Language Model over a 1.7 million words corpus in a few seconds on your laptop*

Generation from N-Gram Models

- You can build a simple trigram Language Model over a 1.7 million words corpus in a few seconds on your laptop*

today the

get probability
distribution



company	0.153
bank	0.153
price	0.077
italian	0.039
emirate	0.039
...	

Sparsity problem: not
much granularity in the
probability distribution

Otherwise, seems reasonable!

Generation from N-Gram Models

- Now we can sample from this mode:

today the

get probability
distribution



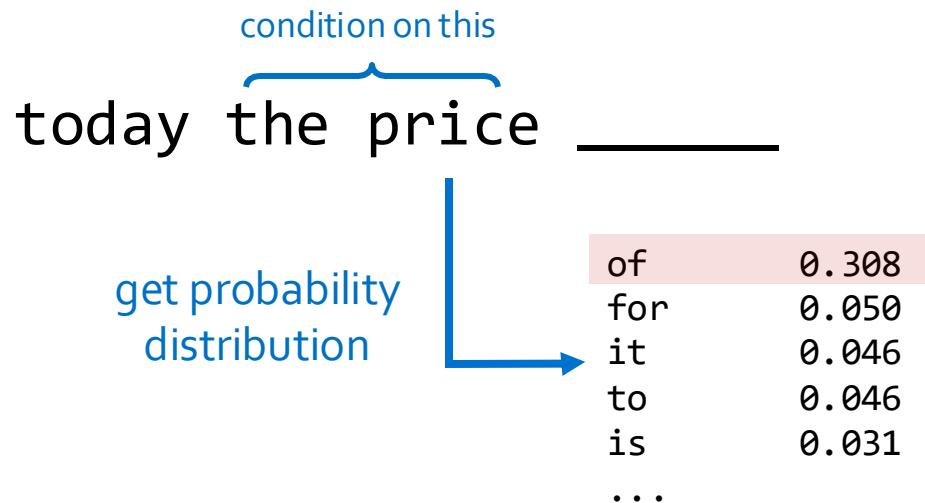
company	0.153
bank	0.153
price	0.077
italian	0.039
emirate	0.039
...	

Sparsity problem: not
much granularity in the
probability distribution

Otherwise, seems reasonable!

Generation from N-Gram Models

- Now we can sample from this mode:

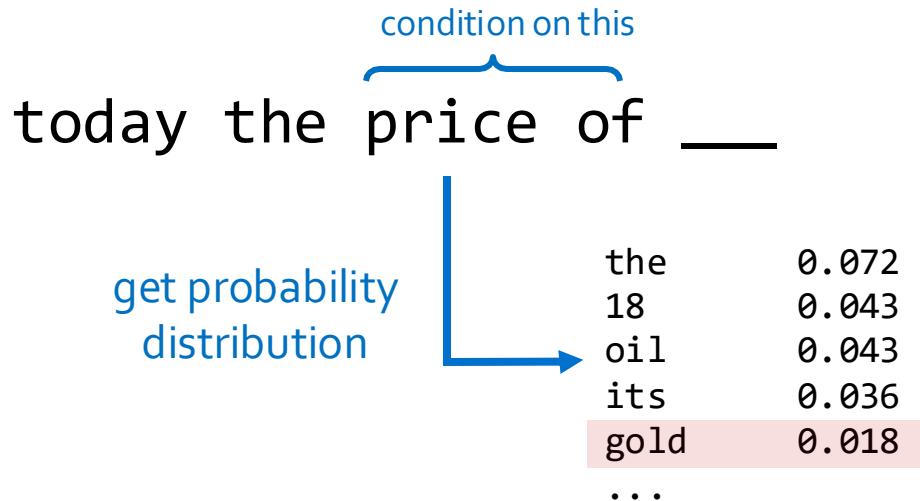


Sparsity problem: not much granularity in the probability distribution

Otherwise, seems reasonable!

Generation from N-Gram Models

- Now we can sample from this mode:



Sparsity problem: not much granularity in the probability distribution

Otherwise, seems reasonable!

N-Gram Models in Practice

- Now we can sample from this mode:

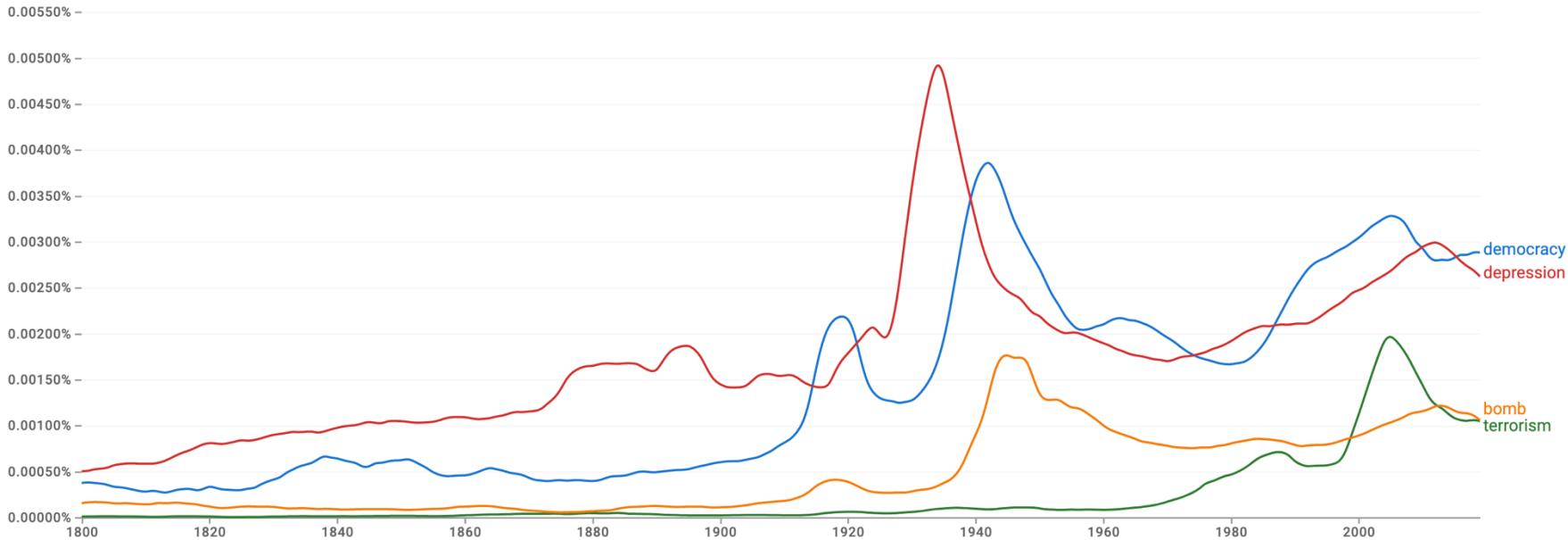
today the price of gold per ton , while production of shoe lasts and shoe industry , the bank intervened just after it considered and rejected an imf demand to rebuild depleted european stocks , sept 30 end primary 76 cts a share .

Surprisingly grammatical!

But quite incoherent! To improve coherence, one may consider increasing larger than 3-grams, but that would worsen the sparsity problem!

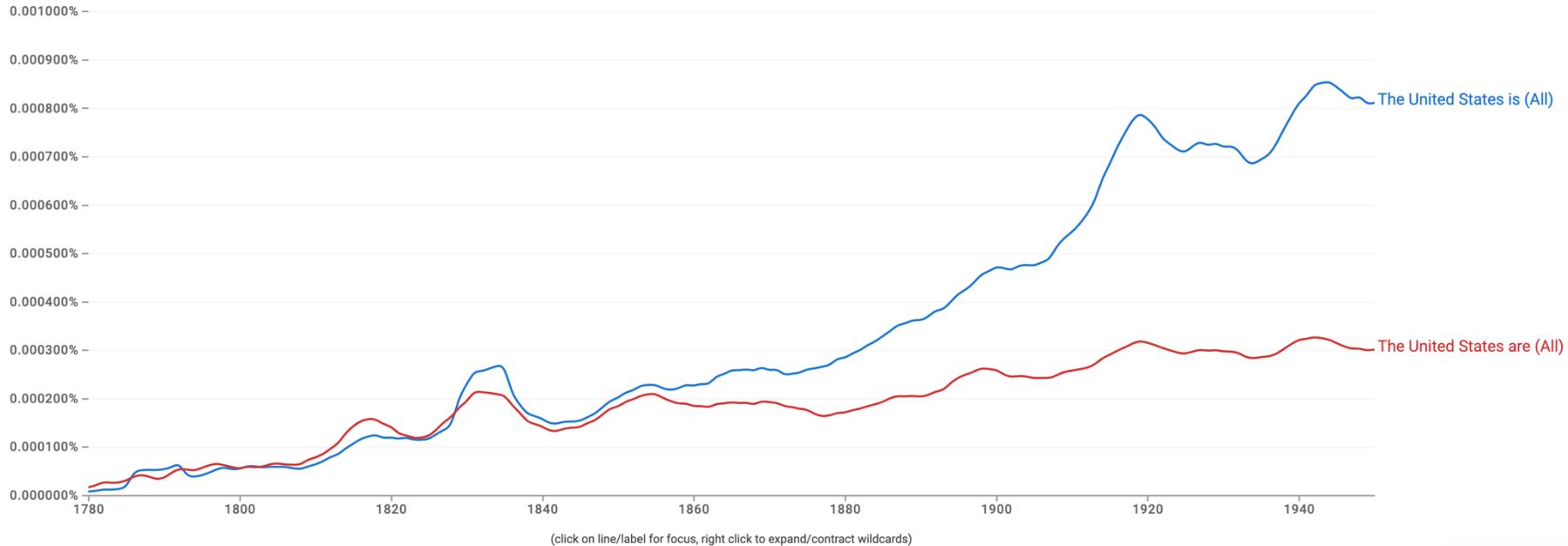
Pre-Computed N-Grams

Google Books Ngram Viewer



Pre-Computed N-Grams

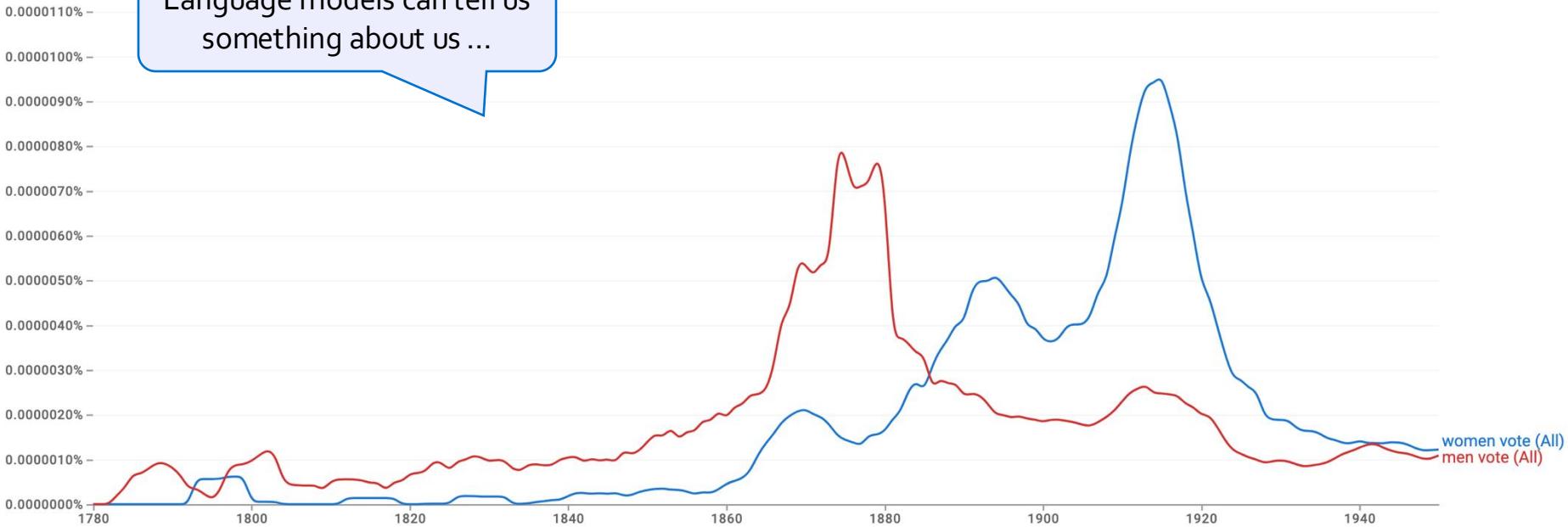
Google Books Ngram Viewer



Pre-Computed N-Grams

Google Books Ngram Viewer

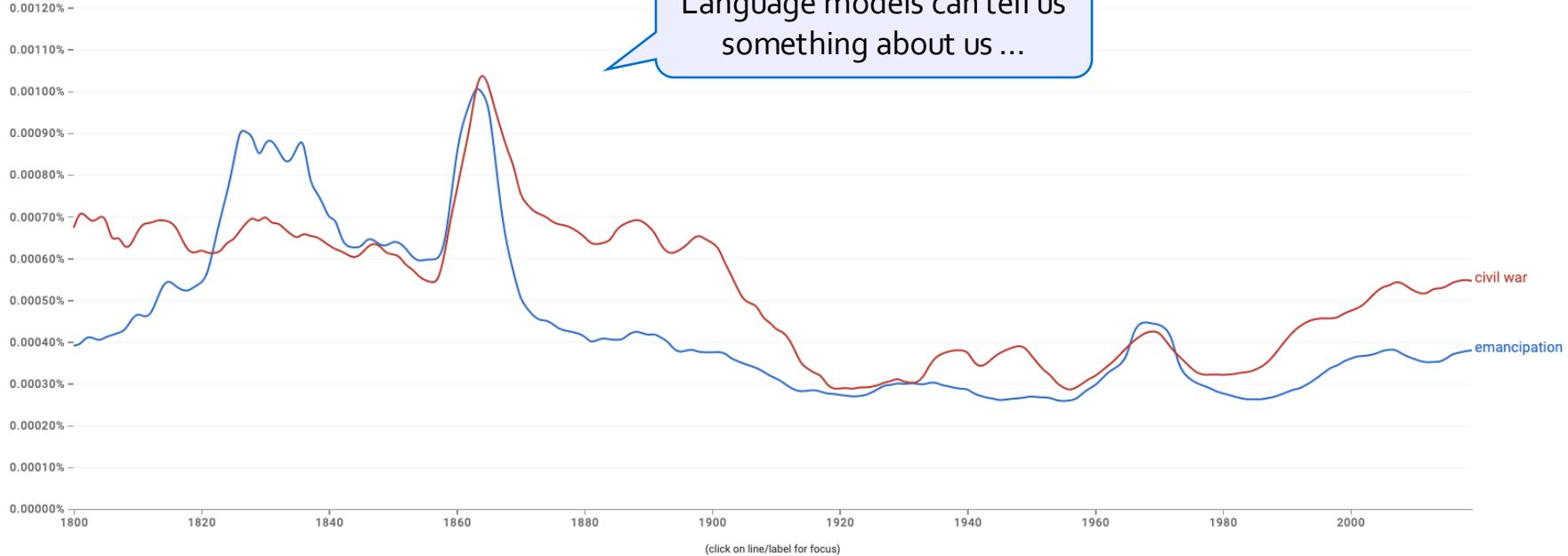
Language models can tell us something about us ...



Pre-Computed N-Grams

Google Books Ngram Viewer

Language models can tell us something about us ...



Limits of N-Grams LMs: Long-range Dependencies

- In general, count-based LMs are insufficient models of language because language has **long-distance dependencies**:

“**The computer** which I had just put into the machine room on the fifth floor **crashed.**”

N-Gram Language Models, A Historical Highlight

"Every time I fire a linguist, the performance of the speech recognizer goes up"!!



Fred Jelinek
(1932-2010)

- Probabilistic n-gram models of text generation [Jelinek+ 1980's, ...]
 - Applications: Speech Recognition, Machine Translation

532

PROCEEDINGS OF THE IEEE, VOL. 64, NO. 4, APRIL 1976

Continuous Speech Recognition by Statistical Methods

FREDERICK JELINEK, FELLOW, IEEE

Abstract—Statistical methods useful in automatic recognition of continuous speech are described. They concern modeling of a speaker and of an acoustic processor, extraction of the models' statistical parameters, and hypothesis search procedures and likelihood computations of linguistic decoding. Experimental results are presented that indicate the power of the methods.

utterance models used will incorporate more grammatical features, and statistics will have been grafted onto grammatical models. Most methods presented here concern modeling of the speaker's and acoustic processor's performance and should, therefore, be universally useful.

Automatic recognition of continuous (English) speech is an

Summary

- Learning a language model ~ learning **conditional probabilities** over language.
- One approach to estimating these probabilities: **counting word co-occurrences**.
- Challenges:
 - Word co-occurrences become **rare** for long sequences. (the sparsity issue)
 - But language understanding requires **long-range** dependencies.
- We need a better alternative! 😱
- **Next:** Measuring quality of language models.

How Good are Language Models?

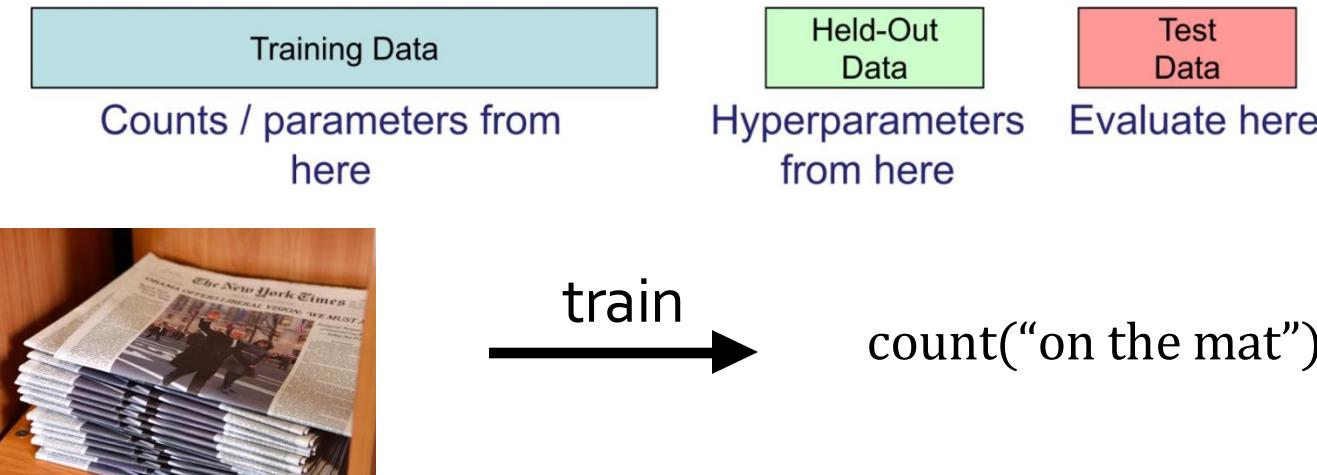
Evaluating Language Models

- Does our language model prefer good sentences to bad ones?
 - Assign higher probability to “real” or “frequently observed” sentences
 - Than “ungrammatical” or “rarely observed” sentences?
- We test the model’s performance on data we haven’t seen.

Evaluating Language Models

Setup:

- Train it on a suitable training documents.
- Evaluate their predictions on different, unseen documents.
- An evaluation metric tells us how well our model does on the test set.



Evaluating Language Models: Example

Setup:

- Train it on a suitable training documents.
- Evaluate their predictions on different, unseen documents.
- An evaluation metric tells us how well our model does on the test set.

Example: I use a bunch of New York Times articles to build a bigram probability table



A good language model
should assign a high
probability to held-out text!

train

count("on the mat")

eval



Now I'm going to evaluate the probability of some heldout data using our bigram table

Be Careful About Data Leakage!

Advice from a grandpa 🧑:

- Don't allow test sentences to leak into training set.
- Otherwise, you will assign it an artificially high probability (==cheating).

Example: I use a bunch of New York Times articles to build a bigram probability table



A good language model
should assign a high
probability to held-out text!

train →

count("on the mat")

eval →



Now I'm going to evaluate the probability of some heldout data using our bigram table

Evaluating Language Models: Intrinsic vs Extrinsic

- **Intrinsic:** measure how good we are at modeling language
- **Extrinsic:** build a new language model, use it for some task (MT, ASR, etc.)

Example: I use a bunch of New York Times articles to build a bigram probability table



train →



Now I'm going to evaluate the probability of some heldout data using our bigram table

eval →



Evaluation Metric for Language Modeling: Perplexity

- **Perplexity** is the inverse probability of the test set, normalized by the number of words:

$$\text{ppl}(w_1, \dots, w_n) = \mathbf{P}(w_1, w_2, \dots, w_n)^{-\frac{1}{n}}$$

- A measure of **predictive quality** of a language model.
- **Minimizing** perplexity is the same as **maximizing** probability

Evaluation Metric for Language Modeling: Perplexity

- **Perplexity** is the inverse probability of the test set, normalized by the number of words:

$$\text{ppl}(w_1, \dots, w_n) = \mathbf{P}(w_1, w_2, \dots, w_n)^{-\frac{1}{n}}$$

Evaluation Metric for Language Modeling: Perplexity

- **Perplexity** is the inverse probability of the test set, normalized by the number of words:

$$\begin{aligned} \text{ppl}(w_1, \dots, w_n) &= \mathbf{P}(w_1, w_2, \dots, w_n)^{-\frac{1}{n}} \\ &= \sqrt[n]{\frac{1}{\mathbf{P}(w_1, w_2, \dots, w_n)}} = \sqrt[n]{\prod_{i=1}^n \frac{1}{\mathbf{P}(w_i | w_{<i})}} \quad \text{chain rule} \\ &= 2^H, \text{ where} \end{aligned}$$

$$H = -\frac{1}{n} \sum_{i=1}^n \log_2 \mathbf{P}(w_i | w_1, \dots, w_{i-1})$$

Evaluation Metric for Language Modeling: Perplexity

- In practice, we prefer to use **log**-probabilities (also known as “logits”)
- We can rewrite perplexity formula in terms of log-probs:

$$\text{ppl}(w_1, \dots, w_n) = 2^H, \text{ where } H = -\frac{1}{n} \sum_{i=1}^n \log_2 \mathbf{P}(w_i | w_1, \dots, w_{i-1})$$

Recap: Definition of **cross-entropy** between two distributions:

$$H(p, q) = - \sum_{x \in \mathcal{X}} p(x) \log q(x)$$

Can be interpreted as **cross-entropy** between LM prob and language prob. **Why?**

Evaluation Metric for Language Modeling: Perplexity

- In practice, we prefer to use **log**-probabilities (also known as “logits”)
- We can rewrite perplexity formula in terms of log-probs:

$$\text{ppl}(w_1, \dots, w_n) = 2^H, \text{ where } H = -\frac{1}{n} \sum_{i=1}^n \log_2 \mathbf{P}(w_i | w_1, \dots, w_{i-1})$$

- **Perplexity** for n-grams:
 - Unigrams: $H = -\frac{1}{n} \sum_{i=1}^n \log_2 \mathbf{P}(w_i)$
 - Bigrams: $H = -\frac{1}{n} \sum_{i=1}^n \log_2 \mathbf{P}(w_i | w_{i-1})$
 - Trigrams: $H = -\frac{1}{n} \sum_{i=1}^n \log_2 \mathbf{P}(w_i | w_{i-2}, w_{i-1})$
 - ...

Intuition-building Quizzes (1)

- In practice, we prefer to use **log**-probabilities (also known as “logits”)
- We can rewrite perplexity formula in terms of log-probs:

$$\text{ppl}(w_1, \dots, w_n) = 2^H, \text{ where } H = -\frac{1}{n} \sum_{i=1}^n \log_2 \mathbf{P}(w_i | w_1, \dots, w_{i-1})$$

- **Quiz:** let's suppose we have a sentence w_1, \dots, w_n and it's fixed. Our model will correctly guess each word with probability 1/5. What is perplexity of our model?

$$H = -\frac{1}{n} \left[\log_2 \left(\frac{1}{5} \right) + \dots + \log_2 \left(\frac{1}{5} \right) \right] = -\log \left(\frac{1}{5} \right) \Rightarrow \text{ppl}(D) = 5$$

Intuition: the model is indecisive among 5 choices.

Intuition-building Quizzes (2)

- In practice, we prefer to use **log**-probabilities (also known as “logits”)
- We can rewrite perplexity formula in terms of log-probs:

$$\text{ppl}(w_1, \dots, w_n) = 2^H, \text{ where } H = -\frac{1}{n} \sum_{i=1}^n \log_2 \mathbf{P}(w_i | w_1, \dots, w_{i-1})$$

- **Quiz:** let's we evaluate an **exact** (!! model of language, i.e., our model always knows what exact word should follow a given context. What is the perplexity of this model?

$$\forall w \in V: \mathbf{P}(w_i | w_{1:i-1}) = 1 \Rightarrow \text{ppl}(D) = 2^{-\frac{1}{2} n \log_2 1} = 1$$

Intuition: the model is indecisive among 1 (the right!) choice!

Intuition-building Quizzes (3)

- In practice, we prefer to use **log**-probabilities (also known as “logits”)
- We can rewrite perplexity formula in terms of log-probs:

$$\text{ppl}(w_1, \dots, w_n) = 2^H, \text{ where } H = -\frac{1}{n} \sum_{i=1}^n \log_2 \mathbf{P}(w_i | w_1, \dots, w_{i-1})$$

- **Quiz:** let's evaluate a **confused** (!!) model of language, i.e., our model has no idea what word should follow each context—it always chooses a uniformly random word. What is the perplexity of this model?

$$\forall w \in V: \mathbf{P}(w | w_{1:i-1}) = \frac{1}{|V|} \Rightarrow \text{ppl}(D) = 2^{-\frac{1}{n} n \log_2 \frac{1}{|V|}} = |V|$$

Intuition: the model is indecisive among all the vocabulary terms.

Perplexity: Summary

$$\text{ppl}(w_1, \dots, w_n) = 2^H, \text{ where } H = -\frac{1}{n} \sum_{i=1}^n \log_2 \mathbf{P}(w_i | w_1, \dots, w_{i-1})$$

- Perplexity is a measure of model's uncertainty about next word (aka "average branching factor").
 - The larger the number of vocabulary, the more options there to choose from.
 - (the choice of atomic units of language impacts PPL — more on this later)
- Perplexity ranges between 1 and $|V|$.
- We prefer LMs with lower perplexity.

Lower perplexity == Better Model

- Training on 38 million words, test 1.5 million words, Wall Street Journal

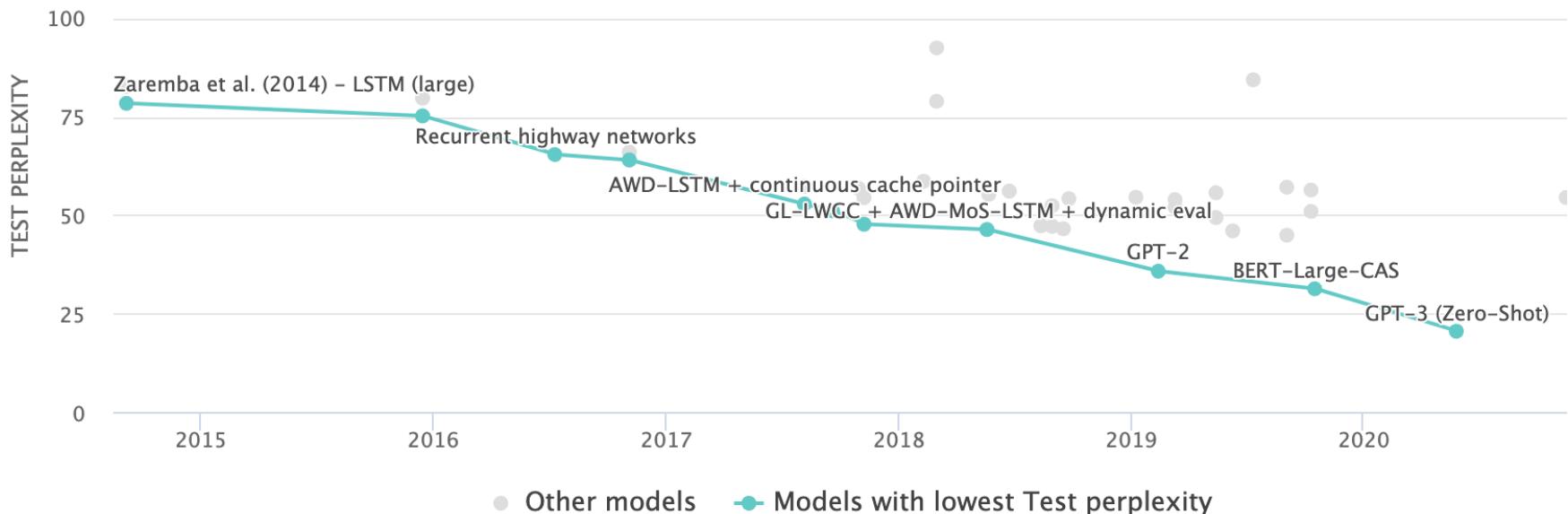
N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

Lower is better

Note these evaluations are done on data that was not used for “counting.” (no cheating!!)

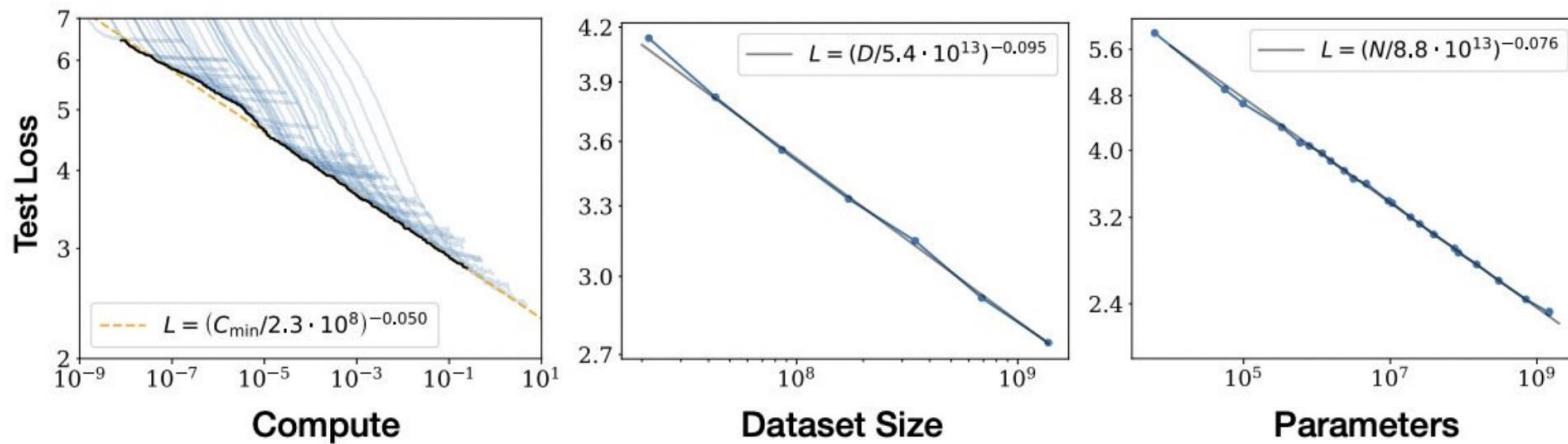
Lower perplexity == Better Model

The PPL of modern language models have consistently been going down.



Lower perplexity == Better Model

The PPL of modern language models have consistently been going down.



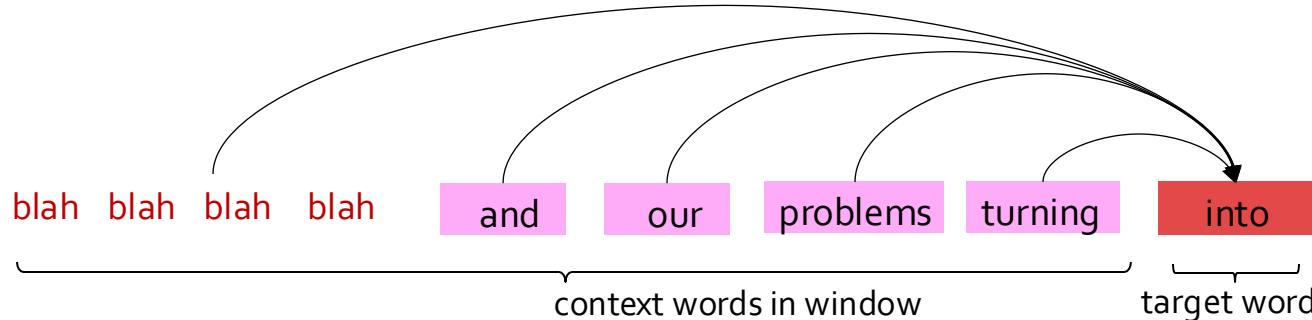
Summary

- Language Models (LM): distributions over language
- Measuring LM quality: use perplexity on held-out data.
- Count-based LMs have limitations.
 - Challenge with large N's: sparsity problem — many zero counts/probs.
 - Challenge with small N's: lack of long-range dependencies.
- Next: Rethinking language modeling as a statistical learning problem.

Beyond Counting: Language Models as a Learning Problem

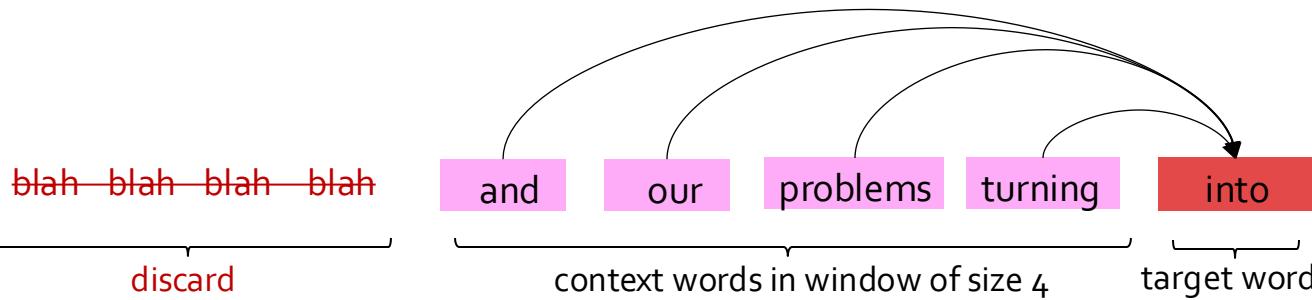
LM as a Machine Learning Problem

- Given the embeddings of the context, predict the word on the right side.
 - Dropping the right context for simplicity -- not a fundamental limitation.
- Discard anything beyond its context window



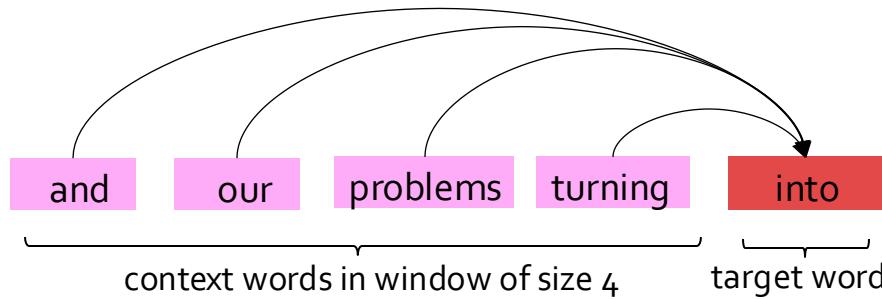
LM as a Machine Learning Problem

- Given the embeddings of the context, predict the word on the right side.
 - Dropping the right context for simplicity -- not a fundamental limitation.
- Discard anything beyond its context window



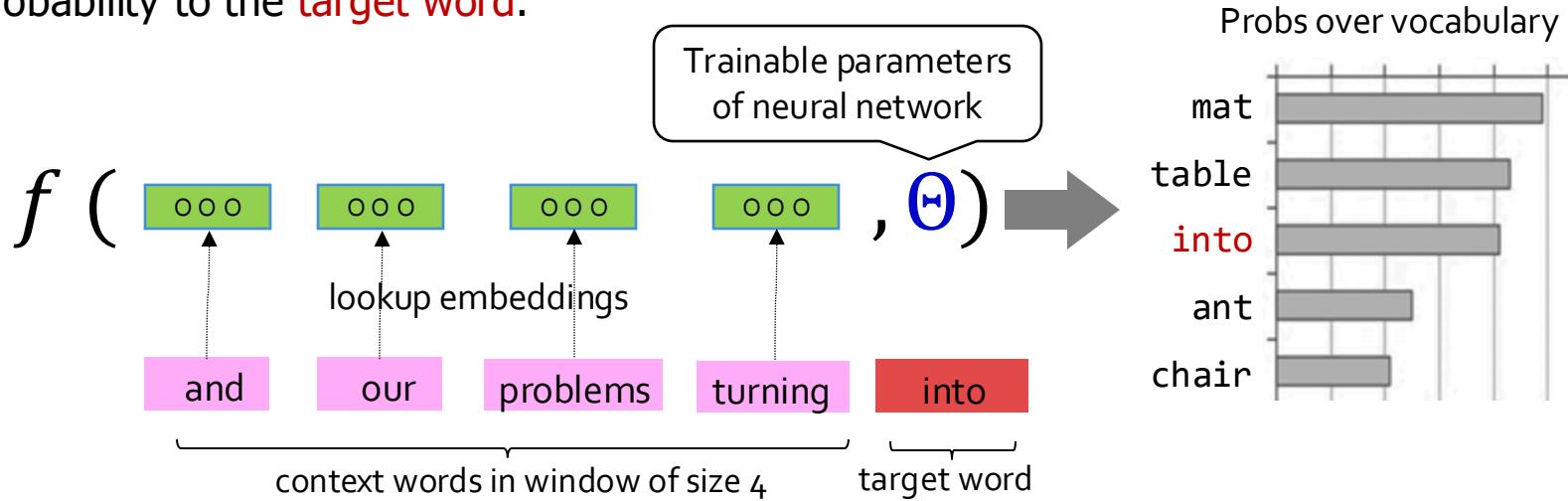
LM as a Machine Learning Problem

- Given the embeddings of the context, predict the word on the right side.
 - Dropping the right context for simplicity -- not a fundamental limitation.
- Discard anything beyond its context window



A Fixed-Window Neural LM

- Given the embeddings of the **context**, predict a **target word** on the right side.
 - Dropping the right context for simplicity -- not a fundamental limitation.
- Training this model is basically optimizing its parameters Θ such that it assigns high probability to the **target word**.



A Fixed-Window Neural LM

- It will also lay the foundation for the future models (recurrent nets, transformers, ...)
- But first we need to figure out how to train neural networks!

How do you build
this function?

$$f($$



Trainable parameters
of neural network



$$, \Theta)$$



Neural Networks
for rescue!

lookup embeddings

our

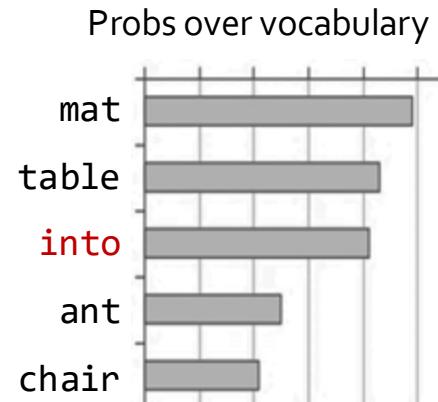
problems

turning

into

context words in window of size 4

target word



From Counting (N-Gram) to Neural Models

- n-gram models of text generation [Jelinek+ 1980's, ...]
 - Applications: Speech Recognition, Machine Translation
 - “Shallow” statistical/neural language models (2000’s) [Bengio+ 1999 & 2001, ...]
-

NeurIPS 2000

A Neural Probabilistic Language Model

Yoshua Bengio*, Réjean Ducharme and Pascal Vincent
Département d’Informatique et Recherche Opérationnelle
Centre de Recherche Mathématiques
Université de Montréal
Montréal, Québec, Canada, H3C 3J7
{bengioy,ducharme,vincentp}@iro.umontreal.ca

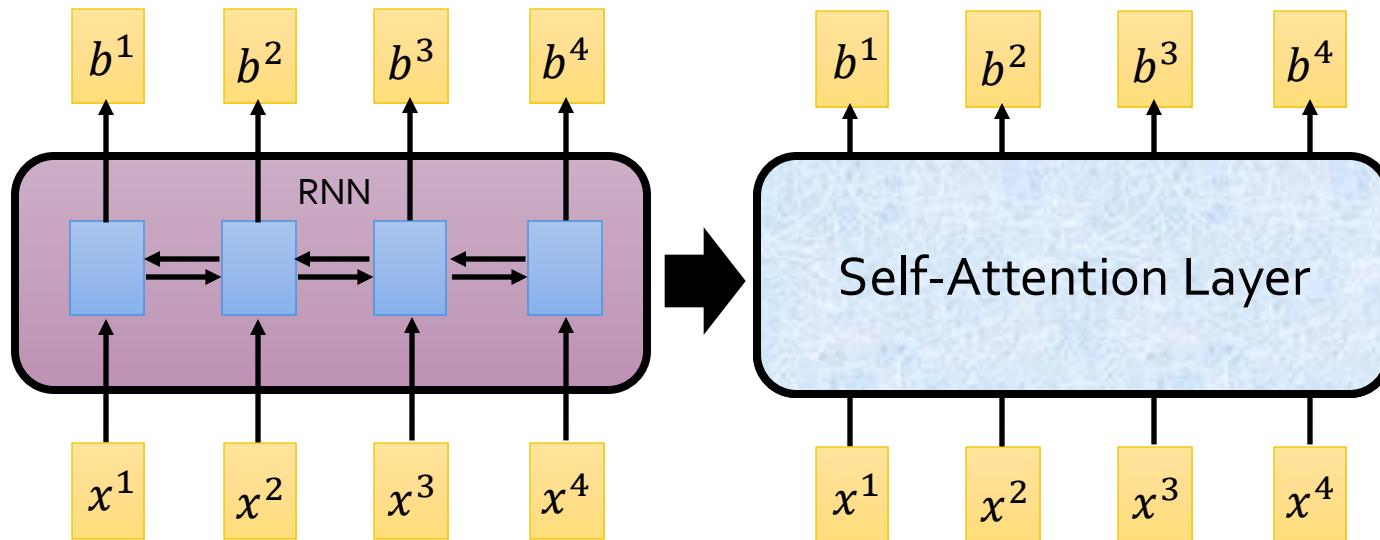
Summary

- **Language Modeling (LM)**, a useful predictive objective for language
- **Perplexity**, a measure of an LM's predictive ability
- **N-gram models** (~1980 to early 2000's),
 - Early instances of LMs
 - Difficult to scale to large window sizes
- **Shallow neural LMs** (early and mid-2000's),
 - These will be effective predictive models based on feed-forward networks
- **Recurrent neural LMs** (2010s),
 - Compact models used recursively

Self-Attention

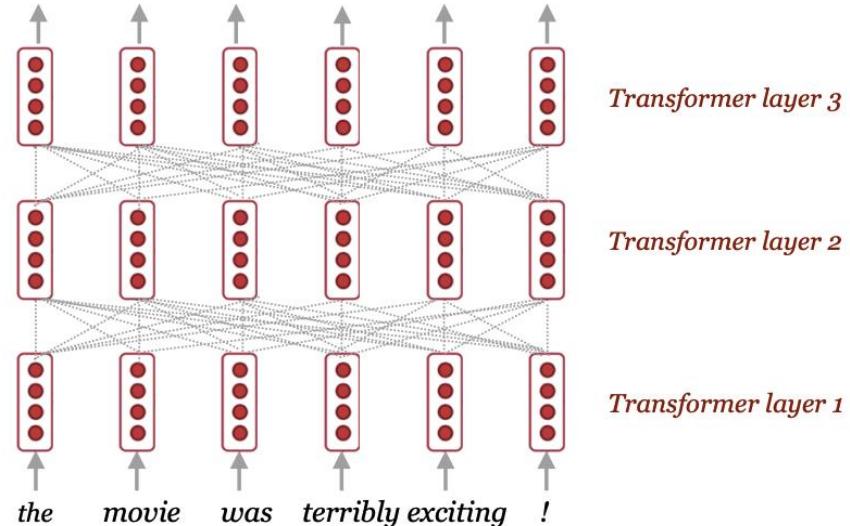
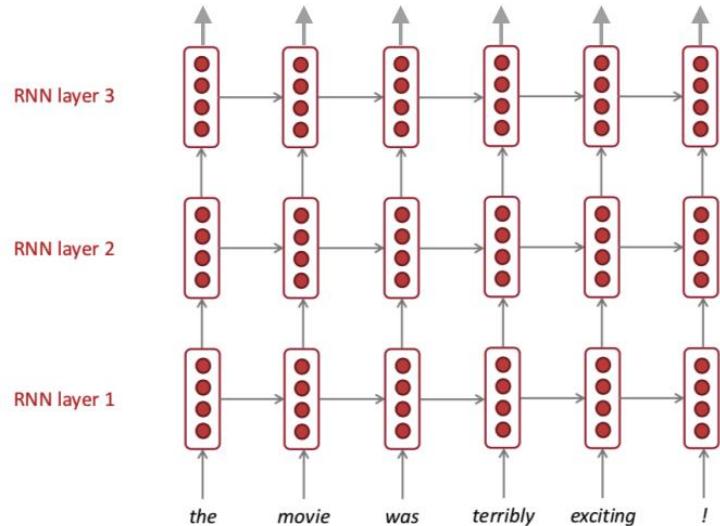
Self-Attention

- b^i is obtained based on the whole input sequence.
- can be parallelly computed.



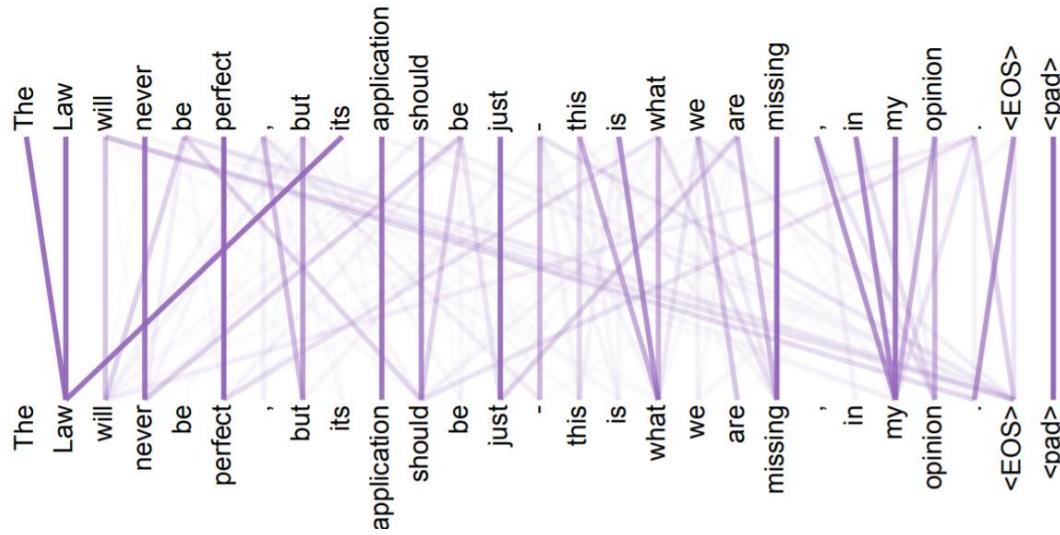
Idea: replace any thing done by RNN with **self-attention**.

RNN vs Transformer



Attention

- Core idea: build a mechanism to focus ("attend") on a particular part of the context.



Defining Self-Attention

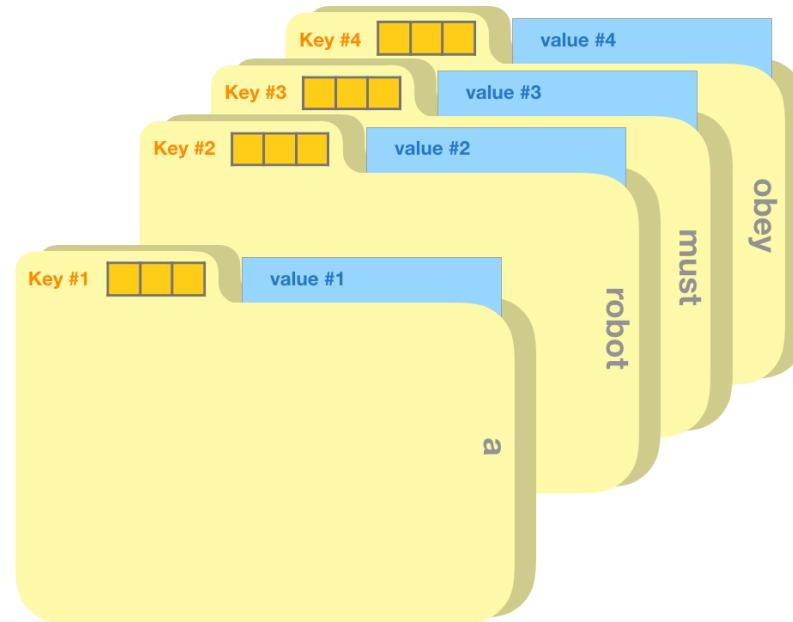
- Terminology:
 - **Query**: to match others
 - **Key**: to be matched
 - **Value**: information to be extracted

Defining Self-Attention

- Terminology:
 - **Query**: to match others
 - **Key**: to be matched
 - **Value**: information to be

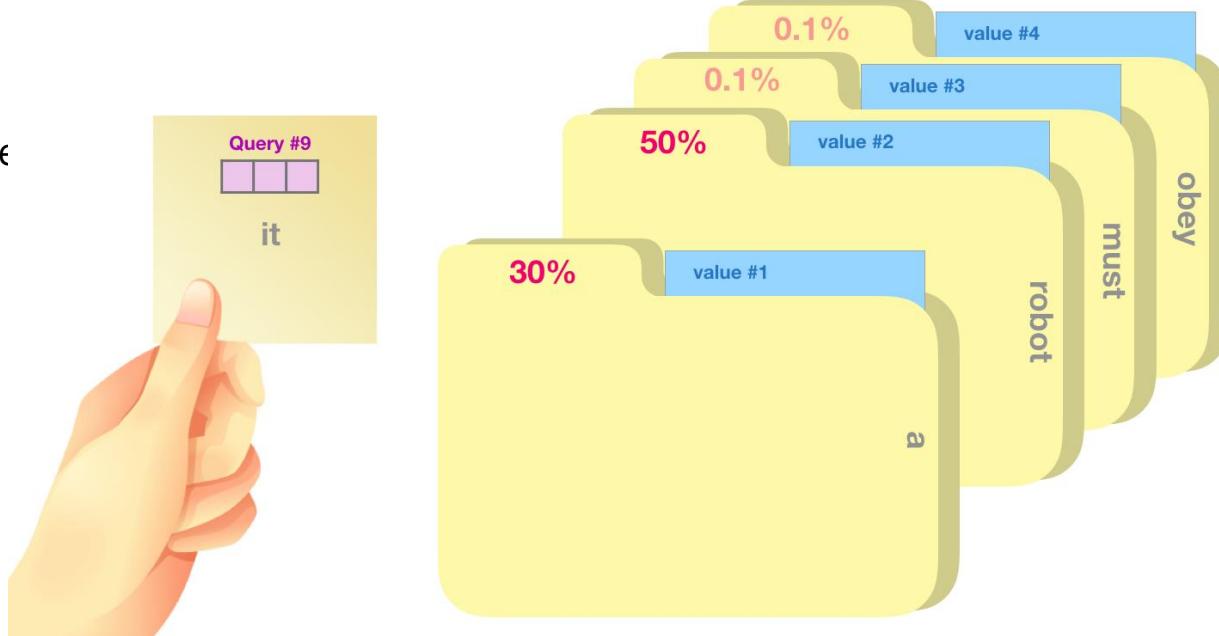


An analogy



Defining Self-Attention

- Terminology:
 - **Query**: to match others
 - **Key**: to be matched
 - **Value**: information to be



q: query (to match others)

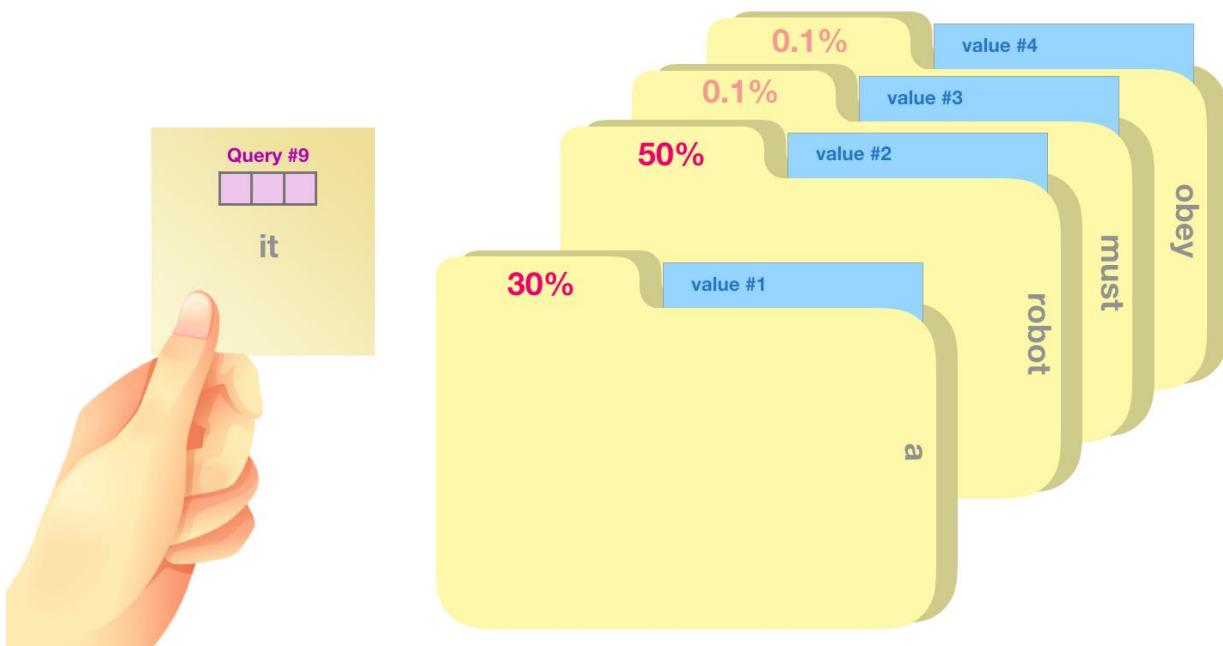
$$q_i = W^q x_i$$

k: key (to be matched)

$$k_i = W^k x_i$$

v: value (information to be extracted)

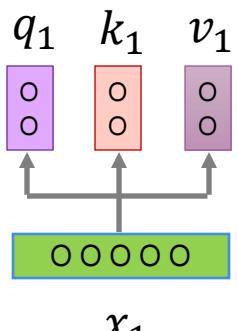
$$v_i = W^v x_i$$



q : query (to match others)
 $q_i = W^q x_i$

k : key (to be matched)
 $k_i = W^k x_i$

v : value (information to be extracted)
 $v_i = W^v x_i$

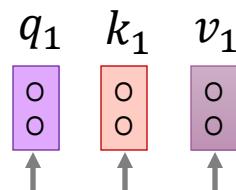


The

q: query (to match others)
 $q_i = W^q x_i$

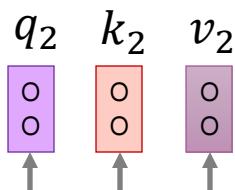
k: key (to be matched)
 $k_i = W^k x_i$

v: value (information to be extracted)
 $v_i = W^v x_i$



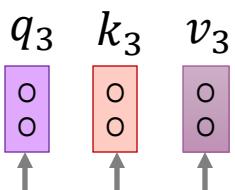
x_1

The



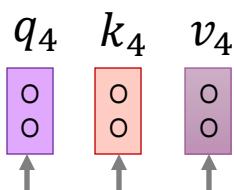
x_2

cat



x_3

sat



x_4

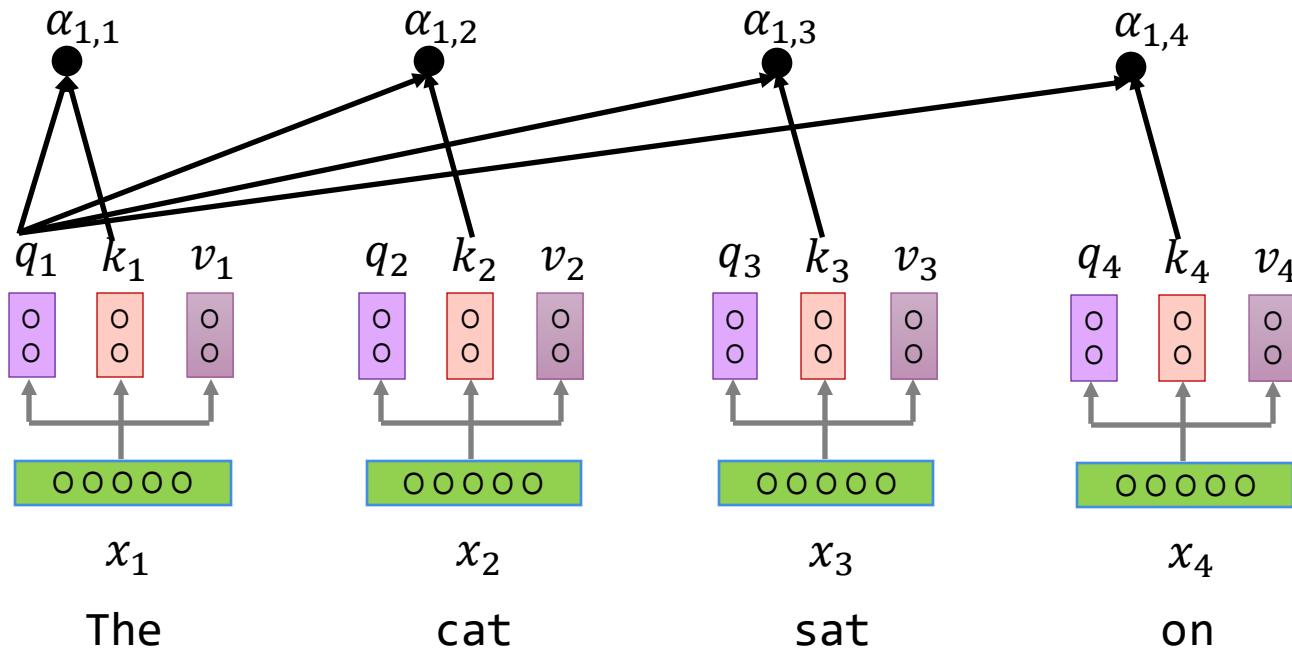
on

$$\alpha_{1,i} = \frac{q^1 \cdot k^i}{\sqrt{d}}$$

Scaled dot product

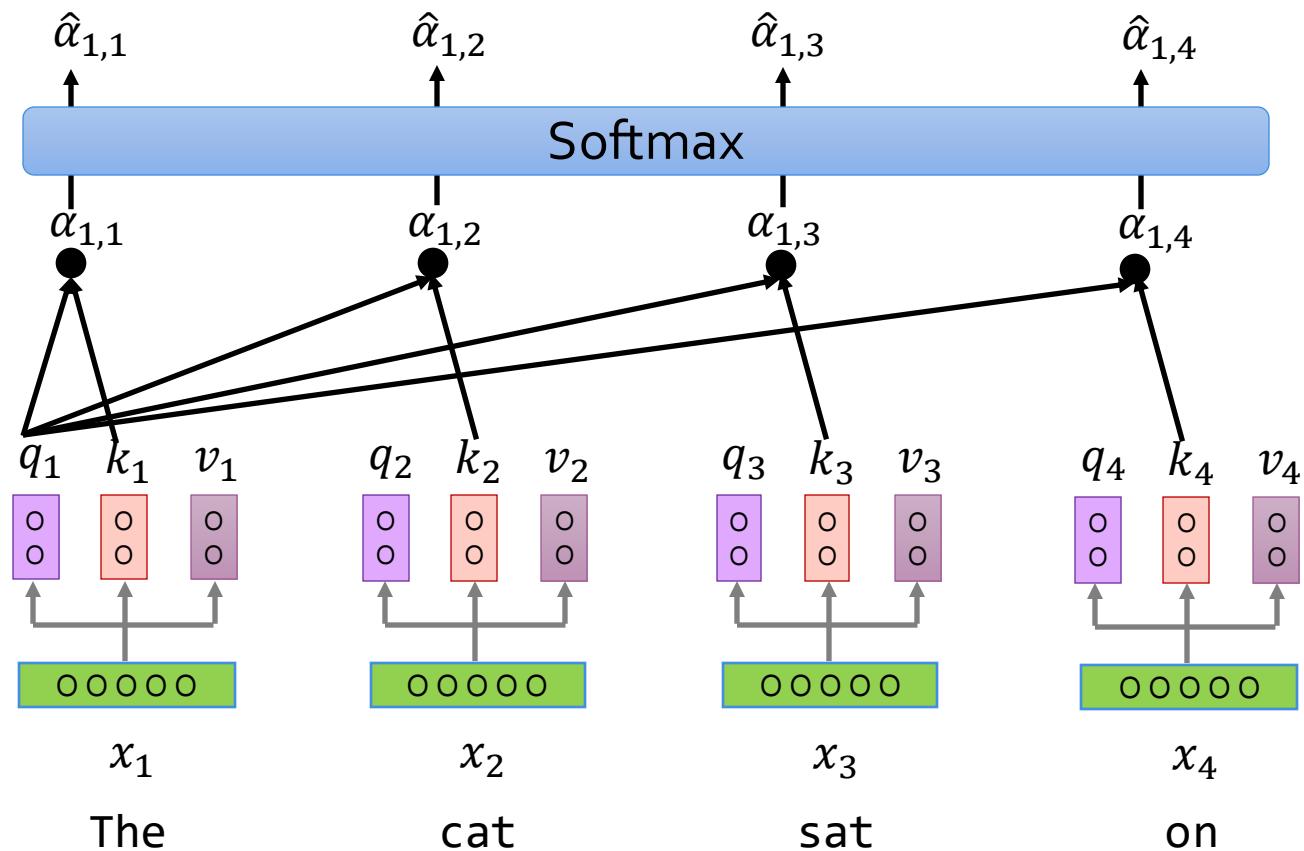
How much
should "The"
attend to other
positions?

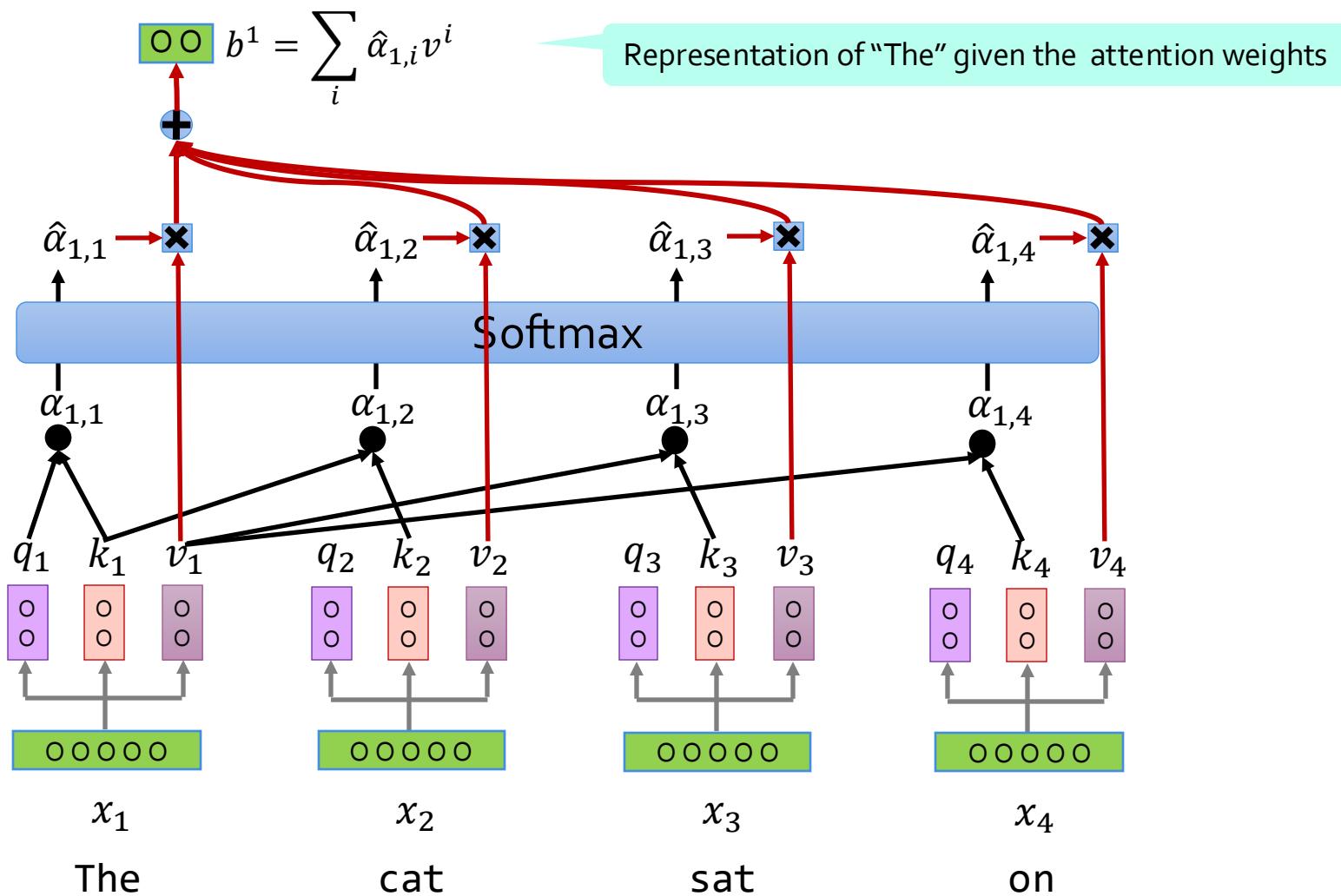
q: query (to match others)
k: key (to be matched)
v: value (information to be extracted)



$$\sigma(z)_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

How much
should "The"
attend to other
positions?





Self-Attention

- Can write it in matrix form:
- Given input \mathbf{x} :

$$Q = \mathbf{W}^q \mathbf{x}$$

$$K = \mathbf{W}^k \mathbf{x}$$

$$V = \mathbf{W}^v \mathbf{x}$$

$$\text{Attention}(\mathbf{x}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$



hardmaru

@hardmaru

...

The most important formula in deep learning after 2018

Self-Attention

What is self-attention? Self-attention calculates a weighted average of feature representations with the weight proportional to a similarity score between pairs of representations. Formally, an input sequence of n tokens of dimensions d , $X \in \mathbf{R}^{n \times d}$, is projected using three matrices $W_Q \in \mathbf{R}^{d \times d_q}$, $W_K \in \mathbf{R}^{d \times d_k}$, and $W_V \in \mathbf{R}^{d \times d_v}$ to extract feature representations Q , K , and V , referred to as query, key, and value respectively with $d_k = d_q$. The outputs Q , K , V are computed as

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V. \quad (1)$$

So, self-attention can be written as,

$$S = D(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right)V, \quad (2)$$

where softmax denotes a *row-wise* softmax normalization function. Thus, each element in S depends on all other elements in the same row.

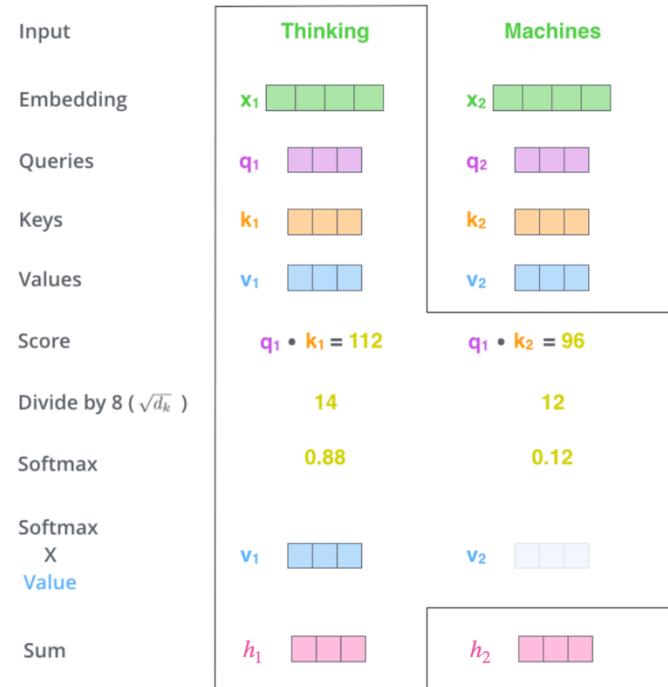
9:08 PM · Feb 9, 2021 · Twitter Web App

553 Retweets 42 Quote Tweets 3,338 Likes

Question

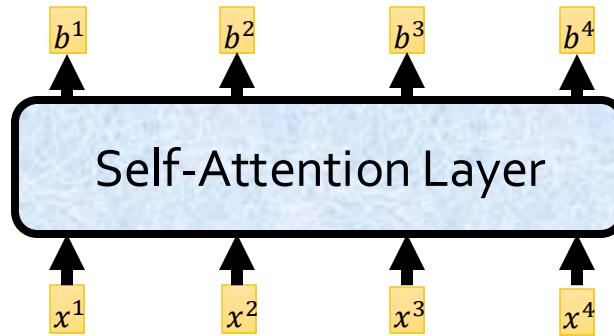
- What would be the output vector for the word “Thinking”?

- (a) $0.5\mathbf{v}_1 + 0.5\mathbf{v}_2$
- (b) $0.54\mathbf{v}_1 + 0.46\mathbf{v}_2$
- (c) $0.88\mathbf{v}_1 + 0.12\mathbf{v}_2$
- (d) $0.12\mathbf{v}_1 + 0.88\mathbf{v}_2$



Self-Attention: Back to Big Picture

- **Attention** is a powerful mechanism to create context-aware representations
- A way to focus on select parts of the input



- Better at maintaining **long-distance dependencies** in the context.

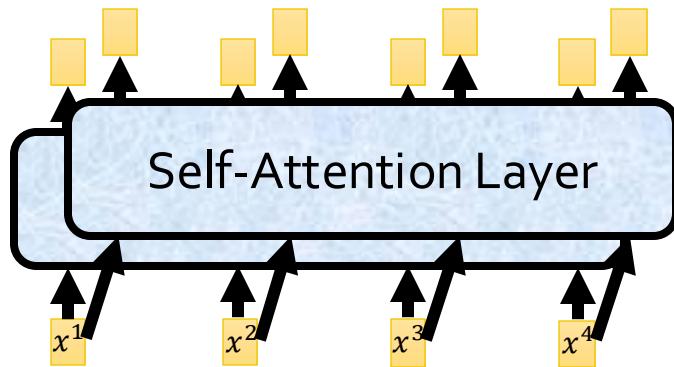
Properties of Self-Attention

Layer Type	Complexity per Layer	Sequential Operations
Self-Attention	$O(n^2 \cdot d)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$

- Per-layer statistics; n = sequence length, d = hidden dimension
- Complexity per layer: Quadratic function of n for SA
- Sequential operations: # of operations that must be performed sequentially
 - $O(1)$ sequential operations for SA.
 - SA layers computes all the operations in parallel across all tokens in the sequence
 - **Efficient** implementations

Multi-Headed Self-Attention

- Multiple parallel attention layers is quite common.
 - Each attention layer has its own parameters.
 - Concatenate the results and run them through a linear projection.



Multi-Headed Self-Attention

- Just concatenate all the heads and apply an output projection matrix.

$$\text{MultiHeadedAttention}(x) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^0$$
$$\text{head}_i = \text{Attention}(xW_i^q, xW_i^k, xW_i^v)$$

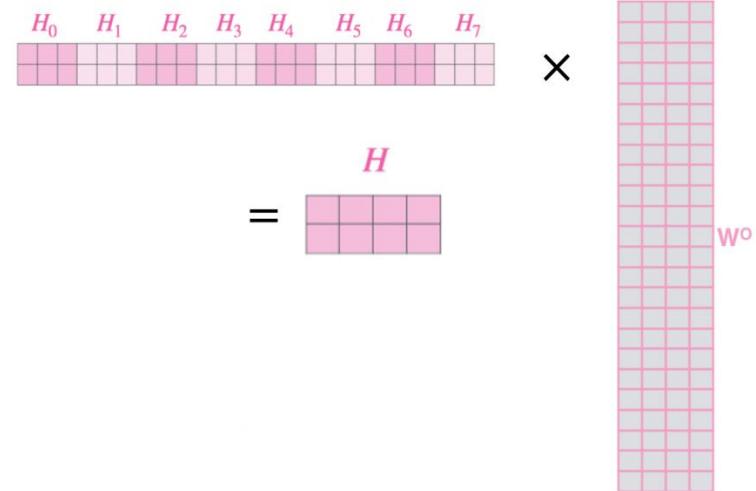
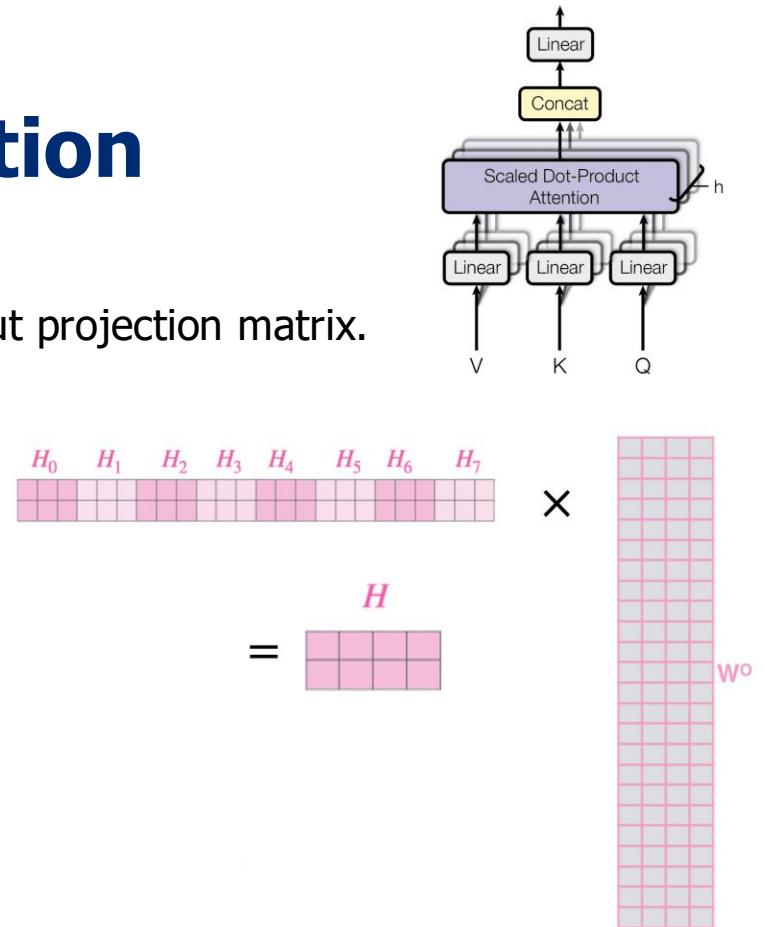
- In practice, we can use a reduced dimension for each head:

$$W_i^q, W_i^k, W_i^v \in \mathbb{R}^{d/h}$$

d = hidden dimension

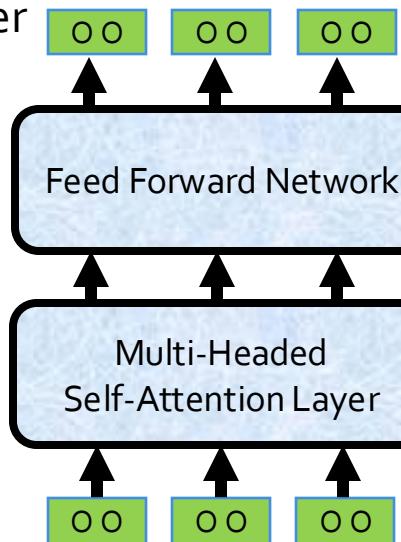
h = # of heads

- The total computational cost is similar to that of a single-head attention with full dimensionality.

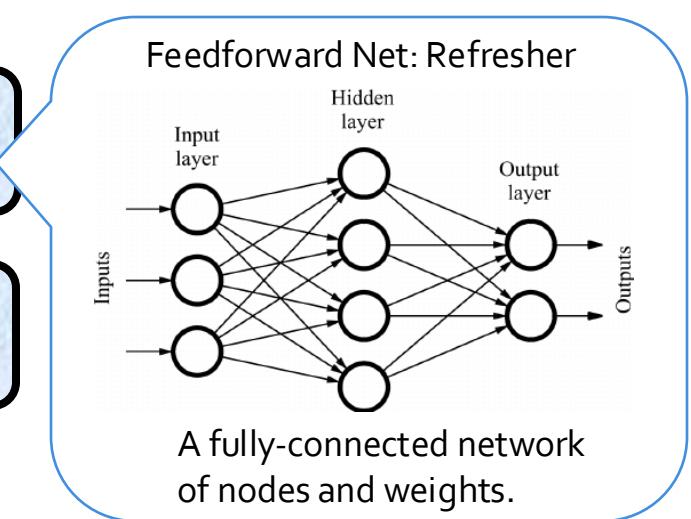


Combine with FFN

- Add a **feed-forward network** on top it to add more expressivity.
 - This allows the model to apply another transformation to the contextual representations (or “post-process” them).
 - Usually, the dimensionality of the hidden feedforward layer is 2-8 times larger than the input dimension.

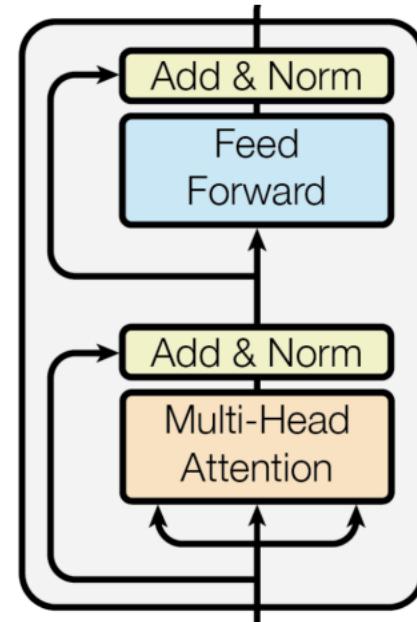


$$\text{FFN}(\mathbf{x}) = f(cW_1 + b_1)W_2 + b_2$$



How Do We Prevent Vanishing Gradients?

- Residual connections let the model “skip” layers
 - These connections are particularly useful for training deep networks
- Use layer normalization to stabilize the network and allow for proper gradient flow



Putting it Together: Self-Attention Block

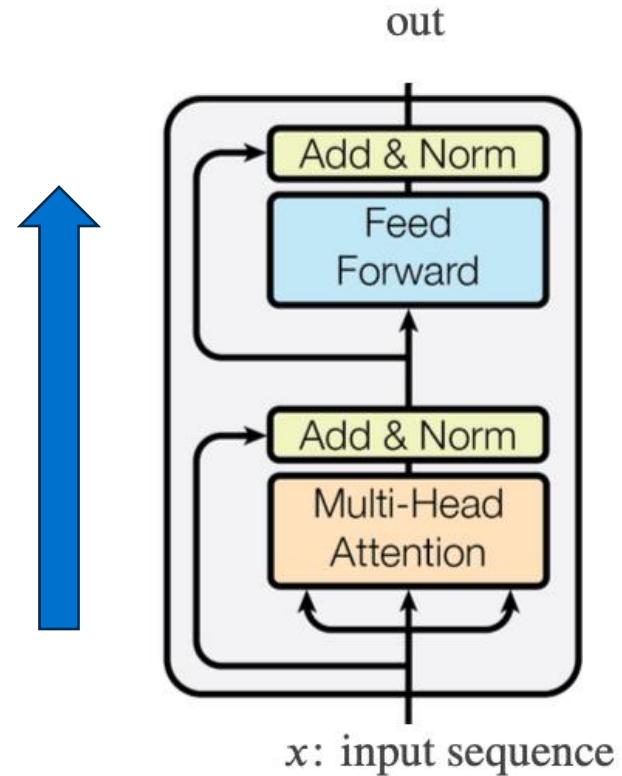
Given input x :

$$\text{out} = \text{LN}(\tilde{\mathbf{c}} + \mathbf{c}')$$

$$\tilde{\mathbf{c}} = \text{FFN}(\mathbf{c}') = f(\mathbf{c}'W_1 + b_1)W_2 + b_2$$

$$\mathbf{c}' = \text{LN}(\mathbf{c} + x)$$

$$\mathbf{c} = \text{MultiHeadedAttention}(x)$$



Summary: Self-Attention Block

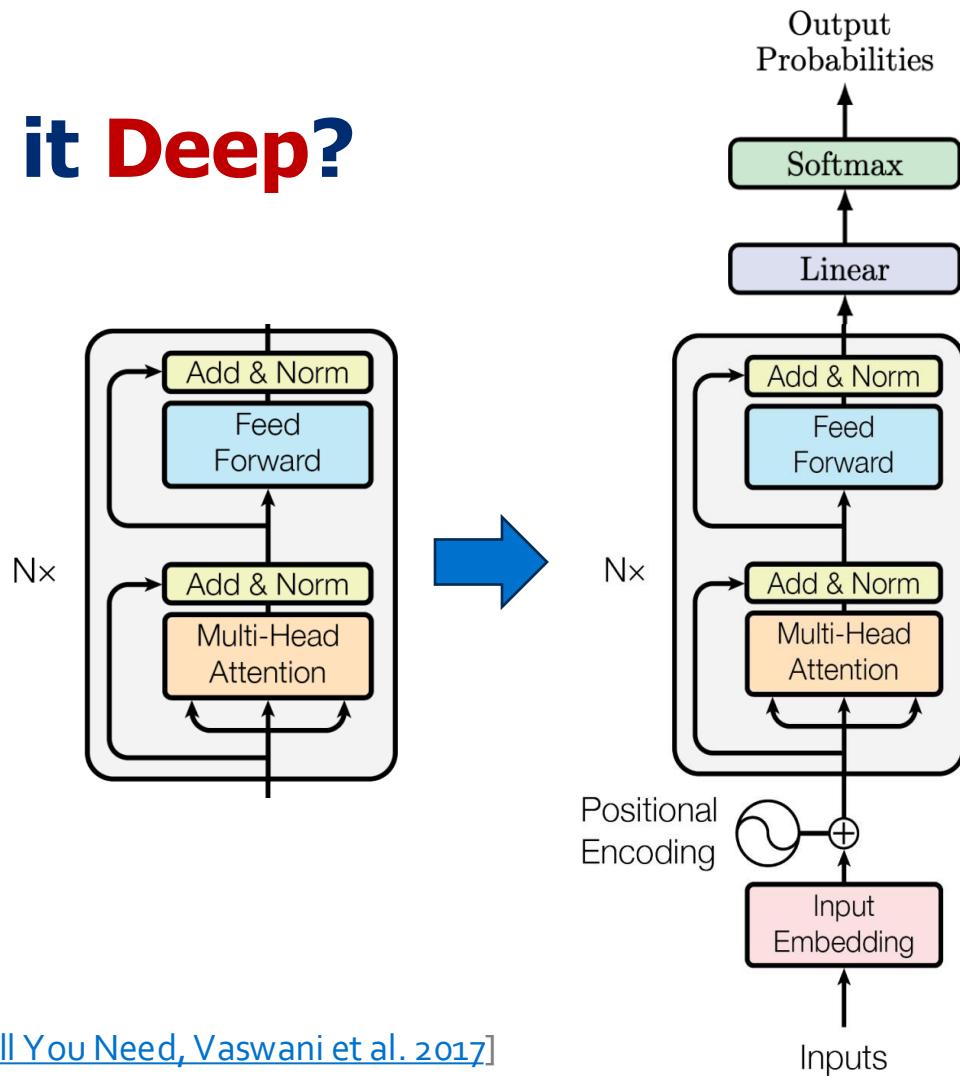
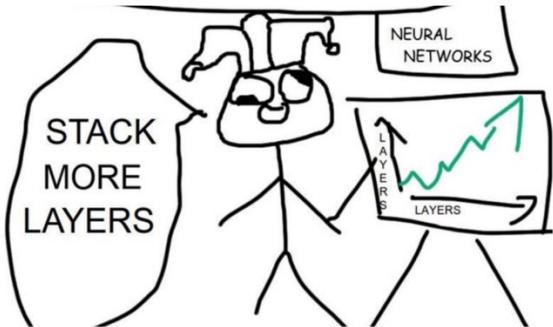
- **Self-Attention:** A critical building block of modern language models.
 - The idea is to compose meanings of words weighted according some similarity notion.
- **Next:** We will combine self-attention blocks to build various architectures known as Transformer.



Transformer

How Do We Make it Deep?

- Stack more layers!



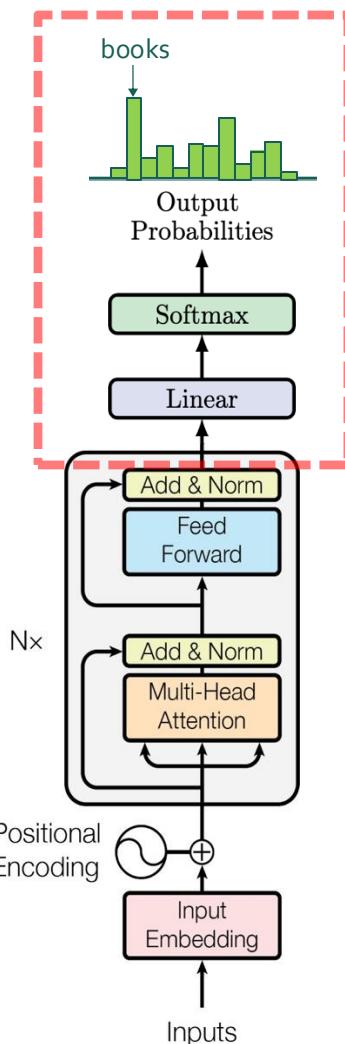
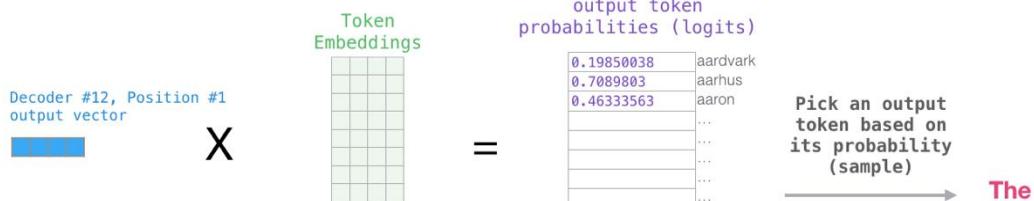
From Representations to Prediction

- To perform prediction, add a classification head on top of the final layer of the transformer.
- This can be per token (Language modeling)
- Or can be for the entire sequence (only one token)

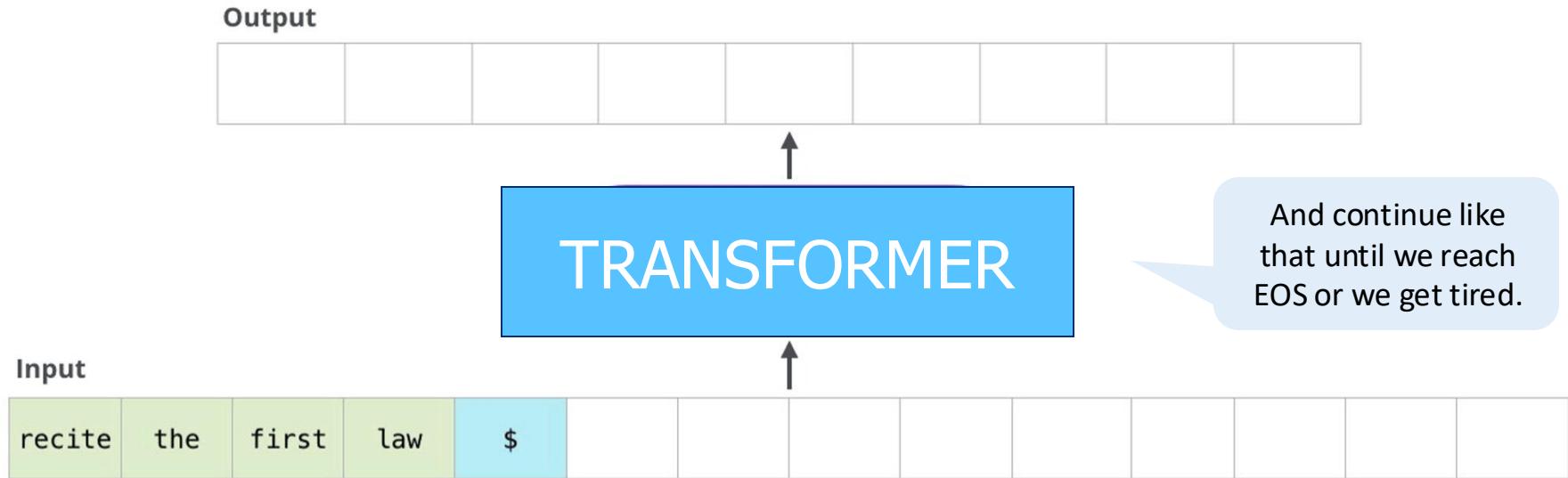
$\text{out} \in \mathbb{R}^{S \times d}$ (S: Sequence length)

$\text{logits} = \text{Linear}_{(d, V)}(\text{out}) = f(\text{out} \cdot W_V) \in \mathbb{R}^{S \times V}$

$\text{probabilities} = \text{softmax}(\text{logits}) \in \mathbb{R}^{S \times V}$



Transformer-based Language Modeling



Training a Transformer Language Model

- **Goal:** Train a Transformer for language modeling (i.e., predicting the next word).
- **Approach:** Train it so that each position is predictor of the next (right) token.
 - We just shift the input to right by one, and use as labels

(gold output) $Y = \text{cat sat on the mat } </s>$

EOS special token



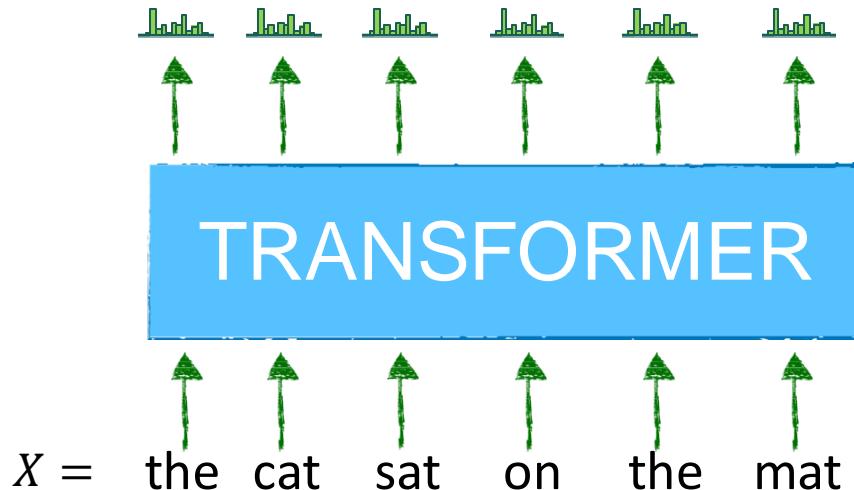
```
X = text[:, :-1]  
Y = text[:, 1:]
```

$X = \text{the cat sat on the mat}$

Training a Transformer Language Model

- For each position, compute their corresponding **distribution** over the whole vocab.

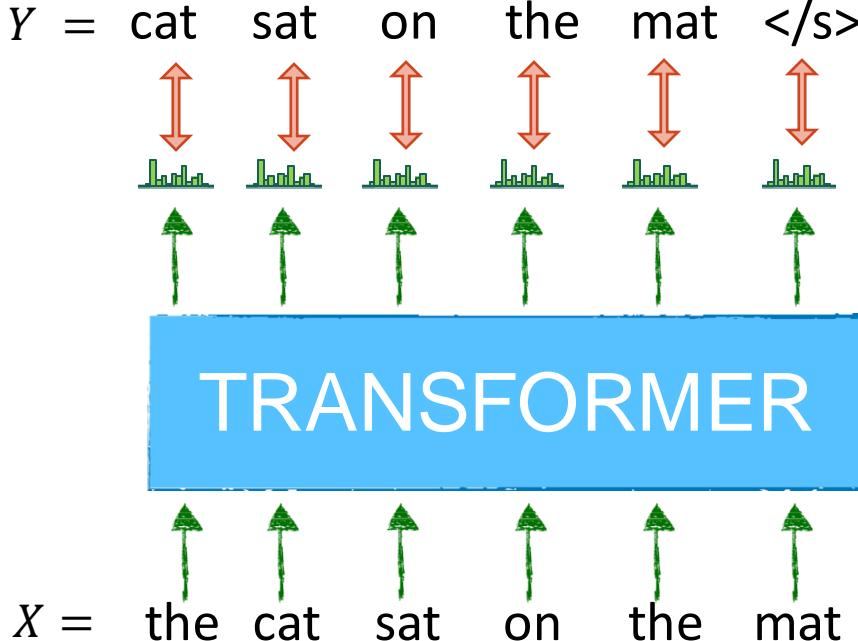
(gold output) $Y = \text{cat sat on the mat } </s>$



Training a Transformer Language Model

- For each position, compute the **loss** between the distribution and the gold output label.

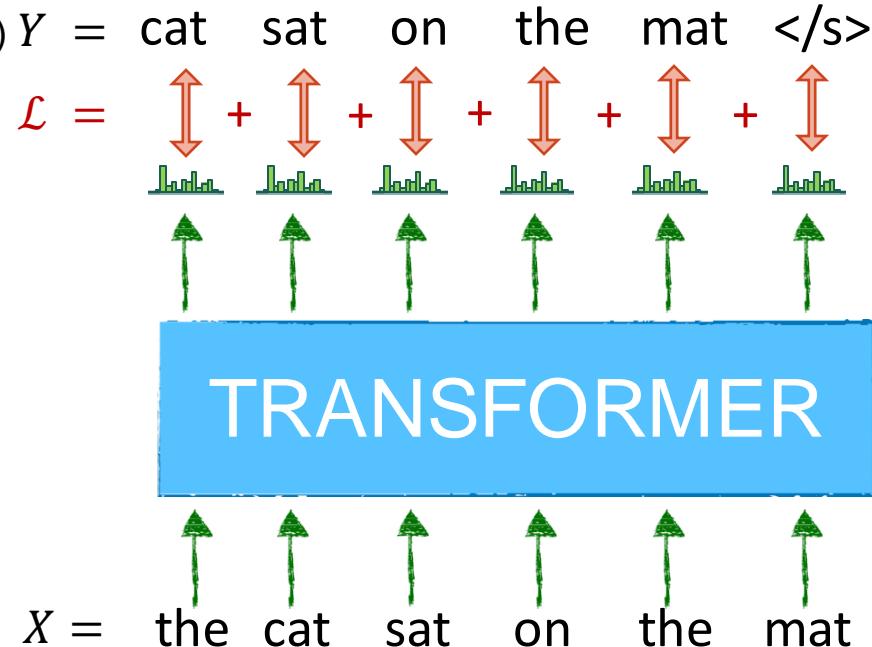
(gold output) $Y = \text{cat sat on the mat } </s>$



Training a Transformer Language Model

- Sum the position-wise loss values to obtain a **global loss**.

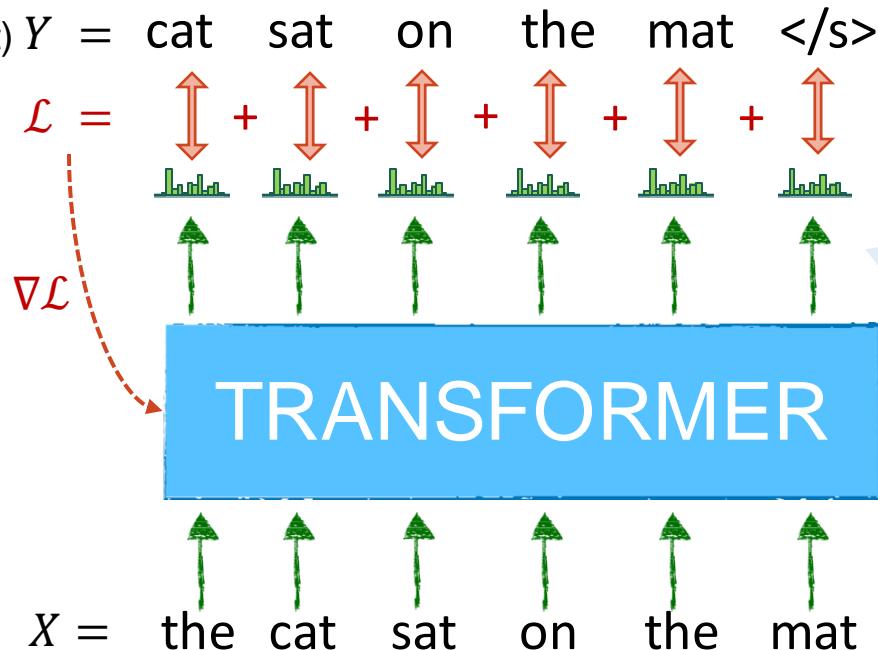
(gold output) $Y = \text{cat sat on the mat } </s>$



Training a Transformer Language Model

- Using this loss, do **Backprop** and **update** the Transformer parameters.

(gold output) $Y = \text{cat sat on the mat } </s>$

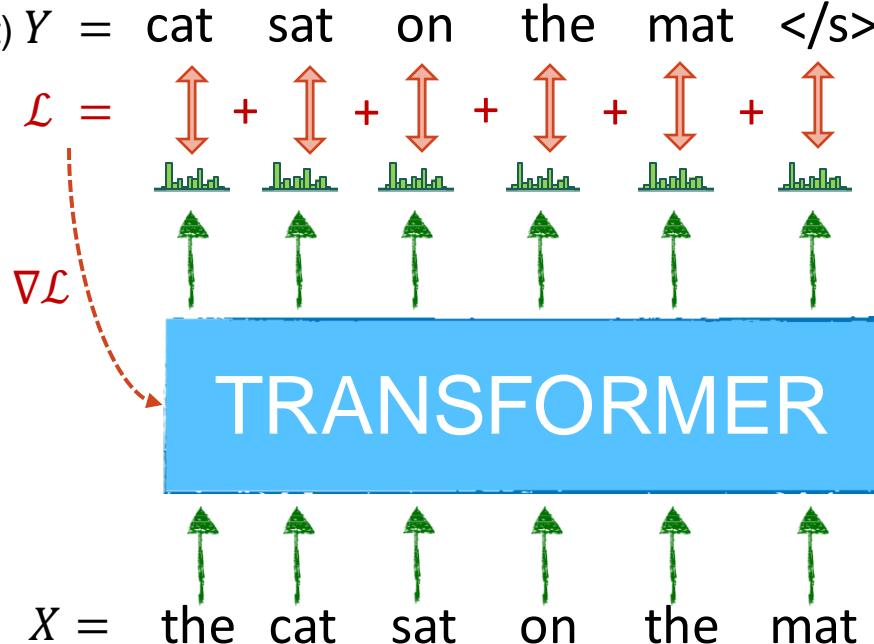


Well, this is not quite right 😊
...
what is the problem with this?

Training a Transformer Language Model

- The model would solve the task by **copying** the next token to output (data leakage).
 - Does **not** learn anything useful

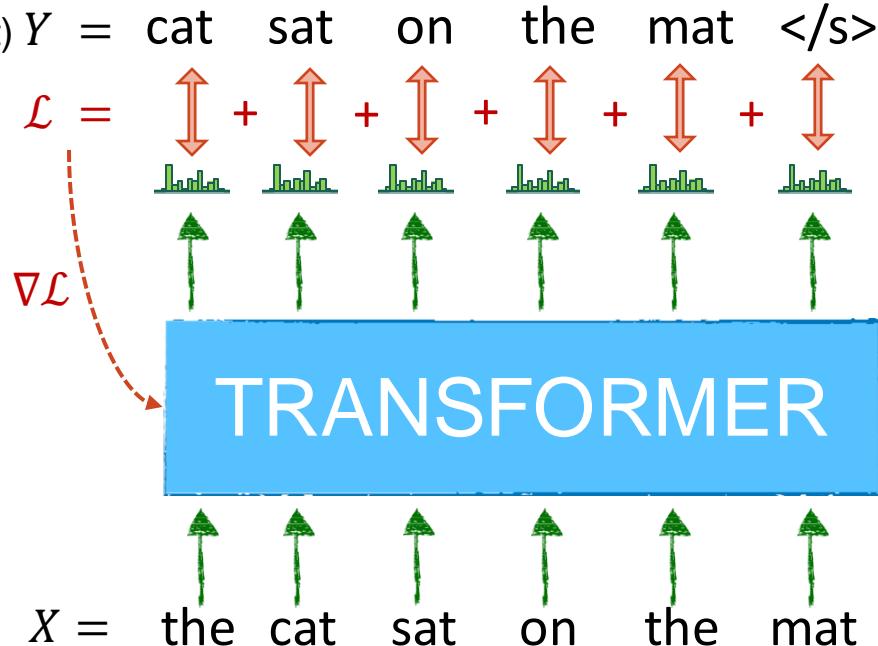
(gold output) $Y = \text{cat sat on the mat } </s>$



Training a Transformer Language Model

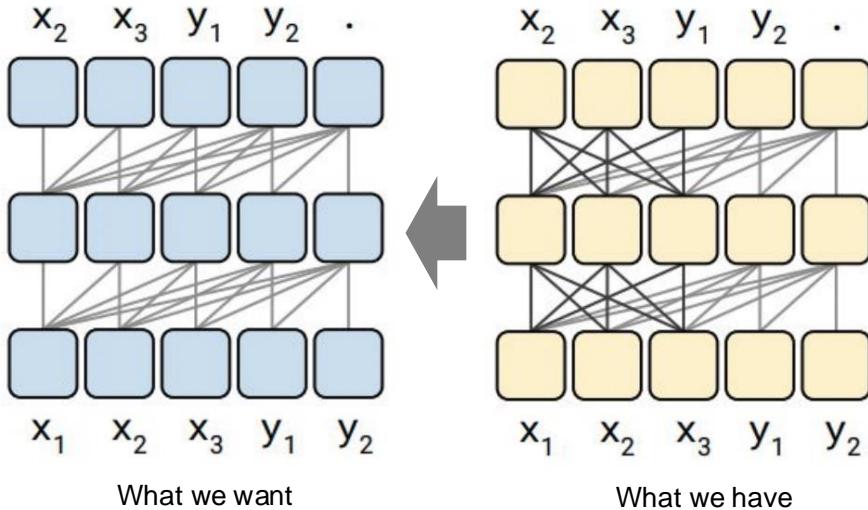
- We need to **prevent information leakage** from future tokens! How?

(gold output) $Y = \text{cat sat on the mat } </s>$

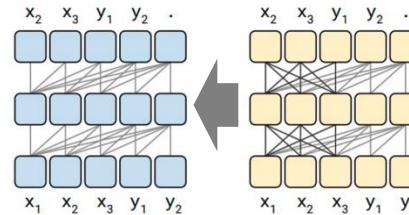


Attention mask

Attention raw scores												
0	-0.08	1.24	0.69	-0.98	1.43	-0.6	0.7	0.16	0.93	1.28	-1.61	-1.1
1	-0.09	-0.0	-0.7	0.06	0.25	0.23	0.26	0.18	0.78	-0.21	-1.01	1.01
2	0.86	1.19	1.59	0.86	-0.13	-0.15	-2.13	-0.98	-0.87	-1.72	1.87	-0.72
3	0.12	-0.03	-0.02	0.88	-0.46	-0.7	0.54	-0.42	-1.89	-0.38	0.04	-0.84
4	0.51	0.17	0.13	-1.64	0.24	-0.02	1.68	-0.36	0.64	0.36	0.27	0.66
5	0.24	-1.44	0.43	0.74	0.96	-1.21	-0.31	1.54	1.66	1.14	0.58	-1.44
6	0.26	-0.1	0.93	0.72	-0.38	1.65	0.47	-0.96	-0.17	-0.9	-1.57	0.22
7	-0.55	0.81	0.71	1.7	-0.8	-1.14	-0.32	1.78	-0.7	-0.04	1.54	0.81
8	0.74	-0.76	-0.44	-0.08	-1.38	-0.13	1.25	-1.37	1.84	0.3	0.57	0.74
9	-0.97	-0.91	0.15	0.35	-0.81	0.11	1.14	-1.52	1.06	1.87	0.5	-0.3
10	1.56	0.9	0.39	1.46	1.44	-1.05	0.9	-0.73	0.36	-0.67	-0.62	-0.43
11	0.32	0.74	0.44	-0.1	1.19	0.83	0.29	2.06	0.51	-0.26	1.51	0.11



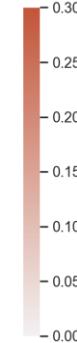
Attention mask



Attention raw scores

	1	2	3	4	5	6	7	8	9	10	11	12
0	-0.08	1.24	0.69	-0.98	1.43	-0.6	0.7	0.16	0.93	1.28	-1.61	-1.1
1	-0.09	-0.0	-0.7	0.06	0.25	0.23	0.26	0.18	0.78	-0.21	-1.01	1.01
2	0.86	1.19	1.59	0.86	-0.13	-0.15	-2.13	-0.98	-0.87	-1.72	1.87	-0.72
3	0.12	-0.03	-0.02	0.88	-0.46	-0.7	0.54	-0.42	-1.89	-0.38	0.04	-0.84
4	0.51	0.17	0.13	-1.64	0.24	-0.02	1.68	-0.36	0.64	0.36	0.27	0.66
5	0.24	-1.44	0.43	0.74	0.96	-1.21	-0.31	1.54	1.66	1.14	0.58	-1.44
6	0.26	-0.1	0.93	0.72	-0.38	1.65	0.47	-0.96	-0.17	-0.9	-1.57	0.22
7	-0.55	0.81	0.71	1.7	-0.8	-1.14	-0.32	1.78	-0.7	-0.04	1.54	0.81
8	0.74	-0.76	-0.44	-0.08	-1.38	-0.13	1.25	-1.37	1.84	0.3	0.57	0.74
9	-0.97	-0.91	0.15	0.35	-0.81	0.11	1.14	-1.52	1.06	1.87	0.5	-0.3
10	1.56	0.9	0.39	1.46	1.44	-1.05	0.9	-0.73	0.36	-0.67	-0.62	-0.43
11	0.32	0.74	0.44	-0.1	1.19	0.83	0.29	2.06	0.51	-0.26	1.51	0.11

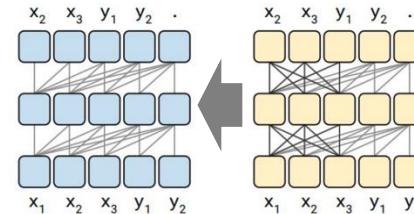
Attention mask



Attention mask

Attention raw scores												
0	-0.08	1.24	0.69	-0.98	1.43	-0.6	0.7	0.16	0.93	1.28	-1.61	-1.1
1	-0.09	-0.0	-0.7	0.06	0.25	0.23	0.26	0.18	0.78	-0.21	-1.01	1.01
2	0.86	1.19	1.59	0.86	-0.13	-0.15	-2.13	-0.98	-0.87	-1.72	1.87	-0.72
3	0.12	-0.03	-0.02	0.88	-0.46	-0.7	0.54	-0.42	-1.89	-0.38	0.04	-0.84
4	0.51	0.17	0.13	-1.64	0.24	-0.02	1.68	-0.36	0.64	0.36	0.27	0.66
5	0.24	-1.44	0.43	0.74	0.96	-1.21	-0.31	1.54	1.66	1.14	0.58	-1.44
6	0.26	-0.1	0.93	0.72	-0.38	1.65	0.47	-0.96	-0.17	-0.9	-1.57	0.22
7	-0.55	0.81	0.71	1.7	-0.8	-1.14	-0.32	1.78	-0.7	-0.04	1.54	0.81
8	0.74	-0.76	-0.44	-0.08	-1.38	-0.13	1.25	-1.37	1.84	0.3	0.57	0.74
9	-0.97	-0.91	0.15	0.35	-0.81	0.11	1.14	-1.52	1.06	1.87	0.5	-0.3
10	1.56	0.9	0.39	1.46	1.44	-1.05	0.9	-0.73	0.36	-0.67	-0.62	-0.43
11	0.32	0.74	0.44	-0.1	1.19	0.83	0.29	2.06	0.51	-0.26	1.51	0.11

X



Attention mask

0	1	2	3	4	5	6	7	8	9	10	11
0	1.0	-inf									
1	1.0	1.0	-inf								
2	1.0	1.0	1.0	-inf							
3	1.0	1.0	1.0	1.0	-inf						
4	1.0	1.0	1.0	1.0	1.0	-inf	-inf	-inf	-inf	-inf	-inf
5	1.0	1.0	1.0	1.0	1.0	1.0	-inf	-inf	-inf	-inf	-inf
6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	-inf	-inf	-inf	-inf
7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	-inf	-inf	-inf
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	-inf	-inf
9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	-inf
10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	-inf
11	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0



Note matrix multiplication is quite fast in GPUs.

Presented by: Arman Cohan

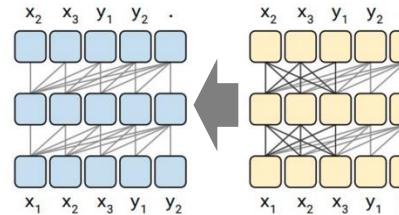
Attention mask

Attention raw scores												
	1	2	3	4	5	6	7	8	9	10	11	12
1	-0.09	1.34	0.69	0.98	3.43	-0.1	0.7	0.16	0.93	1.28	1.81	-1.13
2	-0.09	-0.7	0.06	0.25	0.26	0.26	0.18	0.78	-0.21	1.01	1.01	1.01
3	-0.09	1.19	1.09	0.96	-0.15	0.15	-0.11	-0.06	-0.07	1.73	1.67	-0.23
4	0.12	-0.03	-0.02	0.88	-0.46	-0.1	0.54	-0.42	1.86	-0.38	0.54	-0.84
5	0.51	0.17	0.51	0.34	0.24	-0.02	1.68	-0.36	0.84	0.36	0.27	0.86
6	0.24	-1.64	0.43	0.78	0.96	1.2	-0.31	1.54	1.88	1.14	0.58	-1.44
7	0.26	-0.1	0.63	0.72	-0.36	1.84	0.47	0.96	-0.17	-0.5	0.58	0.22
8	0.55	0.81	0.71	1.7	-0.8	1.16	-0.32	1.78	-0.7	-0.54	1.54	0.81
9	0.74	0.76	-0.48	-0.68	-0.38	0.13	0.13	1.25	1.37	1.84	0.3	0.57
10	0.87	0.91	0.15	0.36	-0.81	0.11	1.14	-0.51	1.08	1.87	0.5	-0.37
11	0.52	0.74	0.44	0.71	1.19	0.89	0.29	2.06	0.31	-0.28	1.51	0.11
12	1	2	3	4	5	6	7	8	9	10	11	12

X

=

Raw attention scores												
	1	2	3	4	5	6	7	8	9	10	11	12
1	1.11	inf										
2	0.1	1.0	inf									
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
11	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
12	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0



Masked attention raw scores

0	-0.08	-inf	-inf	-inf	-inf	-inf						
1	-0.09	-0.0	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf
2	0.86	1.19	1.59	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf
3	0.12	-0.03	-0.02	0.88	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf
4	0.51	0.17	0.13	-1.64	0.24	-inf	-inf	-inf	-inf	-inf	-inf	-inf
5	0.24	-1.44	0.43	0.74	0.96	-1.21	-inf	-inf	-inf	-inf	-inf	-inf
6	0.26	-0.1	0.93	0.72	-0.38	1.65	0.47	-inf	-inf	-inf	-inf	-inf
7	-0.55	0.81	0.71	1.7	-0.8	-1.14	-0.32	1.78	-inf	-inf	-inf	-inf
8	0.74	-0.76	-0.44	-0.08	-1.38	-0.13	1.25	-1.37	1.84	-inf	-inf	-inf
9	-0.97	-0.91	0.15	0.35	-0.81	0.11	1.14	-1.52	1.06	1.87	-inf	-inf
10	1.56	0.9	0.39	1.46	1.44	-1.05	0.9	-0.73	0.36	-0.67	-0.62	-inf
11	0.32	0.74	0.44	-0.1	1.19	0.83	0.29	2.06	0.51	-0.26	1.51	0.11
12	1	2	3	4	5	6	7	8	9	10	11	12

Slide credit: Arman Cohan

Attention mask

The effect is more than just pruning out some of the wirings in self-attention block.

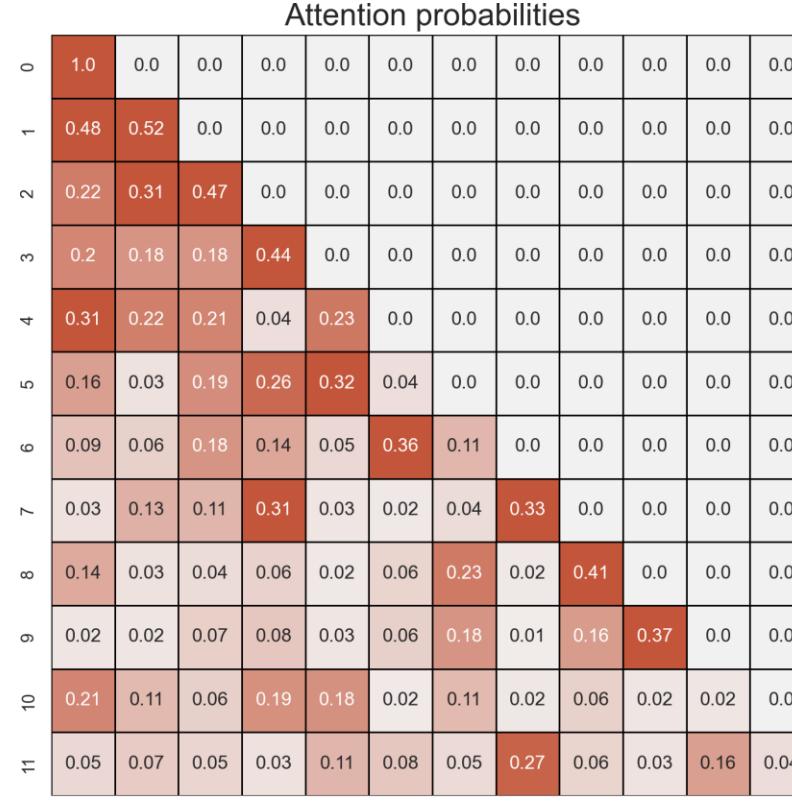
Attention raw scores											
-0.09	1.24	0.69	0.38	3.43	-0.1	0.7	1.16	0.93	1.28	1.81	-1.13
-0.09	-0.0	-0.7	0.06	0.25	0.23	0.26	0.18	0.78	-0.21	1.01	1.01
-0.06	1.19	1.59	0.98	-0.15	0.15	1.11	0.96	-0.97	1.22	1.67	-0.72
-0.12	-0.03	-0.42	0.88	-0.46	-0.1	0.54	-0.42	1.86	-0.38	0.94	-0.84
-0.51	0.17	0.55	0.34	0.24	0.02	1.68	0.36	0.04	0.37	0.66	-0.01
-0.24	-1.64	0.43	0.78	0.96	1.2	-0.31	1.54	1.88	1.14	0.58	-1.44
-0.26	-0.1	0.63	0.72	0.36	1.84	0.47	1.96	-0.17	-0.5	0.67	0.22
-0.55	0.81	0.71	1.7	-0.18	0.32	1.78	-0.7	-0.94	1.54	0.81	-0.74
-0.74	-0.48	-0.48	0.08	-0.33	-0.13	1.25	1.37	1.84	0.3	0.57	0.74
-0.87	0.91	0.15	0.30	-0.81	0.11	1.14	-0.50	1.08	1.87	0.5	-0.37
-0.52	0.74	0.44	0.1	1.19	0.89	0.29	2.06	0.31	-0.28	1.51	0.11

X

Raw attention scores											
0	1.0	ad									
1	1.0	1.2	ad								
2	1.0	1.2	1.0	1.0	ad						
3	1.0	1.0	1.0	1.0	1.0	1.0	ad	ad	ad	ad	ad
4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	ad	ad	ad	ad
5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	ad	ad	ad
6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	ad	ad
7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	ad
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	ad
9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
11	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Masked attention raw scores											
-0.08	ad	ad	ad	ad	ad	ad	ad	ad	ad	ad	ad
-0.09	-0.0	ad	ad	ad	ad	ad	ad	ad	ad	ad	ad
-0.06	1.19	1.59	ad	ad	ad	ad	ad	ad	ad	ad	ad
-0.12	-0.03	0.88	ad	ad	ad	ad	ad	ad	ad	ad	ad
-0.51	0.17	0.33	-1.64	0.24	ad	ad	ad	ad	ad	ad	ad
-0.24	1.64	1.43	0.74	0.68	1.21	ad	ad	ad	ad	ad	ad
-0.26	-0.1	0.93	0.72	-0.38	1.65	0.67	ad	ad	ad	ad	ad
-0.56	0.81	0.71	1.7	-0.8	1.14	-0.22	1.78	ad	ad	ad	ad
-0.74	-0.76	-0.44	-0.36	-0.36	1.13	1.37	1.84	ad	ad	ad	ad
-0.87	-0.91	0.15	0.35	-0.81	0.11	1.14	1.52	1.06	1.87	ad	ad
-1.56	0.8	0.39	1.46	1.44	1.06	0.8	-0.73	1.36	-0.67	-0.62	ad
-0.32	0.74	0.44	-0.1	1.19	0.89	0.29	2.06	0.51	-0.28	1.51	0.11

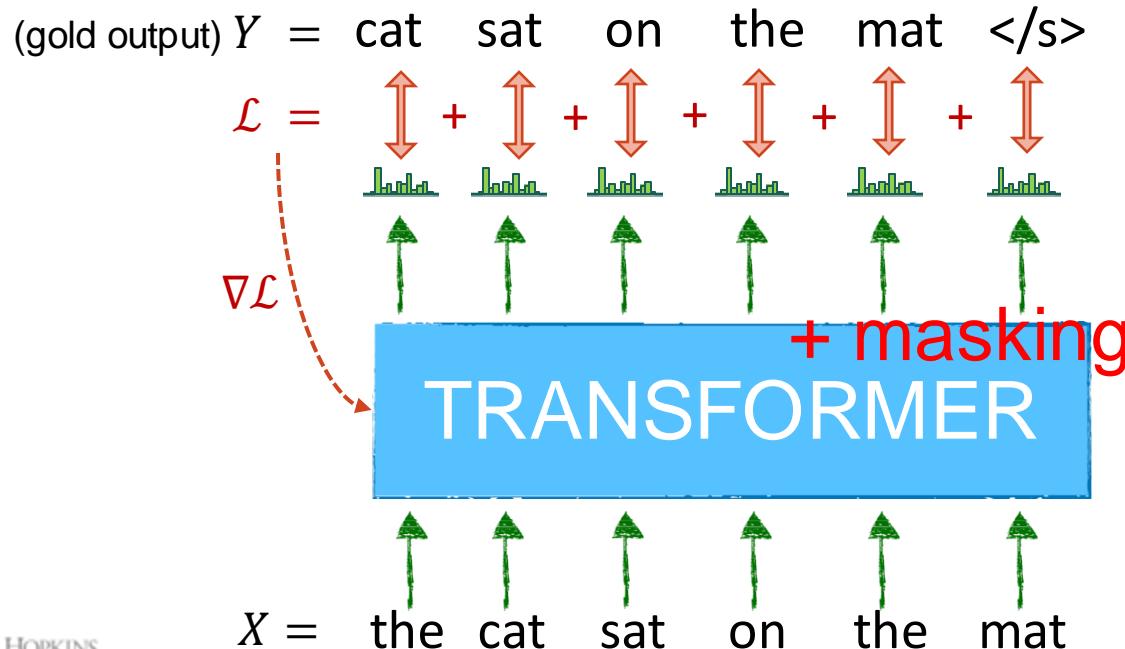
softmax



Slide credit: Alman Cohan

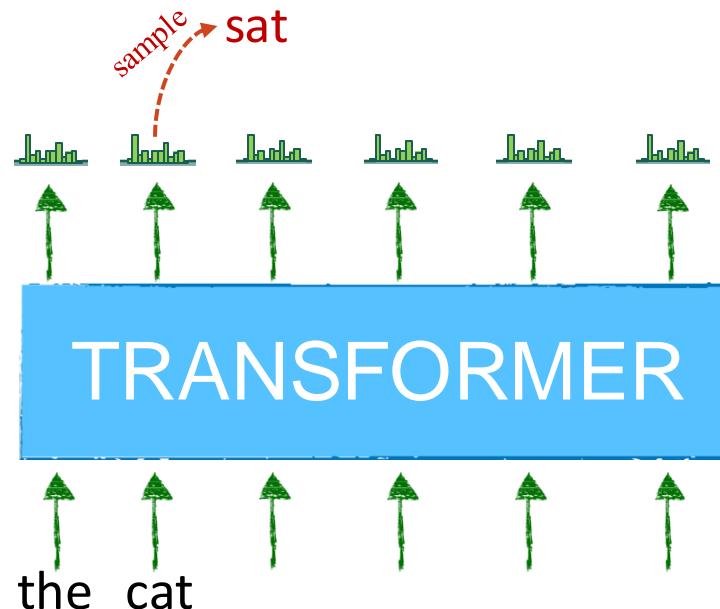
Training a Transformer Language Model

- We need to **prevent information leakage** from future tokens! How?



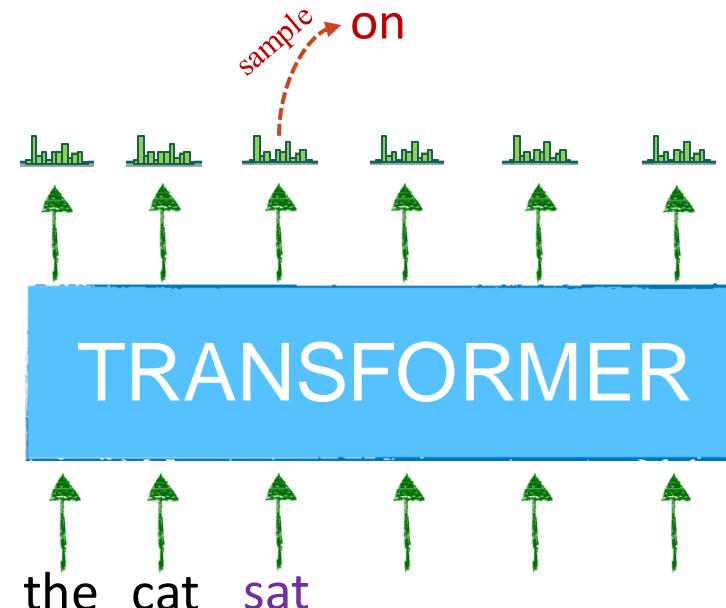
How to use the model to generate text?

- Use the output of previous step as input to the next step repeatedly



How to use the model to generate text?

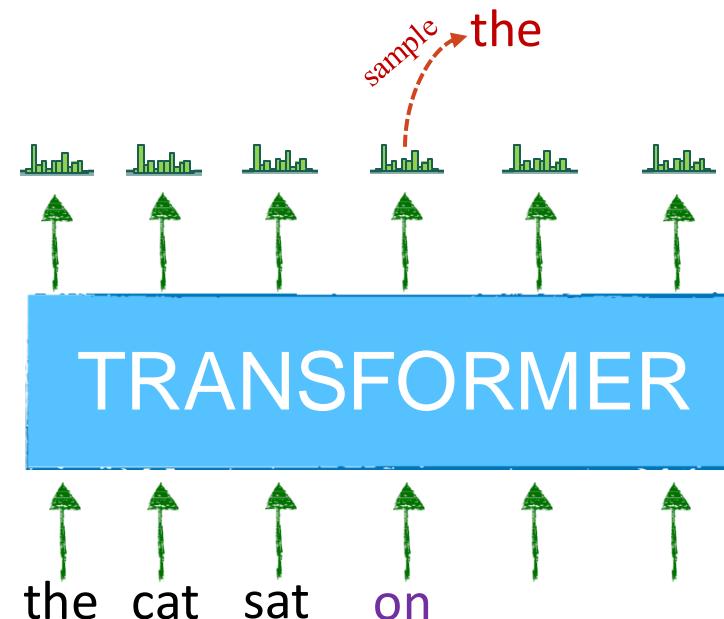
- Use the output of previous step as input to the next step repeatedly



The probabilities get revised upon adding a new token to the input.

How to use the model to generate text?

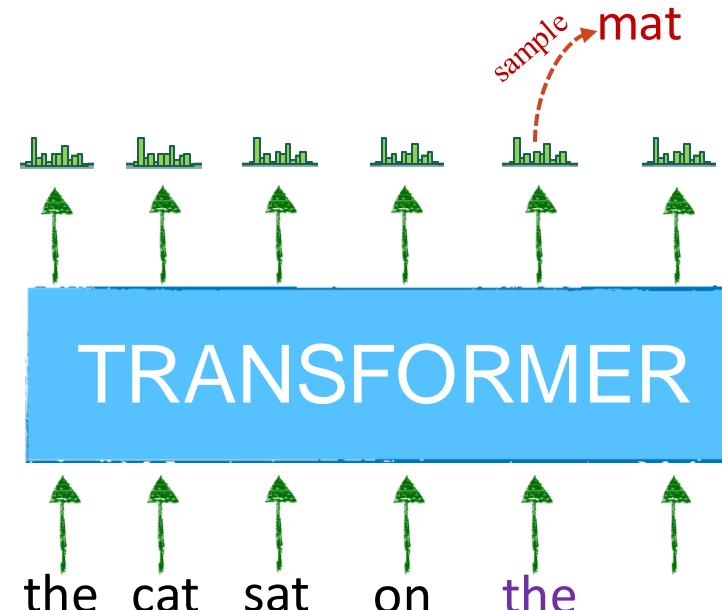
- Use the output of previous step as input to the next step repeatedly



The probabilities get revised upon adding a new token to the input.

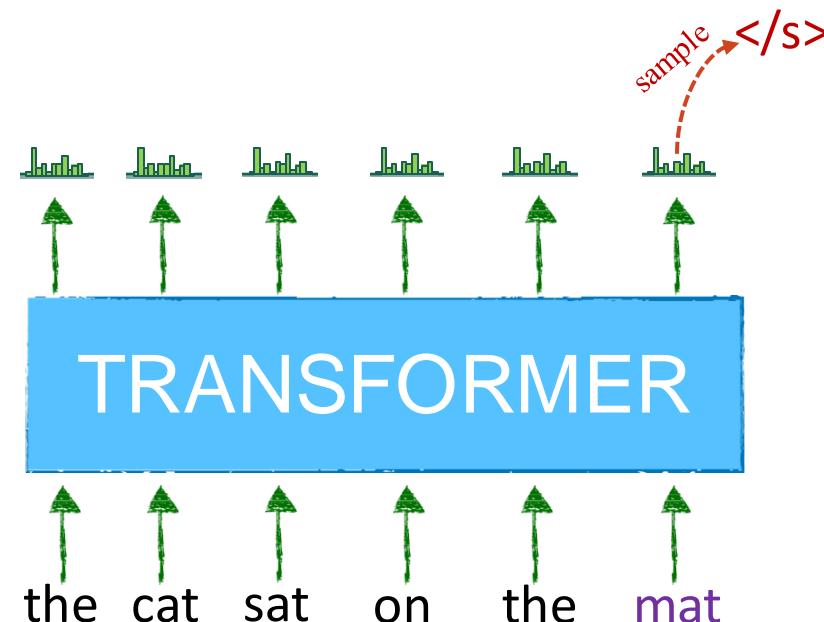
How to use the model to generate text?

- Use the output of previous step as input to the next step repeatedly



How to use the model to generate text?

- Use the output of previous step as input to the next step repeatedly





How efficient is this decoding?



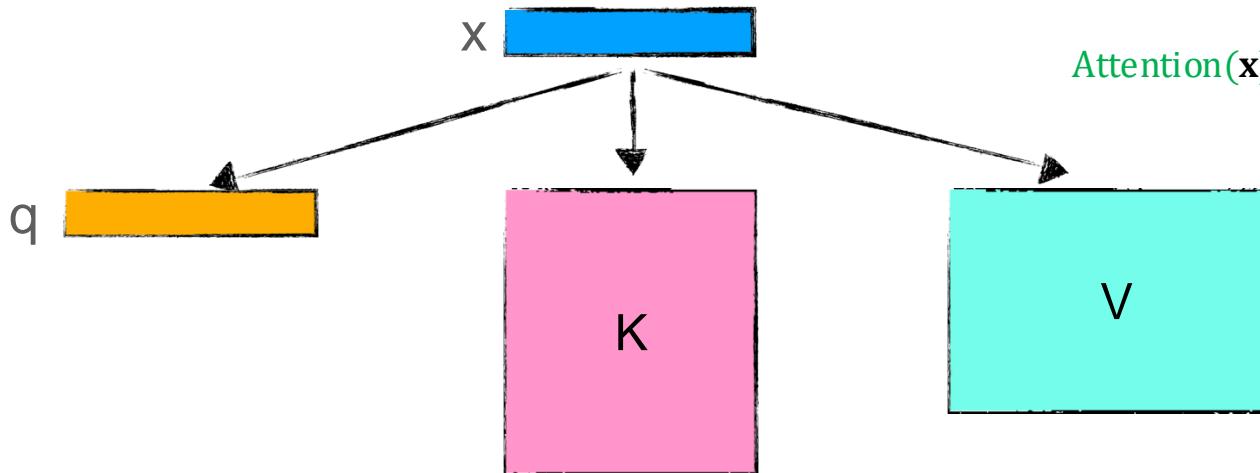
Making decoding more efficient

$$Q = \mathbf{W}^q \mathbf{x}$$

$$K = \mathbf{W}^k \mathbf{x}$$

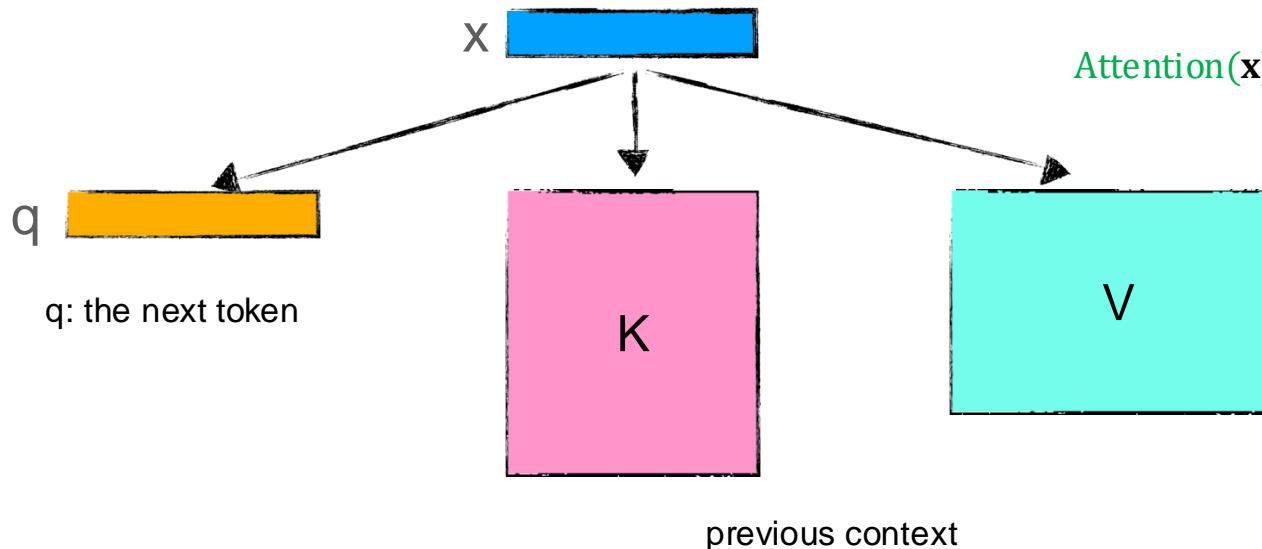
$$V = \mathbf{W}^v \mathbf{x}$$

$$\text{Attention}(\mathbf{x}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$



[Slide credit: Arman Cohan]

Making decoding more efficient



$$\begin{aligned} Q &= \mathbf{W}^q \mathbf{x} \\ K &= \mathbf{W}^k \mathbf{x} \\ V &= \mathbf{W}^v \mathbf{x} \end{aligned}$$

$$\text{Attention}(\mathbf{x}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$$

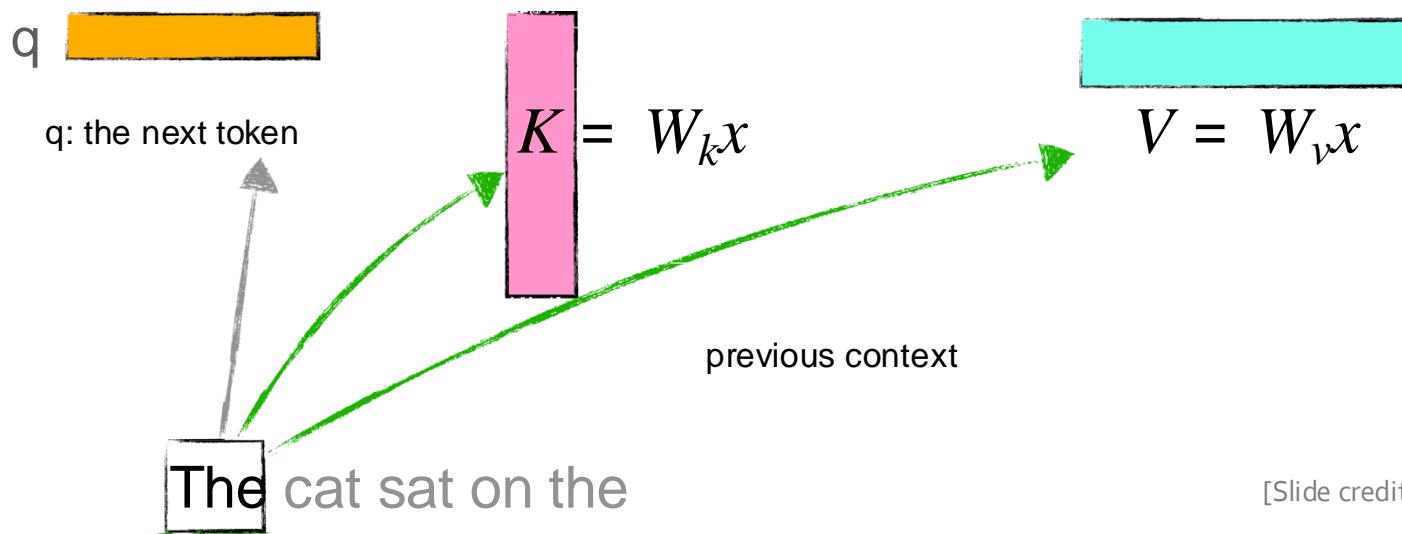
Making decoding more efficient

$$Q = W^q \mathbf{x}$$

$$K = W^k \mathbf{x}$$

$$V = W^v \mathbf{x}$$

$$\text{Attention}(\mathbf{x}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$



[Slide credit: Arman Cohan]

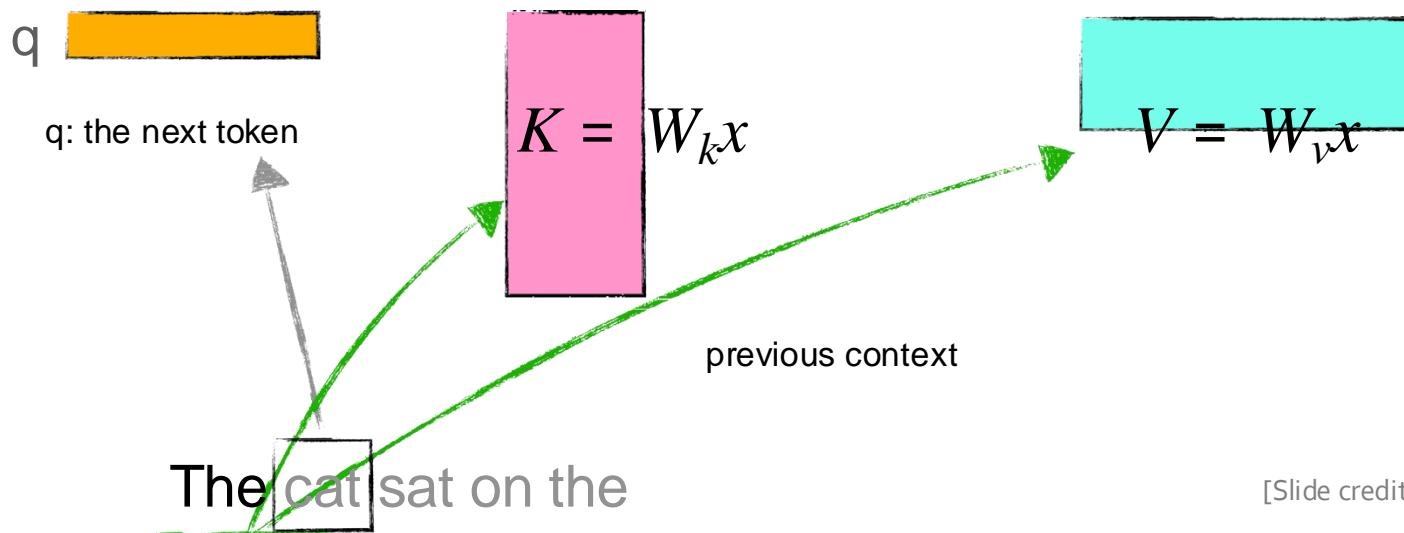
Making decoding more efficient

$$Q = W^q x$$

$$K = W^k x$$

$$V = W^v x$$

$$\text{Attention}(x) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$



[Slide credit: Arman Cohan]

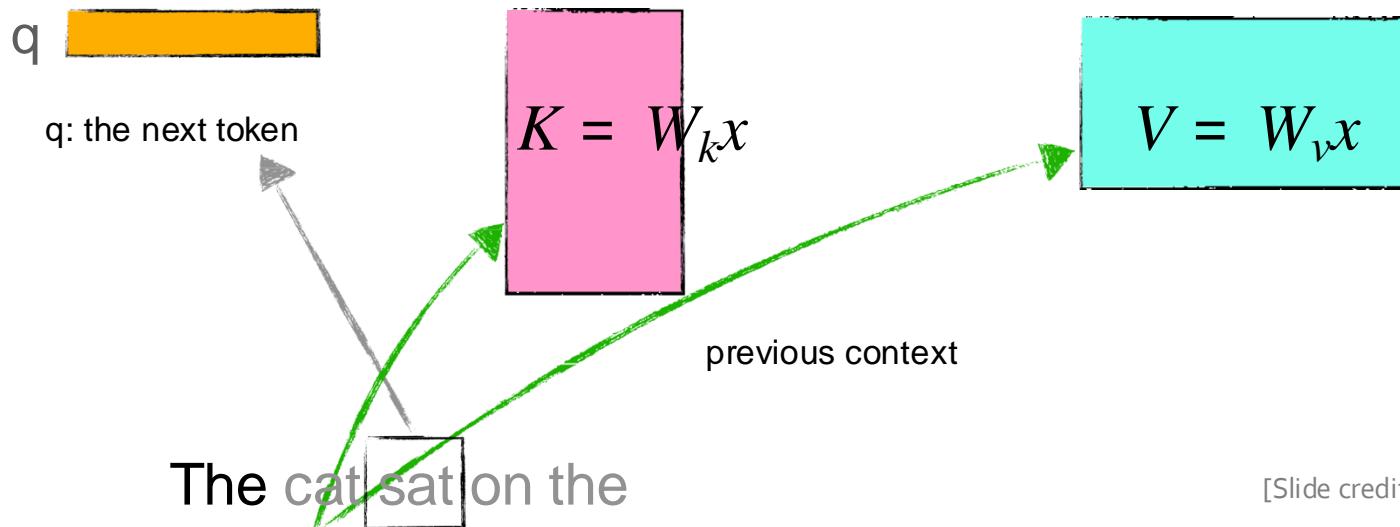
Making decoding more efficient

$$Q = W^q x$$

$$K = W^k x$$

$$V = W^v x$$

$$\text{Attention}(x) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$



[Slide credit: Arman Cohan]

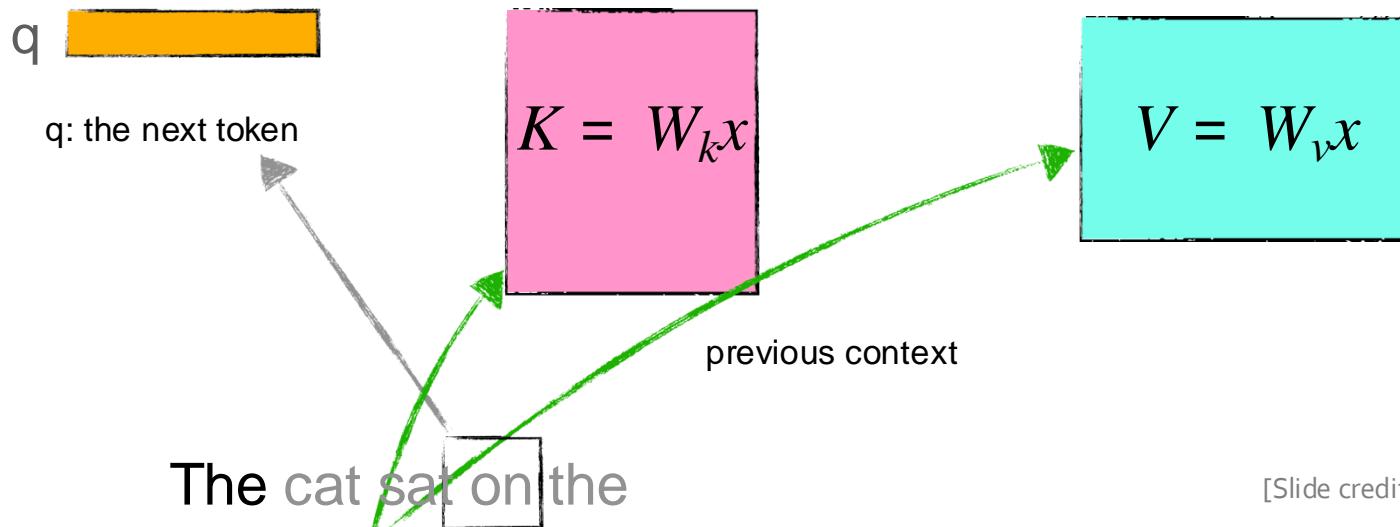
Making decoding more efficient

$$Q = W^q \mathbf{x}$$

$$K = W^k \mathbf{x}$$

$$V = W^v \mathbf{x}$$

$$\text{Attention}(\mathbf{x}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$



[Slide credit: Arman Cohan]

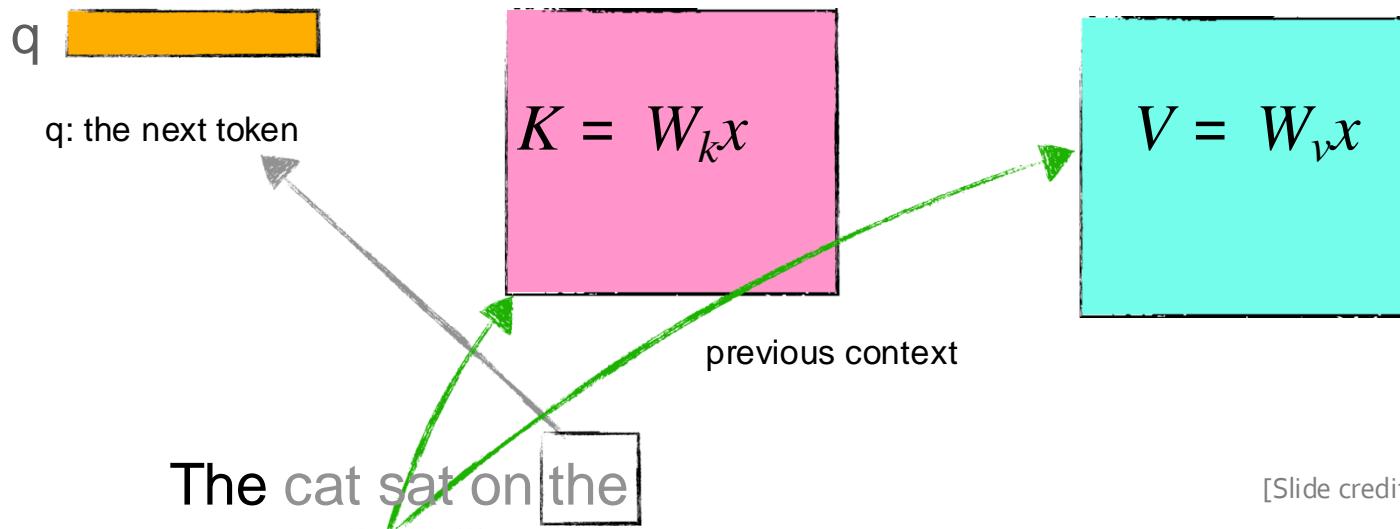
Making decoding more efficient

$$Q = \mathbf{W}^q \mathbf{x}$$

$$K = \mathbf{W}^k \mathbf{x}$$

$$V = \mathbf{W}^v \mathbf{x}$$

$$\text{Attention}(\mathbf{x}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$



[Slide credit: Arman Cohan]

Making decoding more efficient

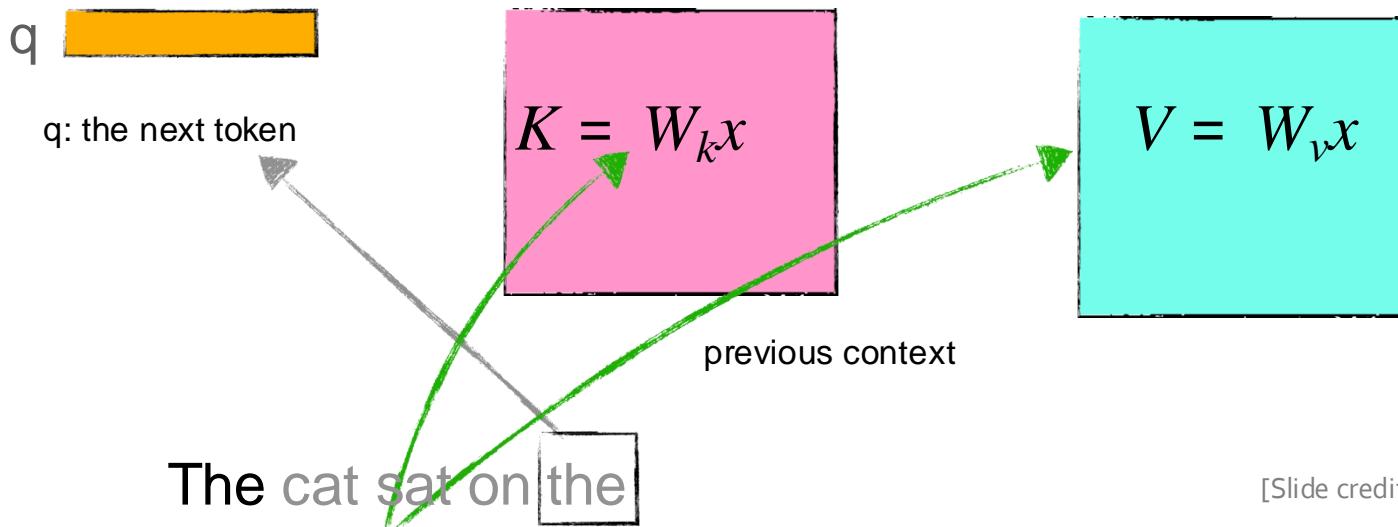
- We are computing the Keys and Values many times!
 - Let's reduce redundancy! 😊

$$Q = \mathbf{W}^q \mathbf{x}$$

$$K = \mathbf{W}^k \mathbf{x}$$

$$V = \mathbf{W}^v \mathbf{x}$$

$$\text{Attention}(\mathbf{x}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$$



[Slide credit: Arman Cohan]

Making decoding more efficient

- We are computing the Keys and Values many times!
 - Let's reduce redundancy! 😭

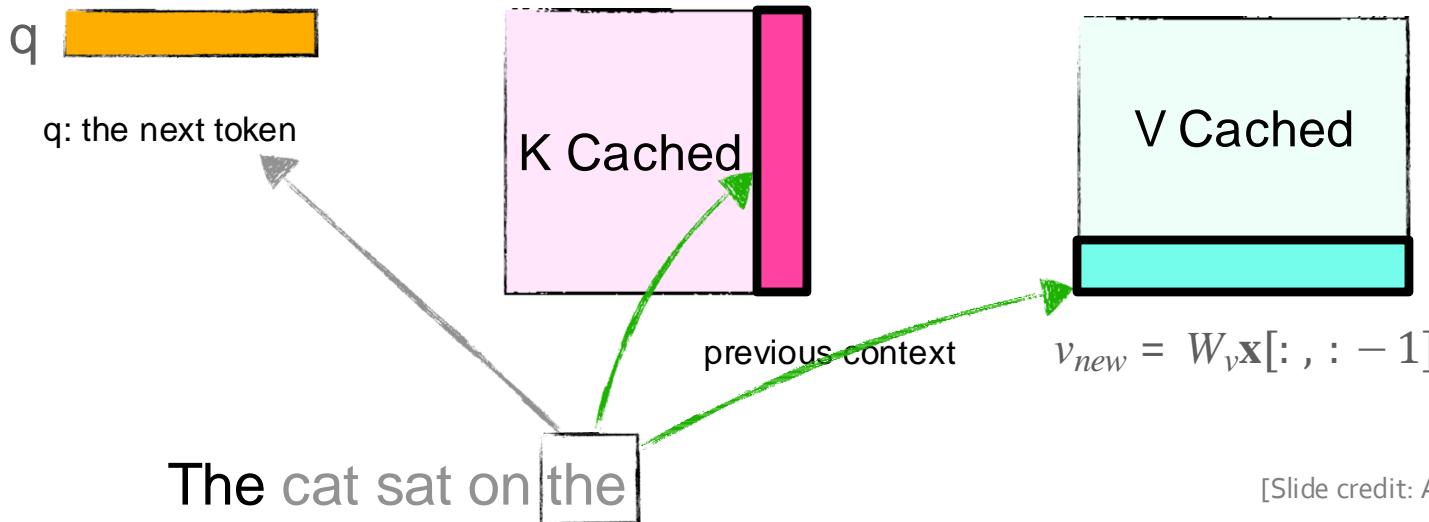
$$Q = \mathbf{W}^q \mathbf{x}$$

$$K = \mathbf{W}^k \mathbf{x}$$

$$V = \mathbf{W}^v \mathbf{x}$$

$$\text{Attention}(\mathbf{x}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$$

$$k_{new} = W_k \mathbf{x}[:, : - 1]$$



[Slide credit: Arman Cohan]

Making decoding more efficient

- **Question:** How much memory does this K, V cache require?

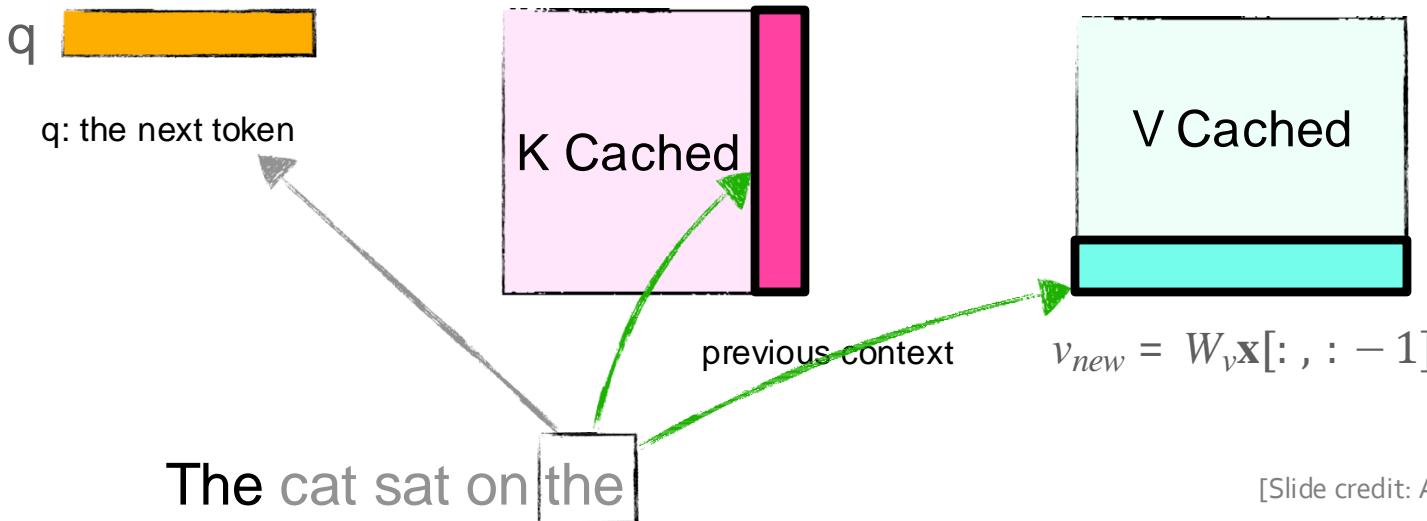
$$Q = \mathbf{W}^q \mathbf{x}$$

$$K = \mathbf{W}^k \mathbf{x}$$

$$V = \mathbf{W}^v \mathbf{x}$$

$$\text{Attention}(\mathbf{x}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$$

$$k_{new} = W_k \mathbf{x}[:, : - 1]$$



$$v_{new} = W_v \mathbf{x}[:, : - 1]$$

[Slide credit: Arman Cohan]

Summary

- This is a very generic Transformer!

- **Next:**
 - Architectural variants
 - ...

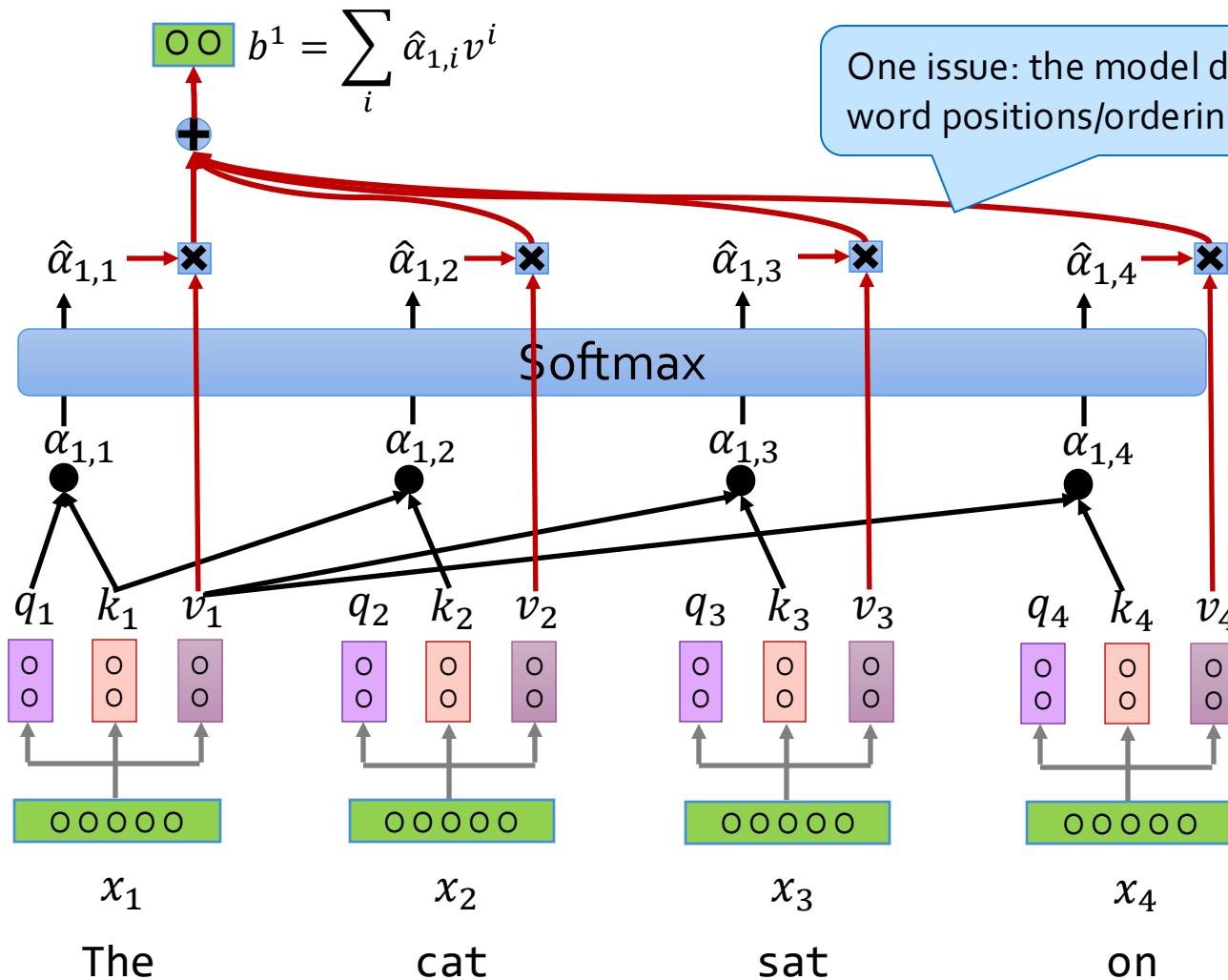


Encoding Positional Information



Why do we need positional encoding?

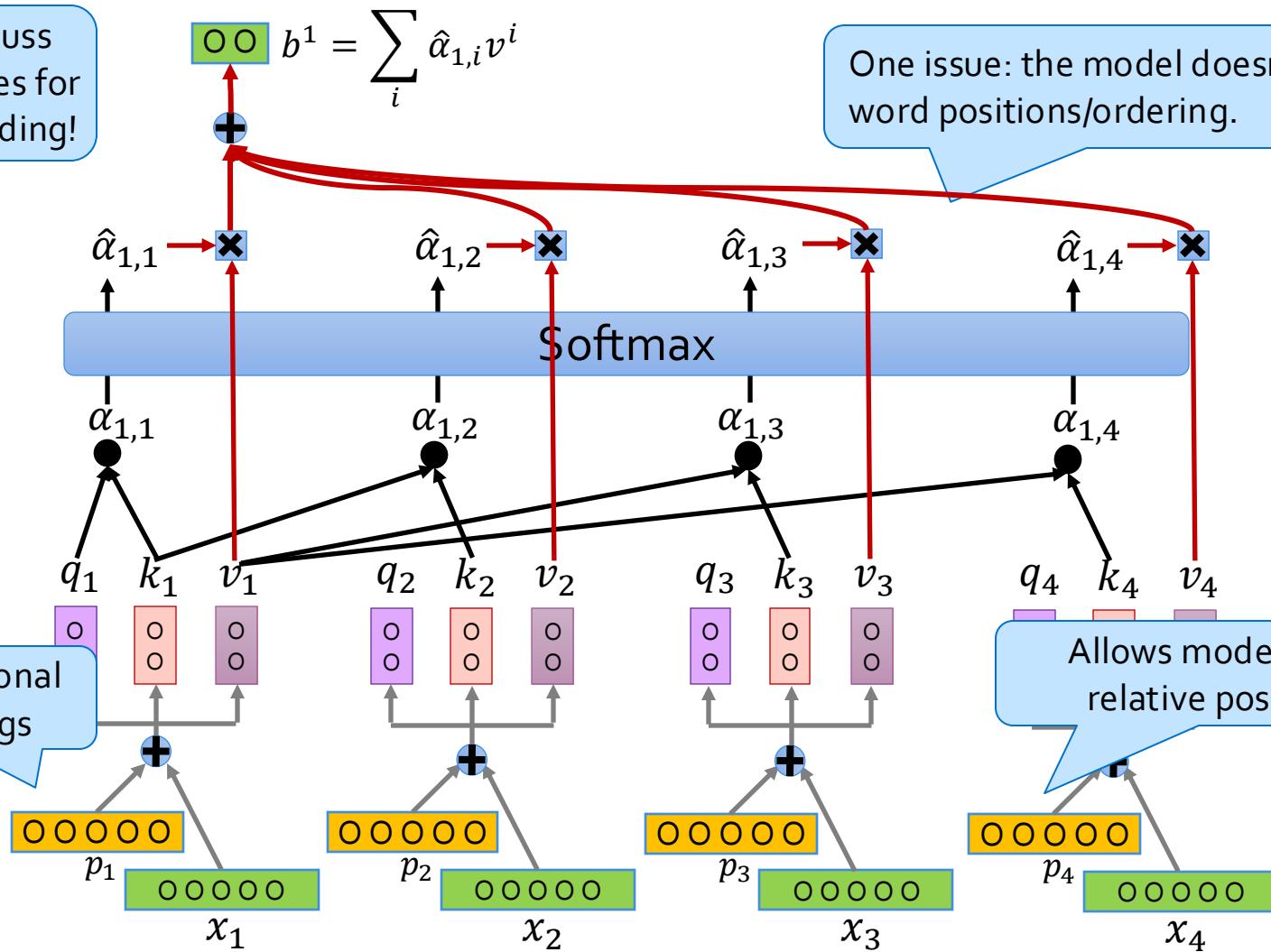




We will discuss various choices for these embedding!

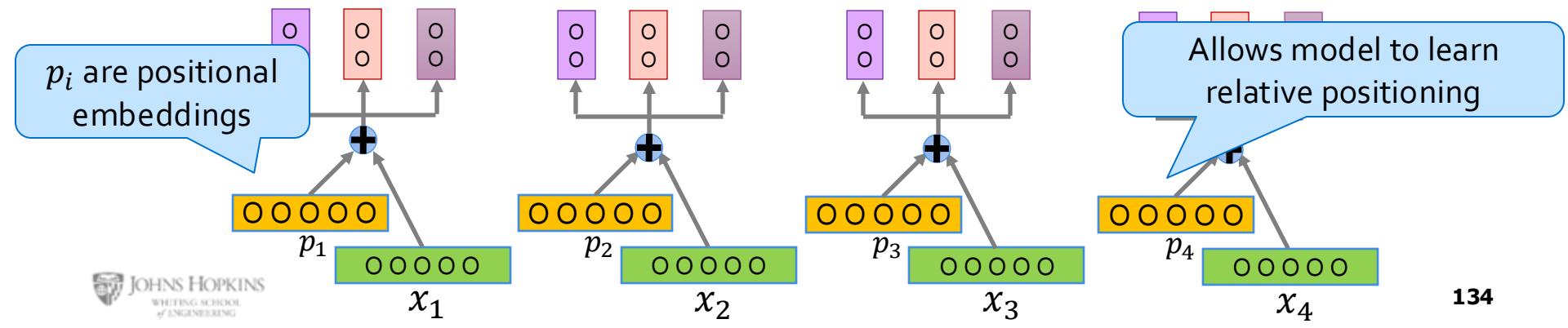
$$OO = b^1 + \sum_i \hat{\alpha}_{1,i} v^i$$

One issue: the model doesn't know word positions/ordering.



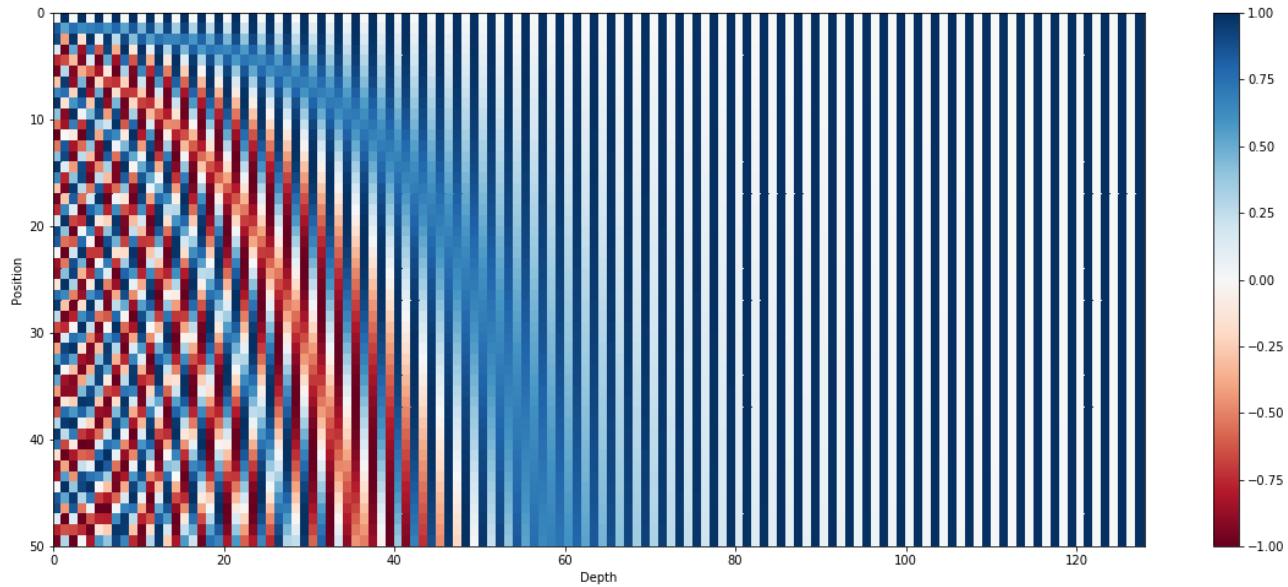
Positional Embeddings: The Flavors

- Absolute encoding: vectors that uniquely encode each position.
- Relative encoding: the positional encoding for each position is determined based on its distance from the other positions it is attending to.

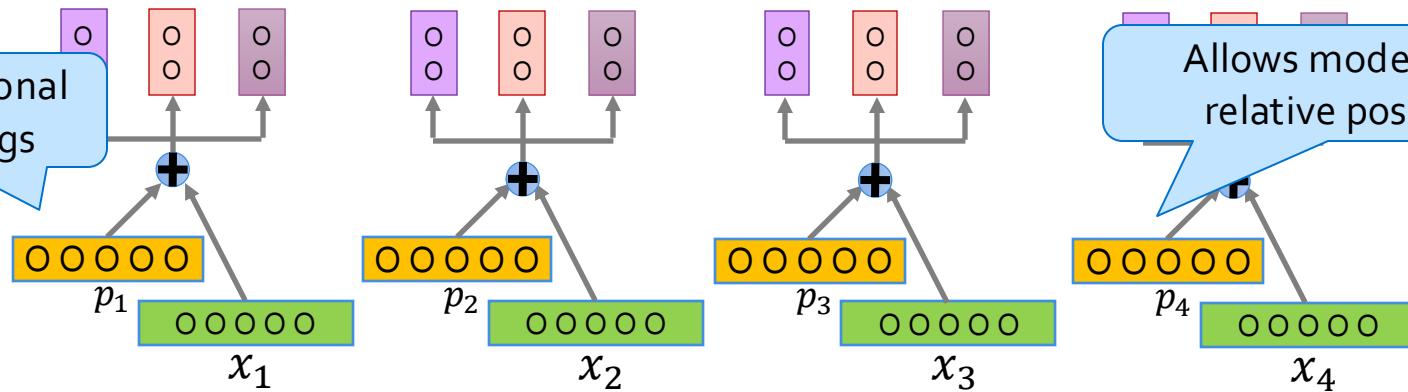


An approach:
Sine/Cosine encoding

$$p_i = \begin{cases} \sin(i/10000^{2*1/d}) \\ \cos(i/10000^{2*1/d}) \\ \vdots \\ \sin(i/10000^{2*\frac{d}{2}/d}) \\ \cos(i/10000^{2*\frac{d}{2}/d}) \end{cases}$$



p_i are positional embeddings



Limits of Absolute Positional Encoding

- We can have fixed positional embeddings for each index training position (e.g., 1, 2, 3, ... 1000).
 - What happens if we get a sequence with 5000 words at test time?
- We want something that can generalize to arbitrary sequence lengths.
 - Approach: encoding the relative positions, for example based on the distance of the tokens in a local window to the current token.

Relative Positional Encoding

- You can rewrite the statement from the previous slide in the following form:

$$QK_{ij} = (W_q[x_i + p_i])^T (W_k[x_j + p_j]) = x_i^T W_q^T W_k x_j + P_{ij}$$

- Note, the values of P_{ij} encode the relative of i and j .
- How should we construct P_{ij} ?

How much attention
should position i should
attend to position j

Relative Positional Encoding

- There have been various choices:
 - T5 models simplify this into learnable relative embeddings \mathbf{P}_{ij} such that:
$$QK_{ij} = \mathbf{x}_i^T \mathbf{W}_q^T \mathbf{W}_k \mathbf{x}_j + \mathbf{P}_{ij}$$
 - DeBERTa learns relative positional embeddings $\tilde{\mathbf{p}}_{i-j}$ such that:
$$QK_{ij} = \mathbf{x}_i^T \mathbf{W}_q^T \mathbf{W}_k \mathbf{x}_j + \mathbf{x}_i^T \mathbf{W}_q^T \mathbf{W}_k \tilde{\mathbf{p}}_{i-j} + \tilde{\mathbf{p}}_{i-j}^T \mathbf{W}_q^T \mathbf{W}_k \mathbf{x}_j$$
 - Transformer-XL learns relative positional embeddings $\tilde{\mathbf{p}}_{i-j}$ and trainable vectors \mathbf{u}, \mathbf{v} s.t.:
$$QK_{ij} = \mathbf{x}_i^T \mathbf{W}_q^T \mathbf{W}_k \mathbf{x}_j + \mathbf{x}_i^T \mathbf{W}_q^T \mathbf{W}_k \tilde{\mathbf{p}}_{i-j} + \mathbf{u}^T \mathbf{W}_q^T \mathbf{W}_k \mathbf{x}_j + \mathbf{v}^T \mathbf{W}_q^T \mathbf{W}_k \tilde{\mathbf{p}}_{i-j}$$
 - ALiBi learns a scalar m such that:
$$QK_{ij} = \mathbf{x}_i^T \mathbf{W}_q^T \mathbf{W}_k \mathbf{x}_j - m |i - j|$$

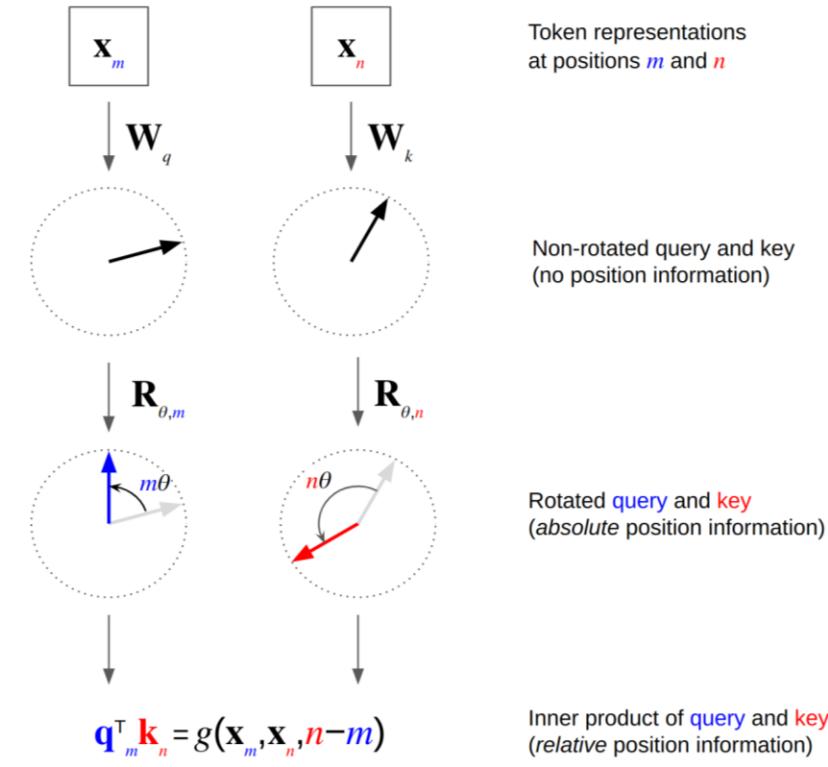
Relative Encoding via Multiplication: Rotary Positional Encoding (RoPE)

- Drop the additive positional encoding and make it multiplicative.

$$\begin{aligned} qk_{mn} &= (R_{\theta,m}W_qx_m)^T(R_{\theta,n}W_kx_n) \\ &= x_m^TW_q^TR_{\theta,m}^TR_{\theta,n}W_kx_n \end{aligned}$$

- θ : the size of rotation
- $R_{\theta,m}$: rotation matrix, rotates a vector it gets multiplied to proportional to θ and the position index m .

- Intuition: **nearby** words have **smaller relative rotation**.



Summary

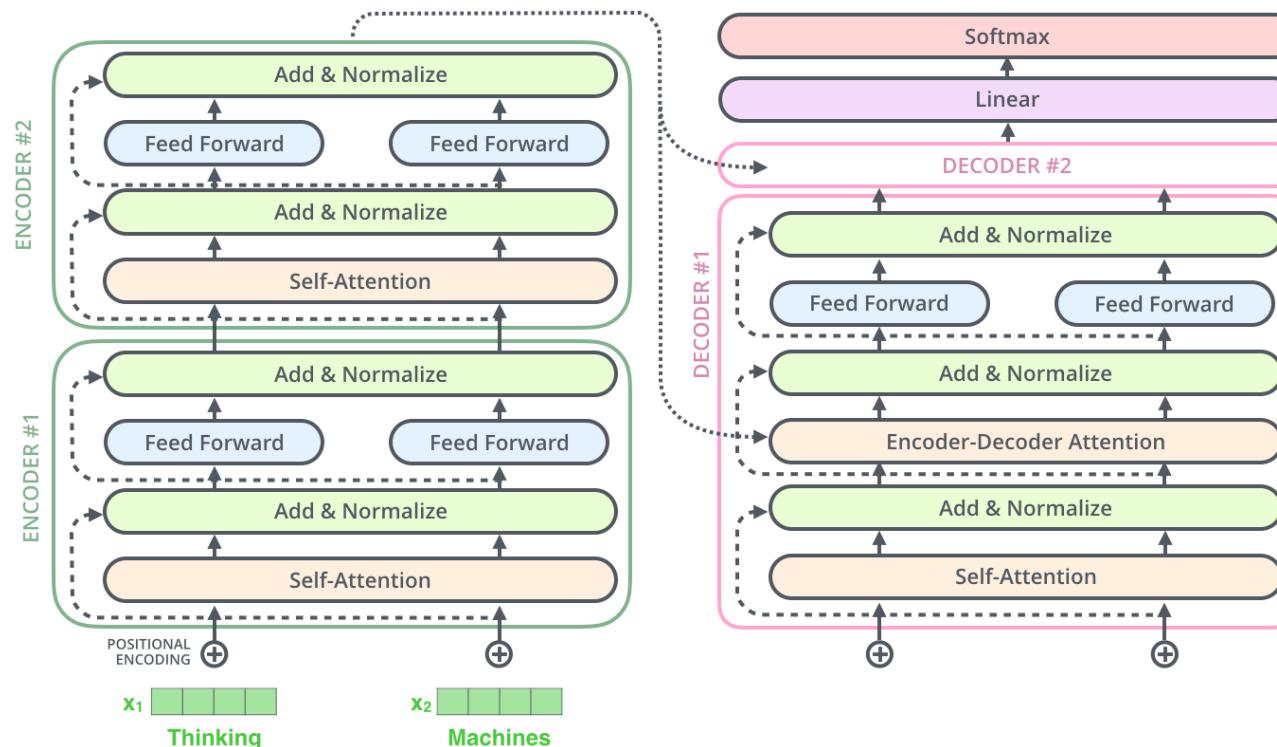
- Encoding positional information in language models is a non-trivial problem.
 - We discussed various proposal: learned, absolute, relative encoding, NoPos, etc.
- This is an important literature related to the length generalization of Transformers.
- This is an active research area and likely to change in the coming years.



Transformer Architectural Variants

Transformer [Vaswani et al. 2017]

- An encoder-decoder architecture built with attention modules.



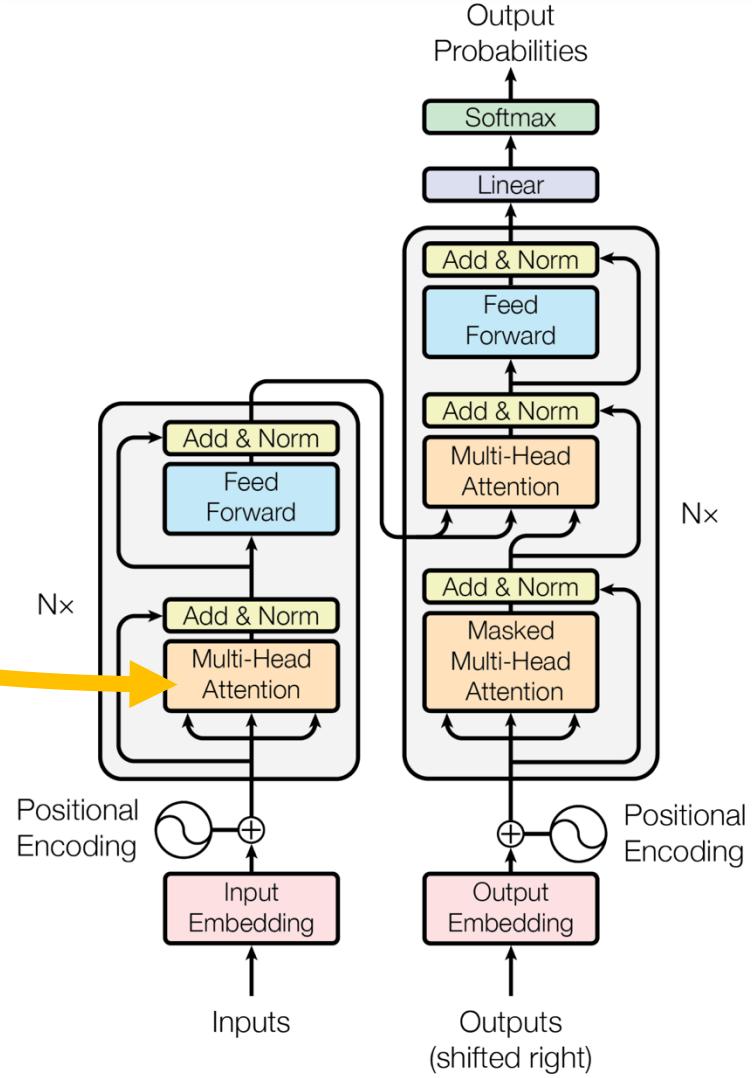
Transformer

[Vaswani et al. 2017]

- Computation of **encoder** attends to both sides.



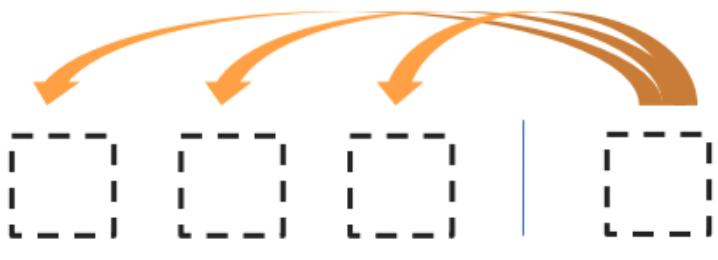
Encoder Self-Attention



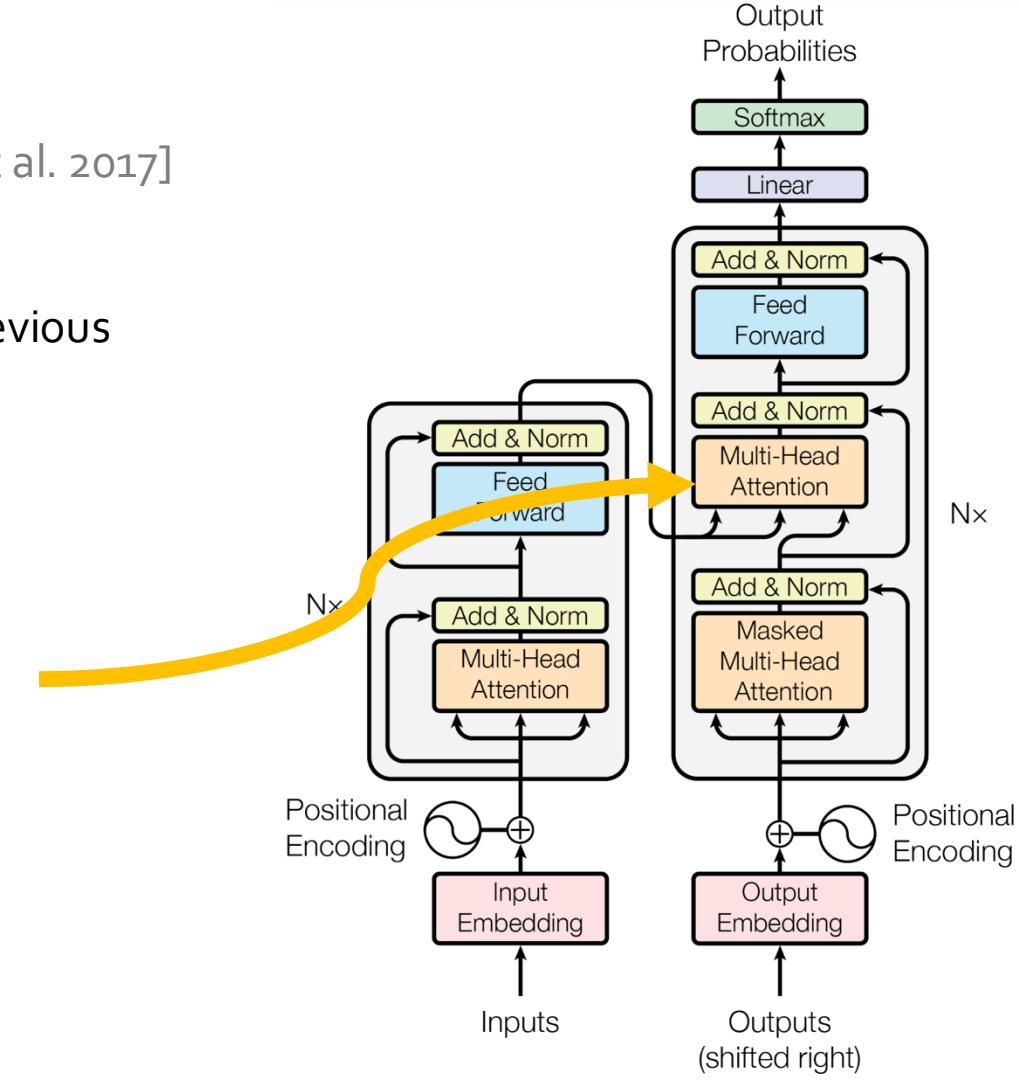
Transformer

[Vaswani et al. 2017]

- At any step of **decoder**, it attends to previous computation of **encoder**



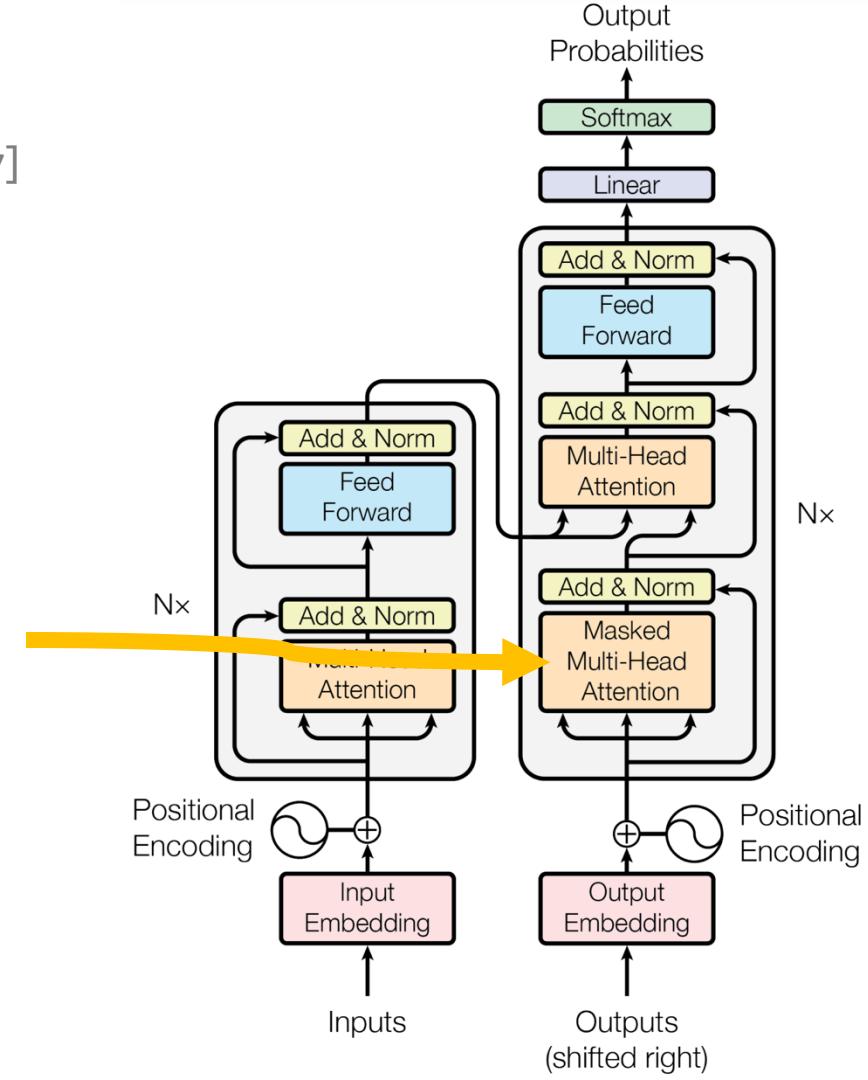
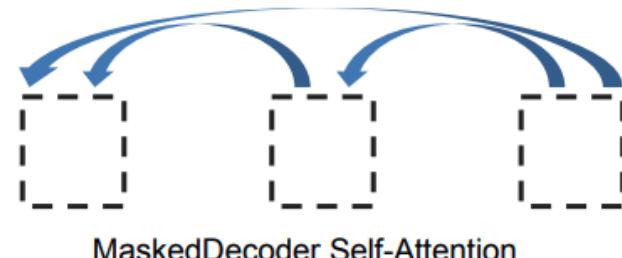
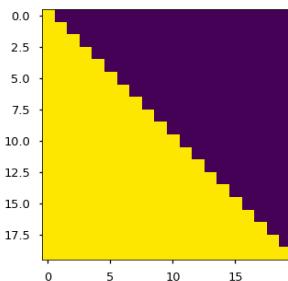
Encoder-Decoder Attention



Transformer

[Vaswani et al. 2017]

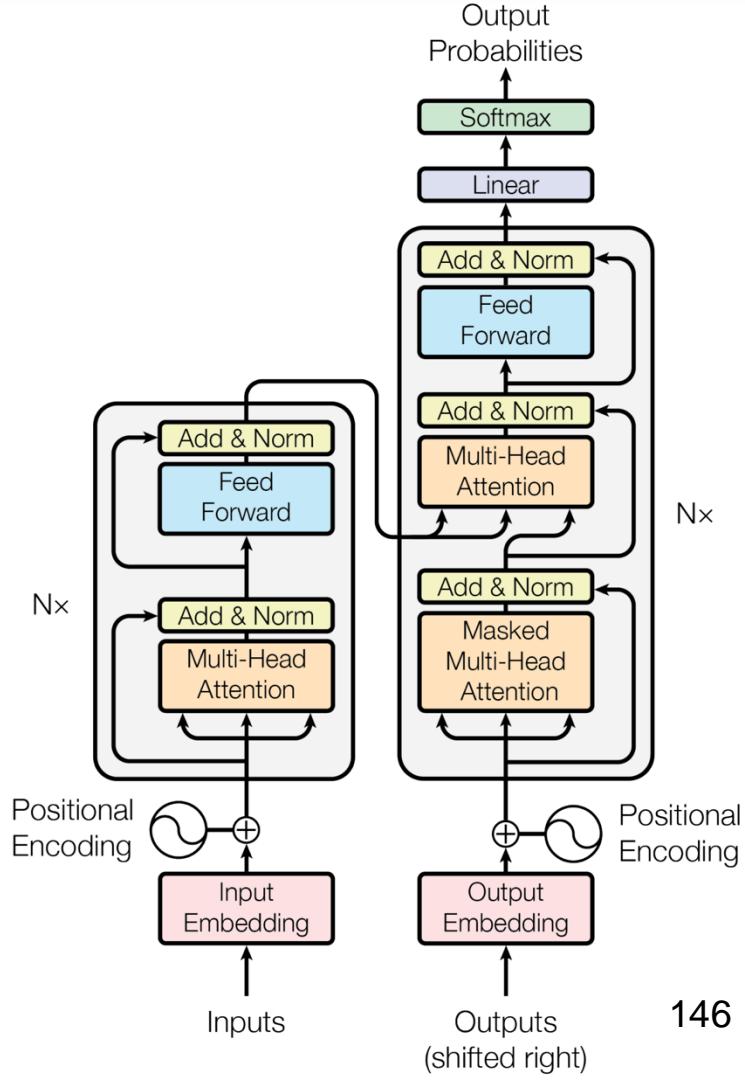
- At any step of **decoder**, it attends to previous computation of **encoder** as well as **decoder's** own generations



Transformer

[Vaswani et al. 2017]

- At any step of **decoder**, it attends to previous computation of **encoder** as well as **decoder's** own generations
- At any step of **decoder**, **re-use** previous computation of **encoder**.
- Computation of **decoder** is **linear**, instead of quadratic.



Quiz: Enc-Dec Cost

- Goal: We are building an encoder-decoder Transformer for summarizing passages to summaries.
- For a passage of length N and a summary of length M , the complexity of the attention is:
 - $O(N) + O(M)$
 - $O(N) + O(M) + O(NM)$
 - $O(N^2) + O(M^2) + O(NM)$
 - $O(N^2) + O(M^2)$

Quiz: Enc-Dec Cost

- Goal: We are building an encoder-decoder Transformer for summarizing passages to summaries.
- For a passage of length N and a summary of length M, the complexity of the attention is:
 - $O(N) + O(M)$
 - $O(N) + O(M) + O(NM)$
 - $O(N^2) + O(M^2) + O(NM)$
 - $O(N^2) + O(M^2)$

No, self attention is all-to-all
and so quadratic.

Quiz: Enc-Dec Cost

- Goal: We are building an encoder-decoder Transformer for summarizing passages to summaries.
- For a passage of length N and a summary of length M , the complexity of the attention is:
 - $O(N) + O(M)$
 - $O(N) + O(M) + O(NM)$
 - $O(N^2) + O(M^2) + O(NM)$
 - $O(N^2) + O(M^2)$

No, self attention is all-to-all
and so quadratic in M and N .

Quiz: Enc-Dec Cost

- Goal: We are building an encoder-decoder Transformer for summarizing passages to summaries.
- For a passage of length N and a summary of length M , the complexity of the attention is:
 - $O(N) + O(M)$
 - $O(N) + O(M) + O(NM)$
 - $O(N^2) + O(M^2) + O(NM)$
 - $O(N^2) + O(M^2)$

No, self attention is all-to-all
and so quadratic in M and N .

Quiz: Enc-Dec Cost

- Goal: We are building an encoder-decoder Transformer for summarizing passages to summaries.
- For a passage of length N and a summary of length M, the complexity of the attention is:
 - $O(N) + O(M)$
 - $O(N) + O(M) + O(NM)$
 - $O(N^2) + O(M^2) + O(NM)$
 - $O(N^2) + O(M^2)$

No, cross attention is missing.

Quiz: Enc-Dec Cost

- Goal: We are building an encoder-decoder Transformer for summarizing passages to summaries.
- For a passage of length N and a summary of length M, the complexity of the attention is:
 - $O(N) + O(M)$
 - $O(N) + O(M) + O(NM)$
 - $O(N^2) + O(M^2) + O(NM)$
 - $O(N^2) + O(M^2)$

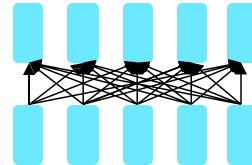
Yes. The three terms are respectively the Encoder self-attention, Decoder self-attention, and Cross attention.

After Transformer ...



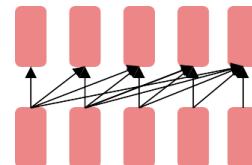
Impact of Transformers

- A building block for a variety of LMs



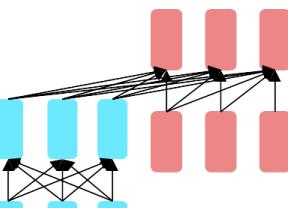
Encoders

- ❖ Examples: BERT, RoBERTa, SciBERT.
- ❖ Captures bidirectional context. Wait, how do we pretrain them?



Decoders

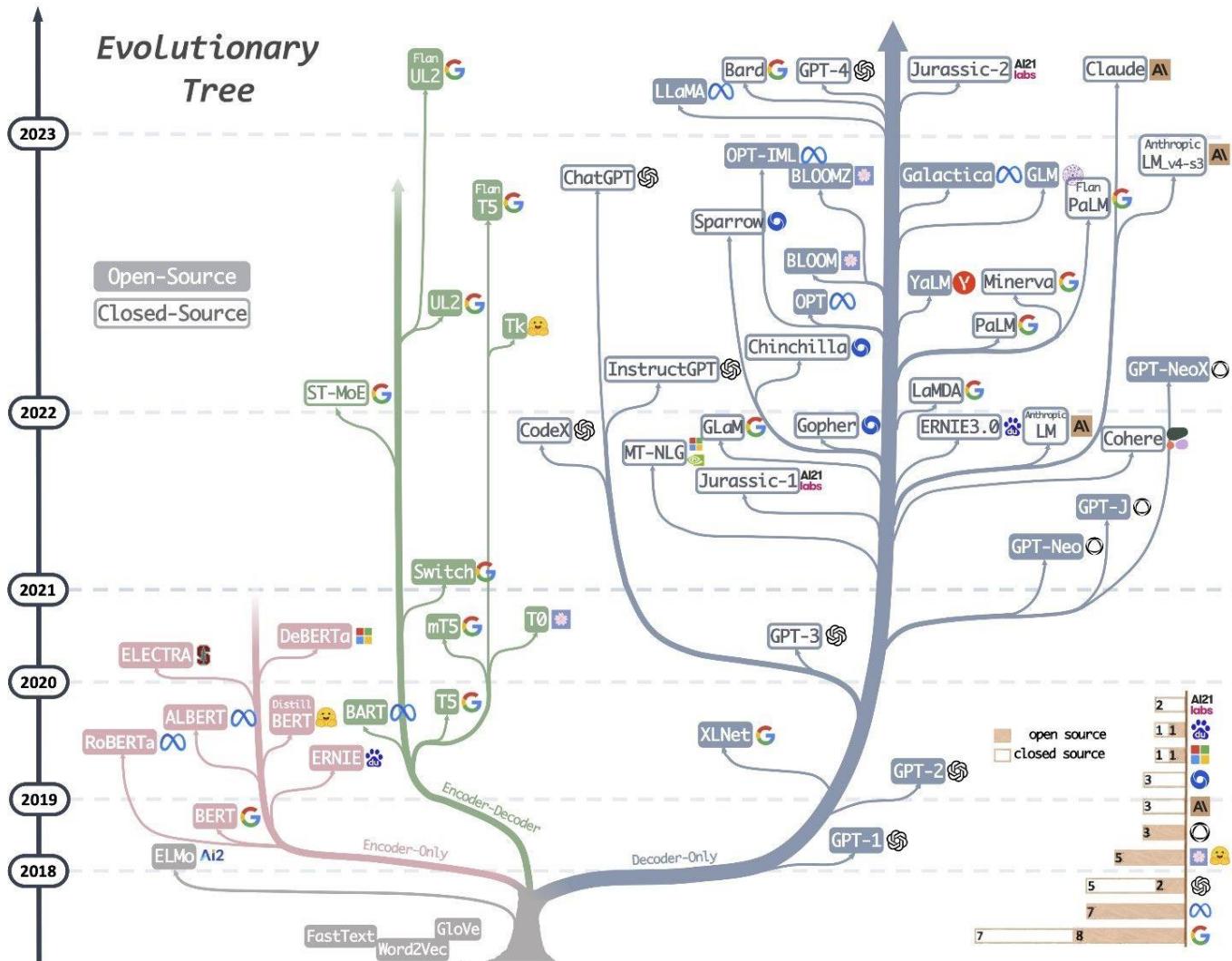
- ❖ Examples: GPT-3, GPT-4, LaMMA, Mistral
- ❖ Other name: causal or auto-regressive language model
- ❖ Nice to generate from; can't condition on future words



Encoder-
Decoders

- ❖ Examples: T5, Meena
- ❖ What's the best way to pretrain them?

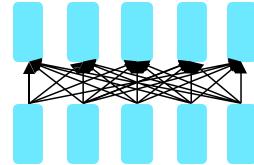
Evolutionary Tree



Yang et al. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond, 2023

Few Notable Models from History Books

Encoder-only Family of Transformers



Probing BERT Masked LM

- BERT is trained to uncover masked tokens.

Paris is the [MASK] of France.

Compute

Computation time on cpu: cached

capital	0.997
heart	0.001
center	0.000
centre	0.000
city	0.000

</> JSON Output

Maximize

Probing BERT Masked LM

- BERT is trained to uncover masked tokens.

Today is Tuesday, so tomorrow is [MASK].

Compute

Computation time on cpu: cached



</> JSON Output

Maximize

Impact of BERT

- In order to have state-of-the-art performance on different tasks, there is no need for coming up with a novel model architecture
 - End of task-specific model architecture engineering
- An early sign that larger scales and self-supervised learning (language modeling) are the key for future performance improvements

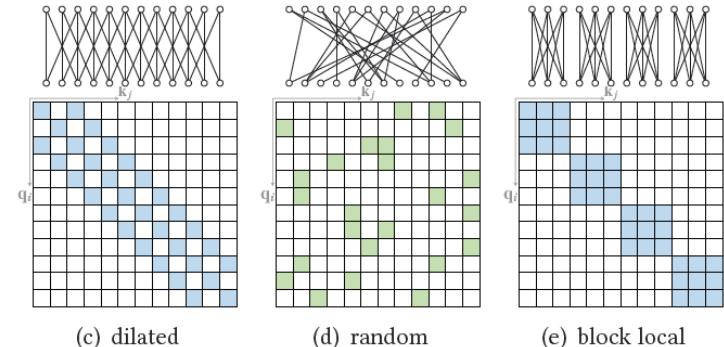
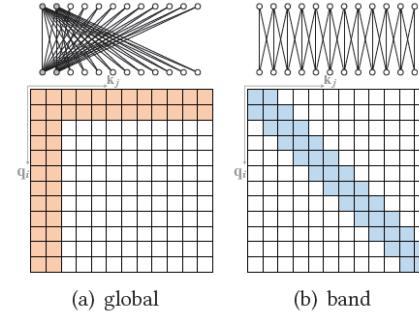
What Happened After BERT?

- RoBERTa (Liu et al., 2019)
 - Exact same architecture as BERT
 - Drops the next sentence prediction loss!
 - Trained on 10x data (the original BERT was actually under-trained)
 - Much stronger performance than BERT (e.g., 94.6 vs 90.9 on SQuAD)

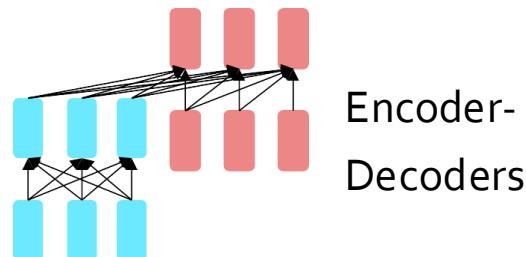
Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	94.6/89.4	90.2	96.4
BERT _{LARGE}						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7

What Happened After BERT?

- Models that handle long contexts
 - Longformer, Big Bird, ...
- Multilingual BERT
 - Trained single model on 104 languages from Wikipedia.
- BERT extended to different domains
 - SciBERT, BioBERT, FinBERT, ClinicalBERT, ...
- Making BERT smaller to use
 - DistillBERT, TinyBERT, ...

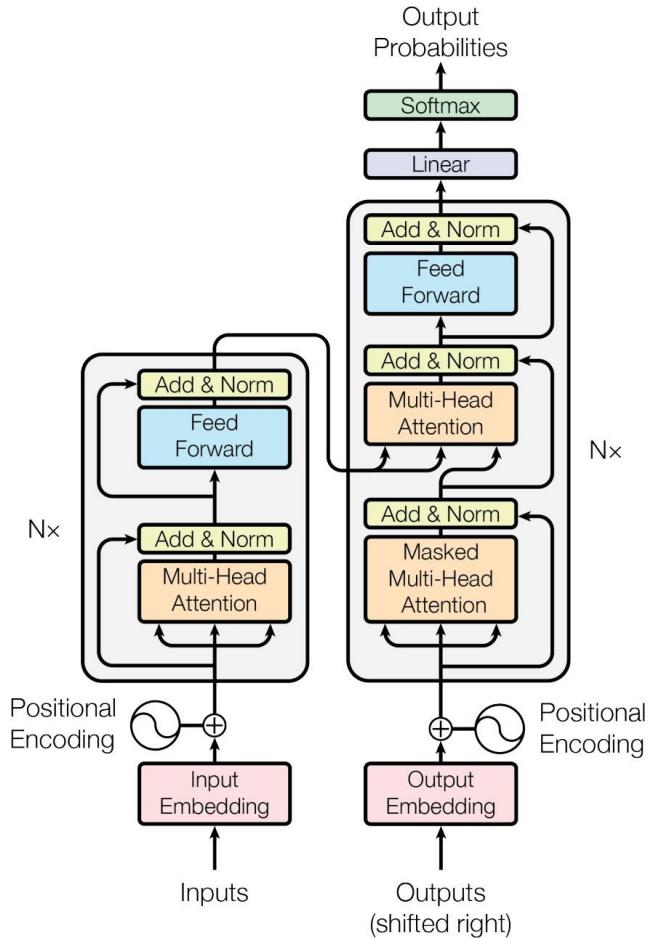


Encoder-Decoder Family of Transformers



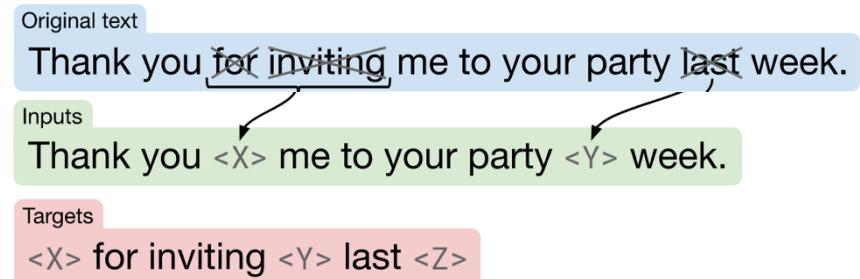
Encoder-decoder Models

- The original transformer architecture was encoder decoder
- Encoder-decoder models are flexible in both generation and classification tasks
- How can we pretrain an encoder-decoder model like BERT to be a good general language pretrained LM?



T5: Text-To-Text Transfer Transformer

- An encoder-decoder architecture
- Pre-training objective:
corrupt and reconstruct objective

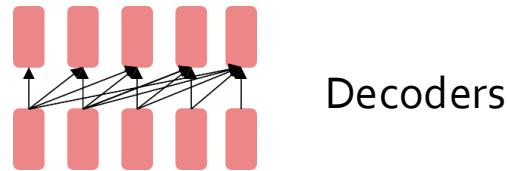


Model	Parameters	No. of layers	d_{model}	d_{ff}	d_{kv}	No. of heads
Small	60M	6	512	2048	64	8
Base	220M	12	768	3072	64	12
Large	770M	24	1024	4096	64	16
3B	3B	24	1024	16384	128	32
11B	11B	24	1024	65536	128	128

- The original paper is an excellent set of in-depth analysis of various parameters of model design. We discuss some of these results in other places.

<https://huggingface.co/t5-base>

Decoder-only Family of Transformers



GPT

Generative Pre-trained Transformer

GPT-2: A Big Language Model (2019)

Language Models are Unsupervised Multitask Learners

Alec Radford ^{*1} Jeffrey Wu ^{*1} Rewon Child ¹ David Luan ¹ Dario Amodei ^{**1} Ilya Sutskever ^{**1}

GPT: An Auto-Regressive LM (2018)

Improving Language Understanding
by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

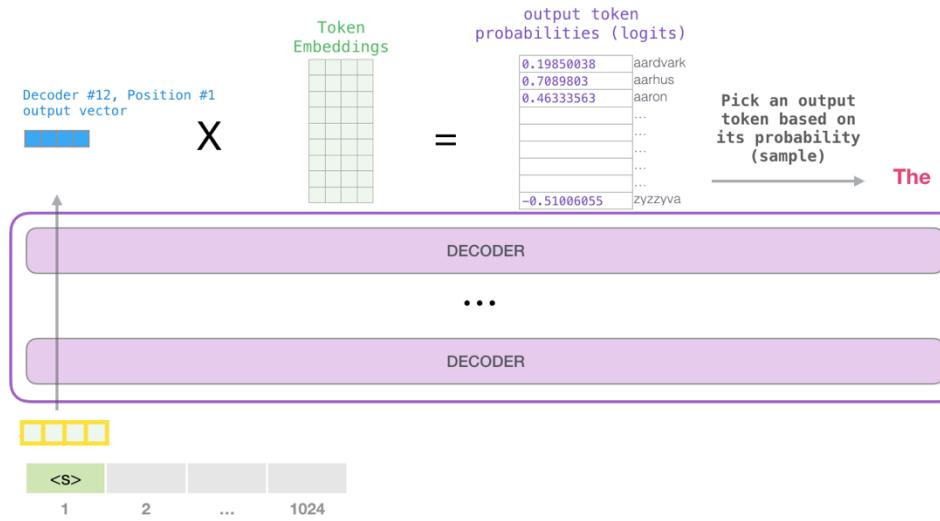
Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com

GPT-2

- GPT-2 uses only Transformer Decoders (no Encoders) to generate new sequences from scratch or from a starting sequence
- As it processes each subword, it masks the “future” words and conditions on and attends to the previous words



GPT-3: Just Scale

- More layers & parameters
- Bigger dataset
- Longer training
- Larger embedding/hidden dimension
- Larger context window

GPT-3 = A **very big** GPT-2



GPT4

- Transformer-based
 - The rest is mystery! 😊
 - If we're going based on costs, GPT4 is ~15-30 times costlier than GPT3. That should give you an idea how its likely size!
- Note, these language models involve more than just pre-training.
 - Pre-training provides the foundation based on which we build the model.
 - We will discuss the later stages (post hoc alignment) in a 2-3 weeks.

Model	Usage	
davinci-002	\$0.0020 / 1K tokens	
Model	Input	Output
gpt-4	\$0.03 / 1K tokens	\$0.06 / 1K tokens

Other Available [Decoder] LMs

EleutherAI: GPT-Neo (6.7B), GPT-J (6B), GPT-NeoX (20B)

<https://huggingface.co/EleutherAI>

<https://6b.eleuther.ai/>

LLaMA, 65B: <https://github.com/facebookresearch/llama>

Mistral and Mixtral:

<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

<https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1>

LMSys ChatArena

<https://lmarena.ai/>

Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	ChatGPT-4o-latest...(2024-08-08)	1317	+5/-5	20805	OpenAI	Proprietary	2023/10
2	Gemini-1.5-Pro-Exp-0801	1298	+4/-4	23232	Google	Proprietary	2023/11
2	Grok-2.08-13	1293	+7/-6	6686	xAI	Proprietary	2024/3
3	GPT-4o-2024-05-13	1286	+3/-3	80741	OpenAI	Proprietary	2023/10
5	GPT-4o-mini-2024-07-18	1275	+5/-4	21621	OpenAI	Proprietary	2023/10
5	Claude-3.5-Sonnet	1271	+3/-3	51097	Anthropic	Proprietary	2024/4
5	Grok-2-Mini-08-13	1268	+7/-7	7266	xAI	Proprietary	2024/3
6	Gemini_Advanced_App_(2024-05-14)	1267	+4/-3	52136	Google	Proprietary	Online
6	Meta_llama-3.1-405b-Instruct	1266	+4/-4	22312	Meta	Llama 3.1 Community	2023/12
7	GPT-4o-2024-08-06	1262	+5/-5	13703	OpenAI	Proprietary	2023/10
8	Gemini-1.5-Pro-001	1260	+3/-2	72623	Google	Proprietary	2023/11
10	Gemini-1.5-Pro-Preview-0409	1257	+3/-3	55604	Google	Proprietary	2023/11
10	GPT-4-Turbo-2024-04-09	1257	+2/-3	86648	OpenAI	Proprietary	2023/12
12	Mistral-Large-2407	1250	+5/-5	14793	Mistral	Mistral Research	2024/7
12	Athene-70b	1250	+5/-5	13655	NexusFlow	CC-BY-NC-4.0	2024/7
14	GPT-4-1106-preview	1251	+3/-3	93540	OpenAI	Proprietary	2023/4

Training Transformer LMs: Empirical Considerations

Pre-training Transformer LMs

- There is so much empirical knowledge/experiences that goes into training these models.
- Various empirical issues about:
 - Preparation/pre-processing data
 - Efficient training of models
 - ...

Data Cleaning: E

Remove any:

- References to Javascript
- “Lorem ipsum” text — placeholder text commonly used to demonstrate the visual form of a document

Menu

Lemon

Introduction

The lemon, Citrus Limon (L.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae. The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses. The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Article

The origin of the le

vn, though

Retain:

- Sentences with terminal punctuation marks
- Pages with at least 5 sentences, sentences with at least 3 words

Please enable JavaScript to use our site.

Home

Products

Shipping

Contact

FAQ

Dried Lemons, \$3.59/pound

Organic dried lemons from our farm in California.

Lemons are harvested and sun-dried for maximum flavor.

Good in soups and on popcorn.

The lemon, Citrus Limon (L.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae.

The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses. The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Curabitur in tempus quam. In mollis et ante at consectetur.

Aliquam erat volutpat.

Donec at lacinia est.

Duis semper, magna tempor interdum suscipit, ante elit molestie urna, eget efficitur risus nunc ac elit.

Fusce quis blandit lectus.

Mauris at mauris a turpis tristique lacinia at nec ante.

Aenean in scelerisque tellus, a efficitur ipsum.

Integer justo enim, ornare vitae sem non, mollis fermentum lectus.

Mauris ultrices nisl at libero porta sodales in ac orci.

```
function Ball(r) {  
    this.radius = r;  
    this.area = pi * r ** 2;  
    this.show = function(){  
        drawCircle(r);  
    }  
}
```

Pre-training Data: Experiment

- Takeaway:
 - Clean and compact data is better than large, but noisy data.
 - Pre-training on in-domain data helps.

Data set	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	19.24	80.88	71.36	26.98	39.82	27.65
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21

Pre-training Data Duplicates

- There is a non-negligible number of duplicates in any pre-training data.

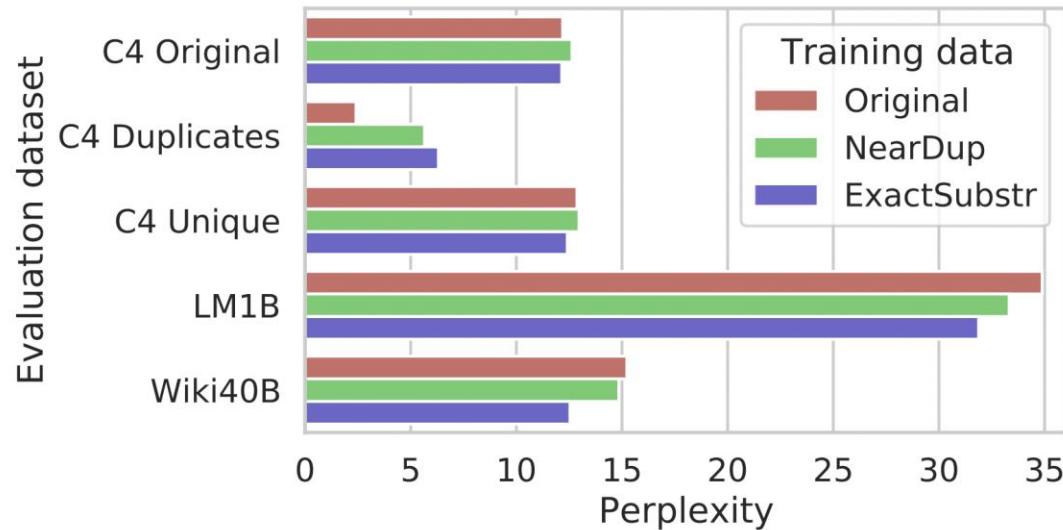
	% train examples with dup in train	% valid with dup in train	% valid with dup in valid
C4	3.04%	1.59%	4.60%
RealNews	13.63%	1.25%	14.35%
LM1B	4.86%	0.07%	4.92%
Wiki40B	0.39%	0.26%	0.72%

Dataset	Example	Near-Duplicate Example
Wiki-40B	\n_START_ARTICLE_\nHum Award for Most Impactful Character\n\n_START_SECTION_\nWinners and nominees\n\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees.\n[...]	\n_Start_ARTICLE_\nHum Award for Best Actor in a Negative Role\n\n_Start_SECTION_\nWinners and nominees\n\n_Start_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]
LM1B	I left for California in 1979 and tracked Cleveland 's changes on trips back to visit my sisters .	I left for California in 1979 , and tracked Cleveland 's changes on trips back to visit my sisters .
C4	Affordable and convenient holiday flights take off from your departure country, "Canada". From May 2019 to October 2019, Condor flights to your dream destination will be roughly 6 a week! Book your Halifax (YHZ) - Basel (BSL) flight now, and look forward to your "Switzerland" destination!	Affordable and convenient holiday flights take off from your departure country, "USA". From April 2019 to October 2019, Condor flights to your dream destination will be roughly 7 a week! Book your Maui Kahului (OGG) - Dubrovnik (DBV) flight now, and look forward to your "Croatia" destination!

Deduplicating Data Improves LMs

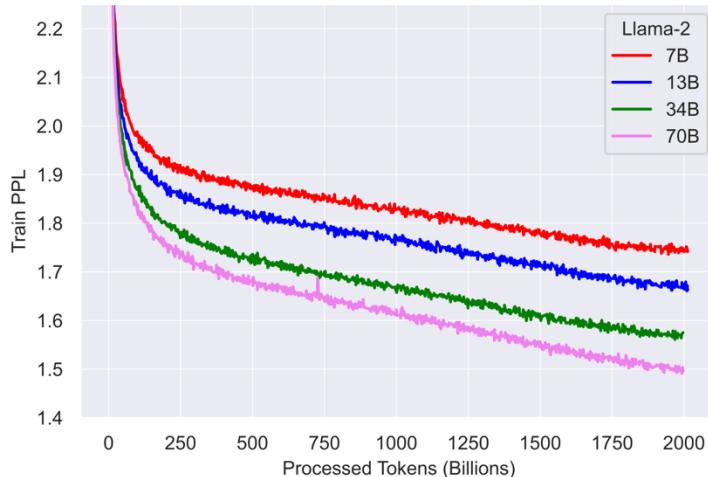
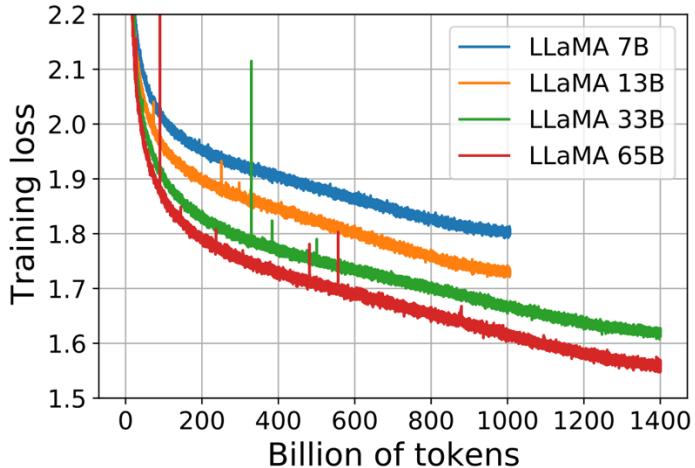
- C4 : the original training data
- C4-NearDup: C4 excluding exact duplicates
- C4-ExactSubs: C4 excluding near-duplicates

Training on deduplicated data almost always leads to lower PPL!



Convergence

- In practice, your model's loss should continue to go down with more training on more data.
- So, the real bottlenecks are:
 - (1) compute
 - (2) data
- Sometimes training diverges (spikes in the loss), at which point practitioners usually restart training from an earlier checkpoint.

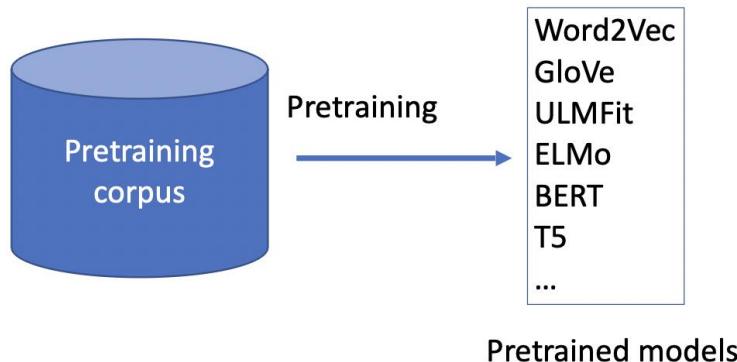


Summary

- There is much empirical knowledge that goes into engineering LMs.
- Here we covered basic topics about data and architecture engineering.
- Various topics are forthcoming: scaling laws, efficient training, etc.

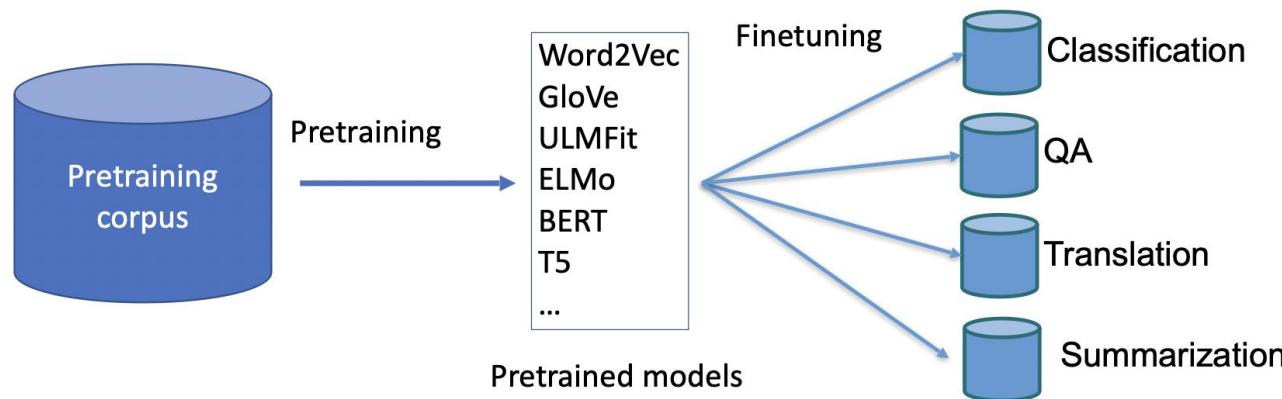
Adaptation via Fine-Tuning

- At this point, we have built a pre-trained model.

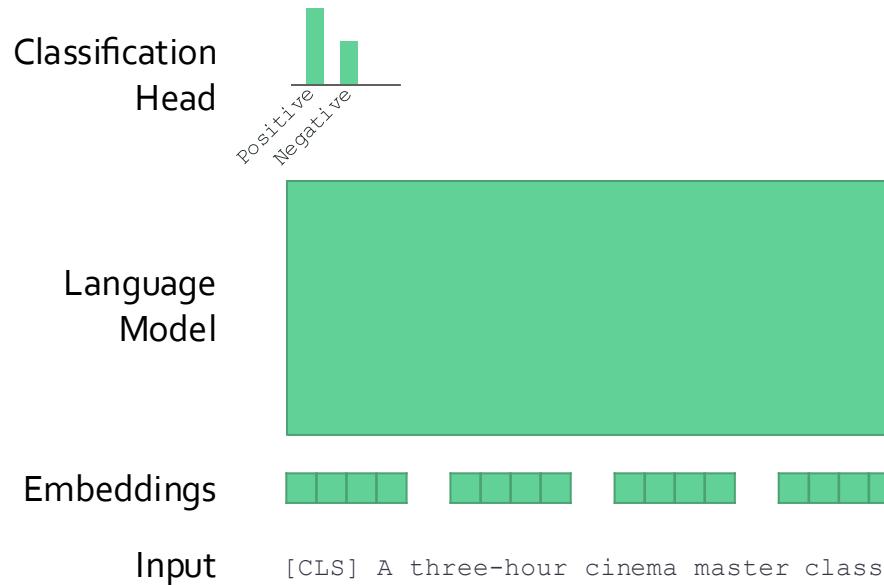


Fine-Tuning for Tasks

- Now we want to “adapt” it for specific tasks with labeled data.



Fine-tuning Pre-trained Models

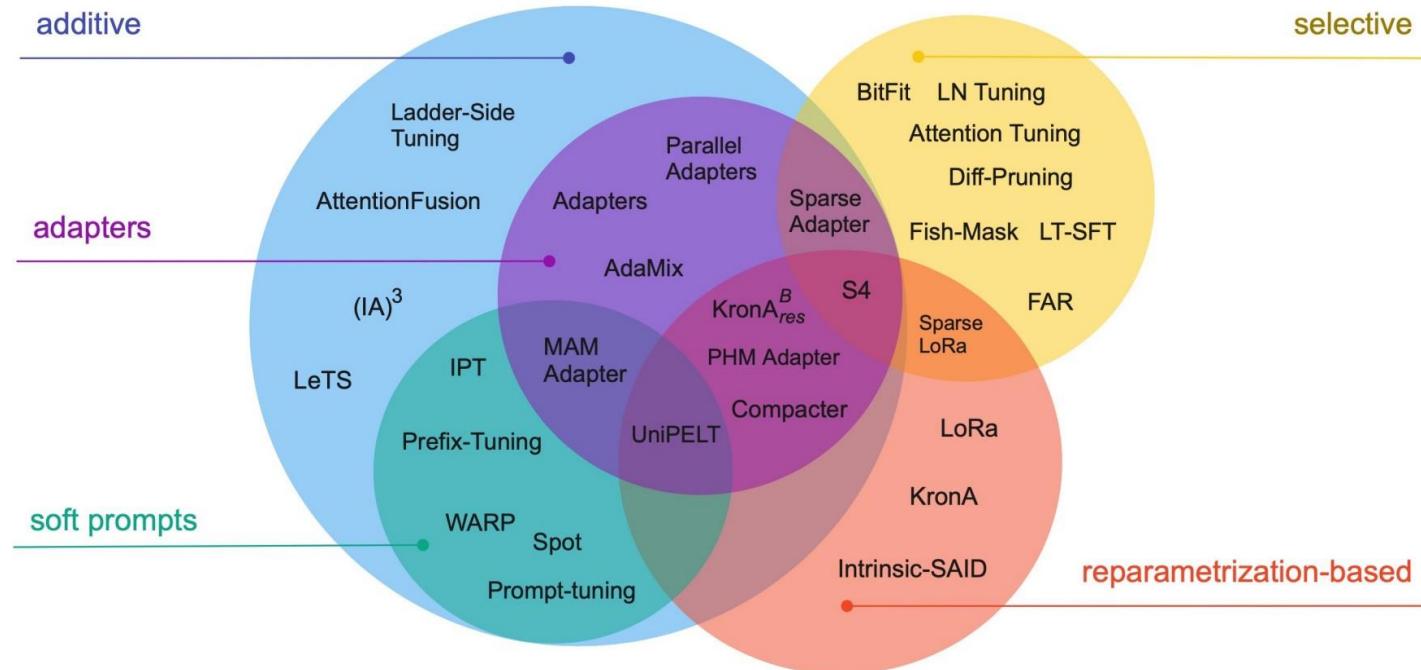


- Whole model tuning:
 - Run an optimization defined on your task data that updates **all** model parameters
- Head-tuning:
 - Run an optimization defined on your task data that updates the parameters of the model “head”

Parameter-efficient Fine-tuning

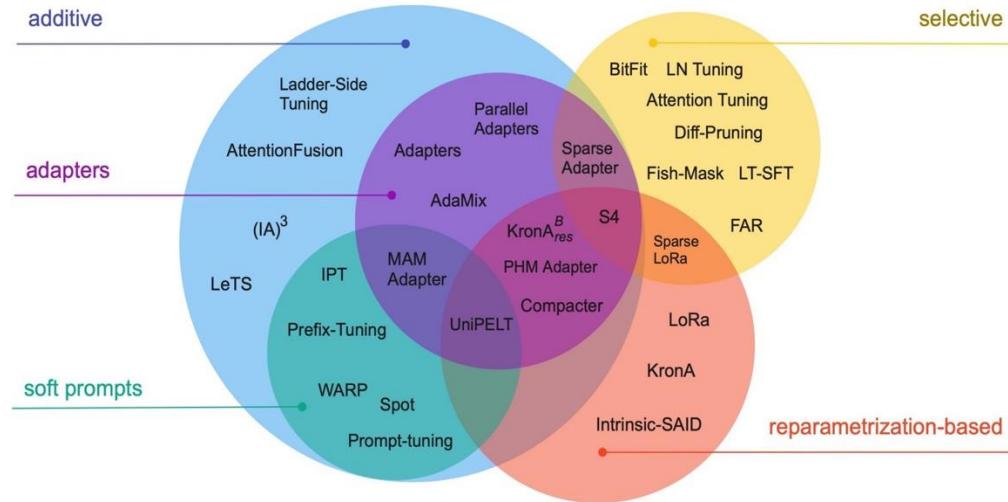
- In fine-tuning we need to updating and storing all the parameters of the LM
 - We would need to store a copy of the LM for each task
- With large models, storage management becomes difficult.
 - E.g., A model of size 170B parameters requires ~340Gb of storage
 - If you fine-tune a separate model for 100 tasks:
 - $340 * 100 = 34 \text{ TB}$ of storage!

Parameter-efficient Fine-tuning



Parameter-efficient Fine-tuning: Adding Models

- Augmenting the existing pre-trained model with extra parameters or layers and training only the new parameters
- Two commonly used methods:
 - Soft prompts
 - **Adapters**



LoRA: Low-Rank Adaptation

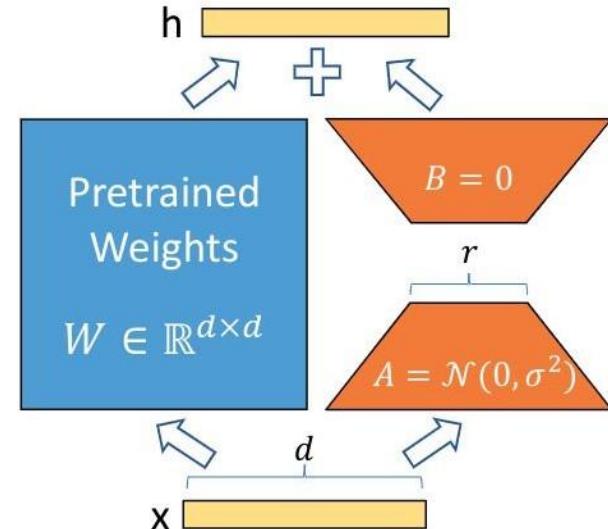
- Hypothesis: the intrinsic rank of the weight matrices in a large language model is low
- Parameter update for a weight matrix is decomposed into a product of two low-rank matrices

$$W \leftarrow W + \Delta W$$

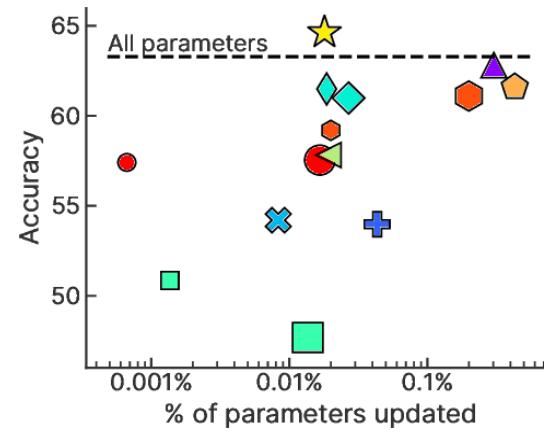
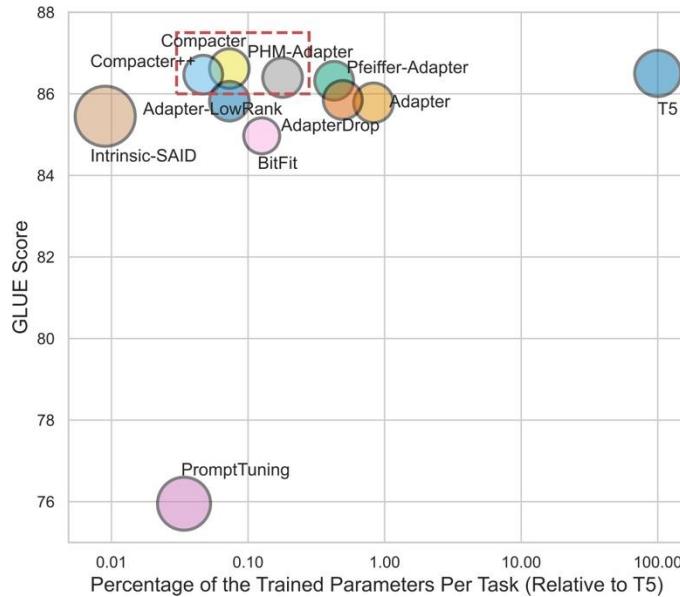
$$\Delta W = BA$$

$$B \in \mathbb{R}^{d,r}, A \in \mathbb{R}^{r,k}, r \ll \min(k, d)$$

- A is initialized with random Gaussian Initialization, B is initialized to zero



Performance/compactness comparison



- ★ (IA)³
- ▲ LoRA
- ✚ BitFit
- ✖ Layer Norm
- ◆ Compacter
- ◆ Compacter++
- Prompt Tuning
- ◀ Prefix Tuning
- Adapter
- FISH Mask
- Intrinsic SAID

Summary

- Parameter efficient optimization — optimize fewer parameters than the whole model.
 - Space efficiency — fewer parameters to store
 - Computation efficiency? Some gains since you're storing less trainable parameters.

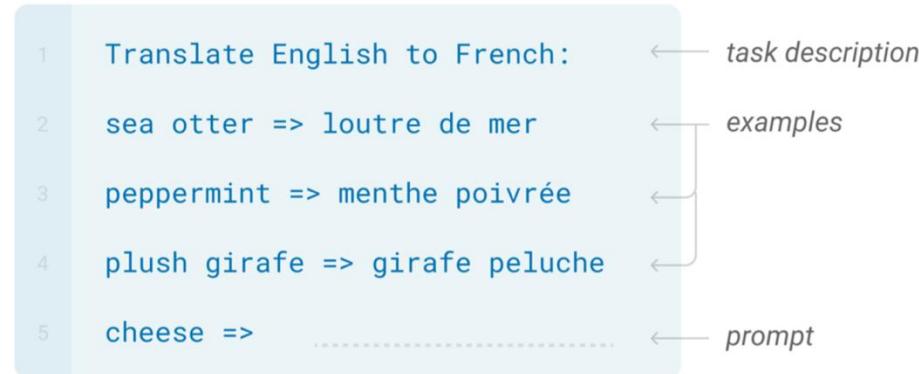
Prompting and In-Context Learning

In-Context Learning

- 1 Translate English to French: ← *task description*
- 2 sea otter => loutre de mer ← *examples*
- 3 peppermint => menthe poivrée ←
- 4 plush girafe => girafe peluche ←
- 5 cheese => ← *prompt*

In-Context Learning

- Learns to do a downstream task by conditioning on input-output examples!
- **No weight update** — our model is not **explicitly pre-trained** to learn from examples
 - The underlying models are quite general



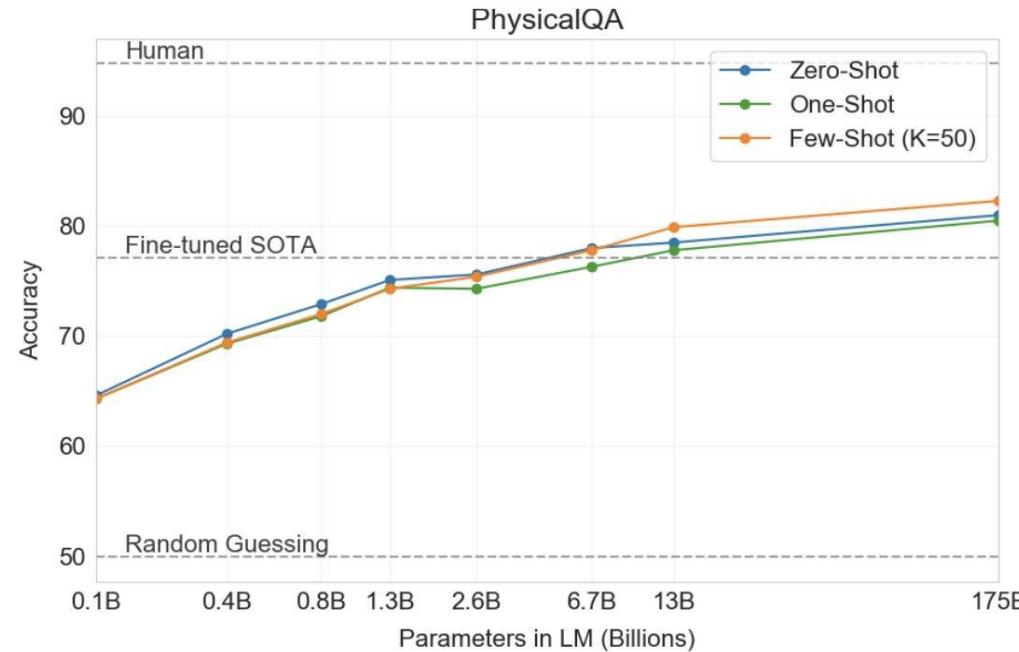
In-Context learning Results



To separate egg whites from the yolk using a water bottle, you should...

a. **Squeeze** the water bottle and press it against the yolk. **Release**, which creates suction and lifts the yolk.

b. **Place** the water bottle and press it against the yolk. **Keep pushing**, which creates suction and lifts the yolk.



Why Do We Care About In-Context Learning?

Practically Useful

Intellectually Intriguing

In-Context Learning: Practically Useful

- Labeling data is costly
 - May require domain expertise
 - Medical, legal, financial
 - You don't want to get more data
 - Emergent, time-sensitive scenarios
 - Something new happened—need to react quickly!
- Finetuning can tricky
 - Training is sensitive to hyperparams
 - Not enough validation data
 - We don't quite understand how finetuning works
 - Expensive to train, time and memory

In-Context Learning: Intellectually Intriguing

- Potential test for “Intelligent Behavior”
 - Generalization from few examples
 - Fundamental piece of intelligence
 - Often used in psychology
 - Quickly adjust to environment
- Insights into Language Modeling
 - What does an LLM “know”?
 - What are the biases/limitations of LLMs?
 - ...

ICL as a General-Purpose Few-Shot Learning Mechanism

Any arbitrary task



Language Model

A few-shot learner



ICL's Sensitivity

LM Prompting: Choices of Encoding

Prompt

Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

Input: Amazing. Sentiment:

LM Prompting: Choices of Encoding

Prompt

Input: Subpar acting. Contimont: negative

Sentence: Subpar acting. Label: bad

Sentence: Beautiful film. Label: good

Sentence: Amazing. Label:

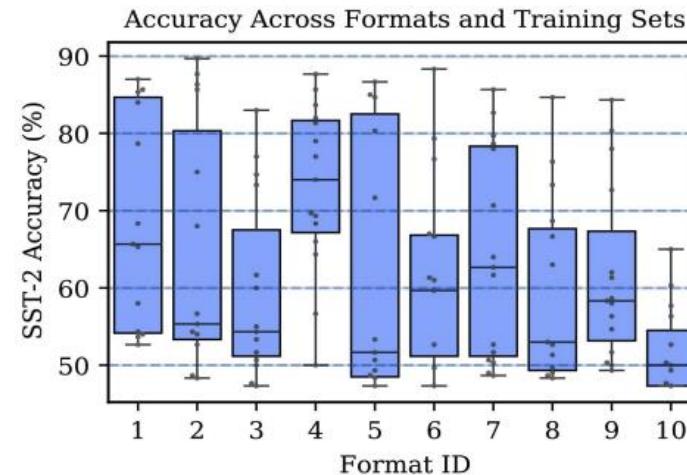
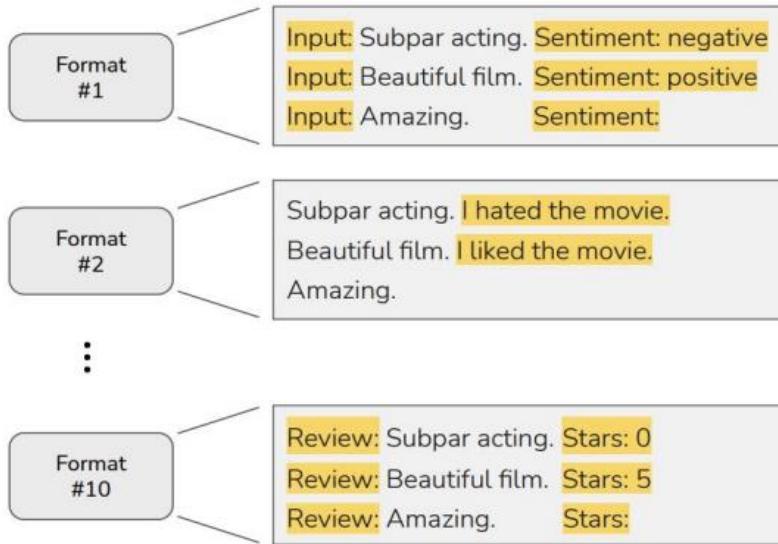
LM Prompting: Choices of Encoding

Prompt

The diagram illustrates a sequence of prompts and responses for sentiment analysis. It starts with a header row containing 'Input: Subpar acting' and 'Sentiment: negative'. Below this, there are three rows, each starting with 'In' followed by a question: 'Q: What's the sentiment of "Subpar acting"?', 'Q: What's the sentiment of "Beautiful film"?', and 'Q: What's the sentiment of "Amazing"?'. Each question is followed by an 'A:' response. The first two questions and their answers are highlighted in yellow boxes, while the third question and its answer are shown in a standard black font.

Input: Subpar acting	Sentiment: negative
In	Q: What's the sentiment of "Subpar acting"?
In	A: negative
In	Q: What's the sentiment of "Beautiful film"?
Se	A: positive
Q: What's the sentiment of "Amazing"?	A:

In-Context Learning: Sensitivity to Encoding



In-context learning is highly sensitive to prompt format (training sets and patterns/verbalizers)

What Causes These Variances?

- Here we will provide several factors ...

Majority Label Bias

Prompt

Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

Input: Amazing. Sentiment:

- Among 4 demonstrations, count how many are “positive”.
- Then check if the model output correlates with the number of “positive” demos.

Majority Label Bias

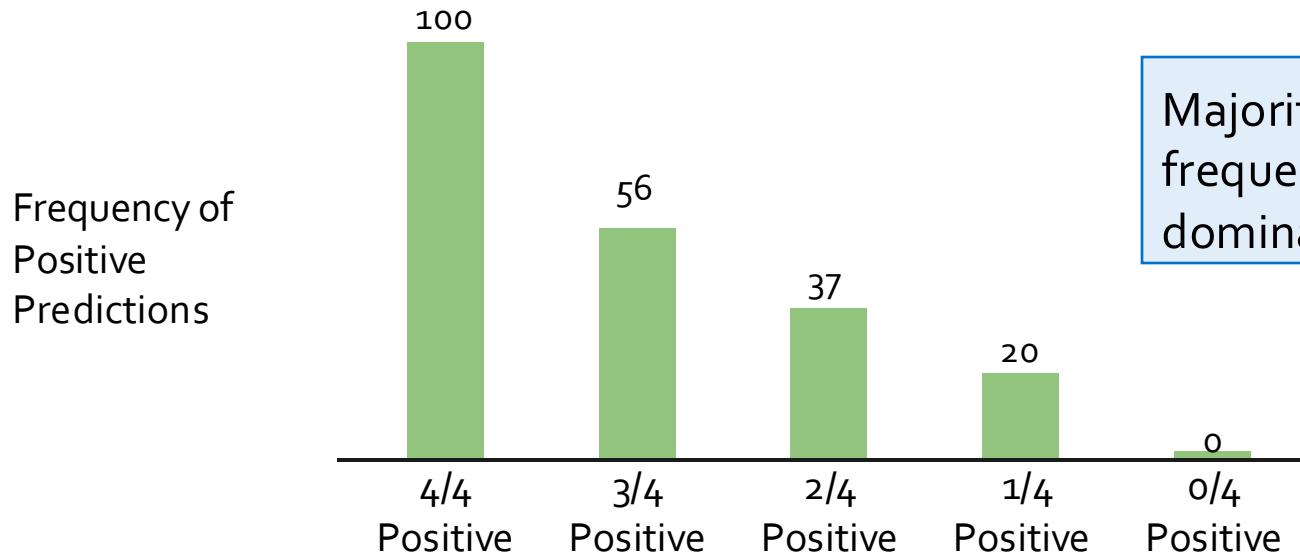
Prompt

Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

Input: Amazing. Sentiment:

- Among 4 demonstrations, count how many are “positive”.
- Then check if the model output correlates with the number of “positive” demos.



Majority label bias:
frequent training answers
dominate predictions.

Recency Bias

Prompt

Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

Input: Amazing. Sentiment:

- Check if the label of the most-recent demo biases the model output.

Recency Bias

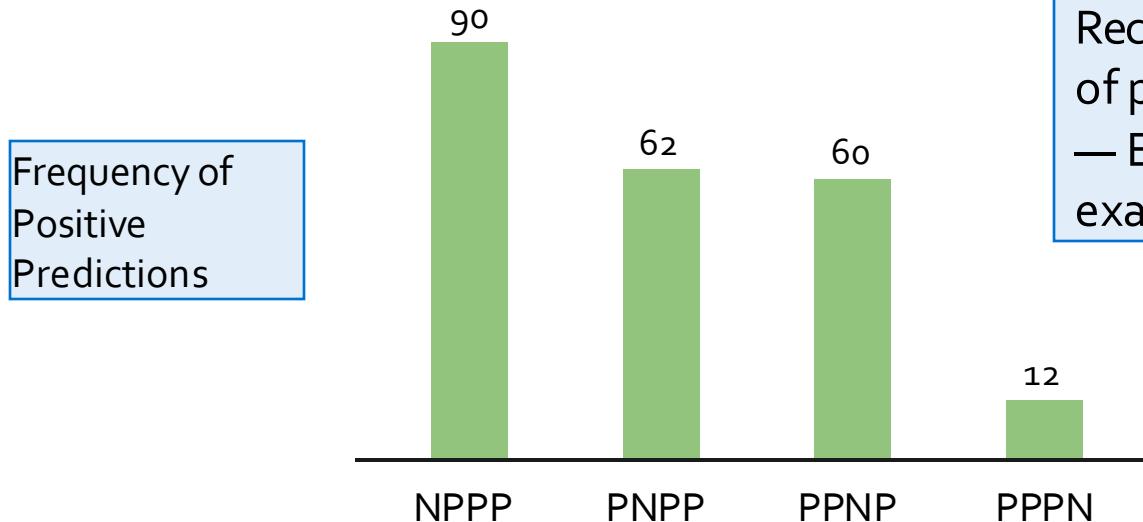
Prompt

Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

Input: Amazing. Sentiment:

- Check if the label of the most-recent demo biases the model output.



Recency bias: examples near end of prompt dominate predictions
— Explains variance across example permutations!

Common Token Bias

- Is a bias toward words that are more frequently observed in pre-training data?

Common Token Bias

Language Model



What topic is the following text about?

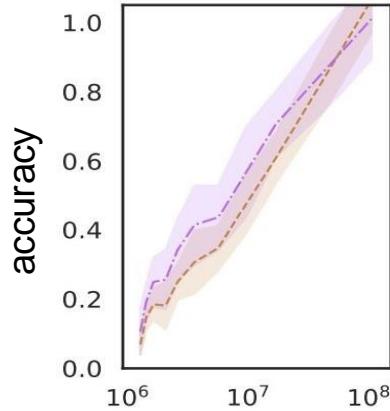
Prompt The Model T was released by Ford in 1908.

Answer:

Impact of Pretraining Term Frequencies

- For each task, identify relevant terms from each instance—numbers and units
- Count co-occurrences of these terms in the pretraining data (term pairs or triples within a fixed window)

Impact of Pretraining Term Frequencies



(a) Arithmetic-Multiplication

Q: What is 24 times 18?

A: 432 ✓

Q: What is 23 times 18?

A: 462 ✗

In-context learning performance is highly correlated with term frequencies during pretraining

Why Does ICL Emerge?

- We don't know!
- We have partial empirical explanations.
- And some theoretical analogies.
- But none of them fully explain ICL.

Prompting to Solve Multi-step Problems

Some Problems Involve Reasoning

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: The answer is **5**

Q: Take the last letters of the words in "Elon Musk" and concatenate them

A: The answer is **nk**.

Q: What home entertainment equipment requires cable?
Answer Choices: (a) radio shack
(b) substation (c) television (d) cabinet

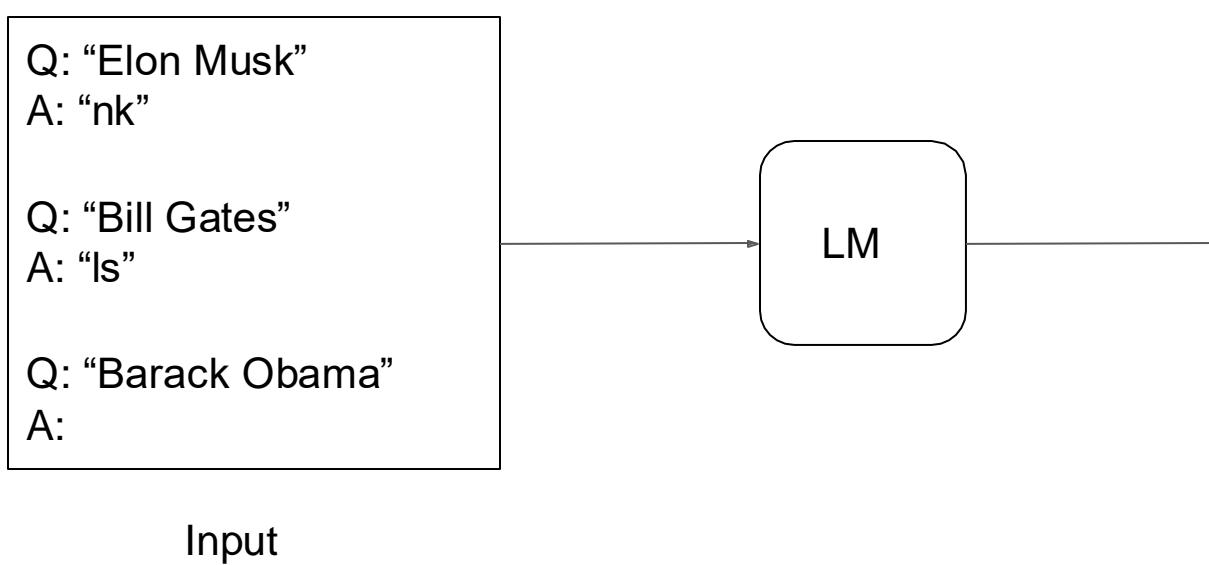
A: The answer is **(c)**.

Arithmetic Reasoning (AR)
 $(+ - \times \div \dots)$

Symbolic Reasoning (SR)

Commonsense Reasoning (CR)

Vanilla ICL on Reasoning Problems



Playground

Load a preset...

Save

View code

Share

...

Q: "Elon Musk"

A: "nk"

Q: "Bill Gates"

A: "Is"

Q: "Barack Obama"

A: "ma"



Complete

Model

text-davinci-003

Temperature 0

Maximum length 256

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

Submit



54

Playground

Load a preset...



Save

View code

Share

...

Q: "Elon Musk"



A: "nk"

Q: "Bill Gates"

A: "Is"

Q: "Barack Obama"

A: "ma"

How about adding more examples?

Submit



Complete



Model

text-davinci-003



Temperature

0

Maximum length

256

Stop sequences

Enter sequence and press Tab

Top P

1

Frequency penalty

0

Presence penalty

0

Playground

Load a preset...

Save

View code

Share

...

Q: "Elon Musk"

A: "nk"

Q: "Bill Gates"

A: "ls"

Q: "Steve Jobs"

A: "es"

Q: "Larry Page"

A: "ye"

Q: "Jeff Bezos"

A: "fs"

Q: "Barack Obama"

A: "ma"



Submit



94

Complete

Model

text-davinci-003

Temperature

0

Maximum length

256

Stop sequences

Enter sequence and press Tab

Top P

1

Frequency penalty

0

Presence penalty

0

CoT: Adding “thought” before “answer”

Q: “Elon Musk”

A: the last letter of "Elon" is "n". the last letter of "Musk" is "k". Concatenating "n", "k" leads to "nk". so the output is "nk".

thought

Q: “Bill Gates”

A: the last letter of "Bill" is "l". the last letter of "Gates" is "s". Concatenating "l", "s" leads to "ls". so the output is "ls".

Q: “Barack Obama”

A:

CoT: Adding “thought” before “answer”

Q: “Elon Musk”

A: the last letter of "Elon" is "n". the last letter of "Musk" is "k". Concatenating "n", "k" leads to "nk". so the output is "nk".

thought

Q: “Bill Gates”

A: the last letter of "Bill" is "l". the last letter of "Gates" is "s". Concatenating "l", "s" leads to "ls". so the output is "ls".

Q: “Barack Obama”

A: the last letter of "Barack" is "k". the last letter of "Obama" is "a". Concatenating "k", "a" leads to "ka". so the output is "ka".

CoT: Adding “thought” before “answer”

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. X

Apply CoT to Any Task

Though each task's demonstrations need to be "engineered" manually! 😞

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm³, which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

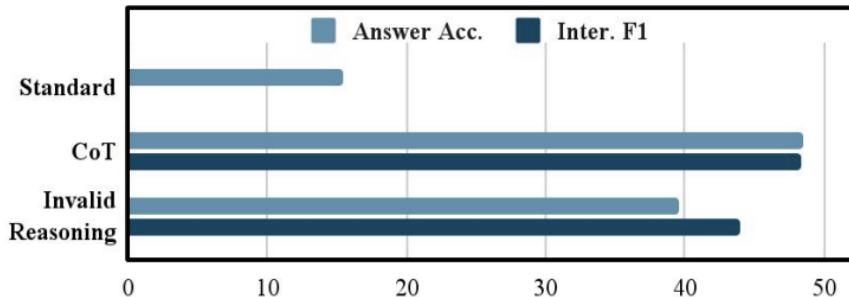
Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

Multi-Step Prompting: Steps Don't Have to Be Correct!

- It is possible even with invalid demonstration
- Prompting with invalid reasoning steps can act with correct reasoning steps.



	In-context Demonstration	Inference by LLM
Query	Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	Julie is reading a 120-page book. Yesterday, she was able to read 12 pages and today, she read twice as many pages as yesterday. If she wants to read half of the remaining pages tomorrow, how many pages should she read?
Standard	The answer is 39.	The answer is 18. ✗
CoT	Originally, Leah had 32 chocolates and her sister had 42. So in total they had $32 + 42 = 74$. After eating 35, they had $74 - 35 = 39$ pieces left in total. The answer is 39.	Julie is reading a 120-page book. Yesterday, she read 12 pages and today, she read 24 pages. So she read a total of $12 + 24 = 36$ pages. Now she has $120 - 36 = 84$ pages left. Since she wants to read half of the remaining pages, she should read $84 / 2 = 42$ pages. The answer is 42. ✓
Invalid Reasoning	Originally, Leah had 32 chocolates and her sister had 42. So her sister had $42 - 32 = 10$ chocolates more than Leah has. After eating 35, since $10 + 35 = 45$, they had $45 - 6 = 39$ pieces left in total. The answer is 39.	Yesterday, Julie read 12 pages. Today, she read $12 * 2 = 24$ pages. So she read a total of $12 + 24 = 36$ pages. Now she needs to read $120 - 36 = 84$ more pages. She wants to read half of the remaining pages tomorrow, so she needs to read $84 / 2 = 42$ pages tomorrow. The answer is 42. ✓

Summary

- Prompting language models is a powerful way to adapt them to our desired tasks.
 - We saw prompting via in-context demonstrations
 - We also saw various variants and extensions
- They also serve as a gateway to understand the underlying dynamics inside models.
- Lots of activity in this area and room for a lot of research progress.

Prompt Engineering

- Reformulating tasks to a language that is easier to for the models.
 - Show demonstrations
 - Decompose your problem
 - Ask for rationales (a la CoT)
 - Check for consistency
 - ...
- **Question for you:** will “prompt engineering” be relevant topic in the coming years?

AI 'prompt engineer' jobs can pay up to \$375,000 a year and don't always require a background in tech

Britney Nguyen May 1, 2023, 11:34 PM GMT+8

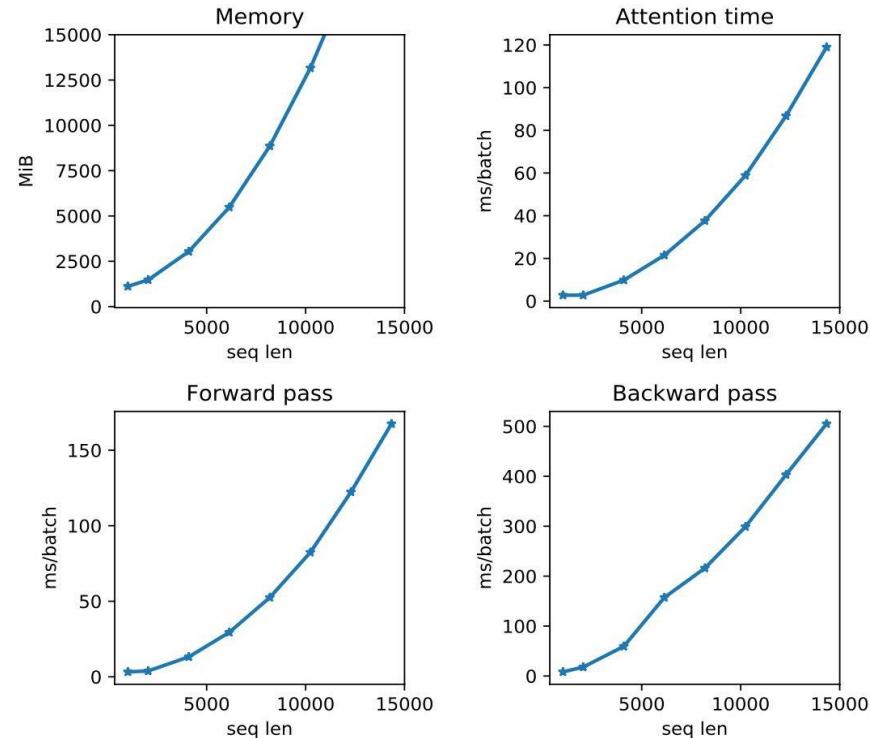


The rise of generative AI tools like ChatGPT is creating a hot market for "prompt engineers" who test and improve chatbot answers. Getty Images

Long Context: Efficiency and Generalization

Transformer LMs and Long Inputs

- **Length generalization:** Do Transformers work accurately on long inputs?
- **Efficiency considerations:** How efficient are LMs on long inputs?



Transformer LMs and Long Inputs

- A wide variety of topics.
- We will cover few only.

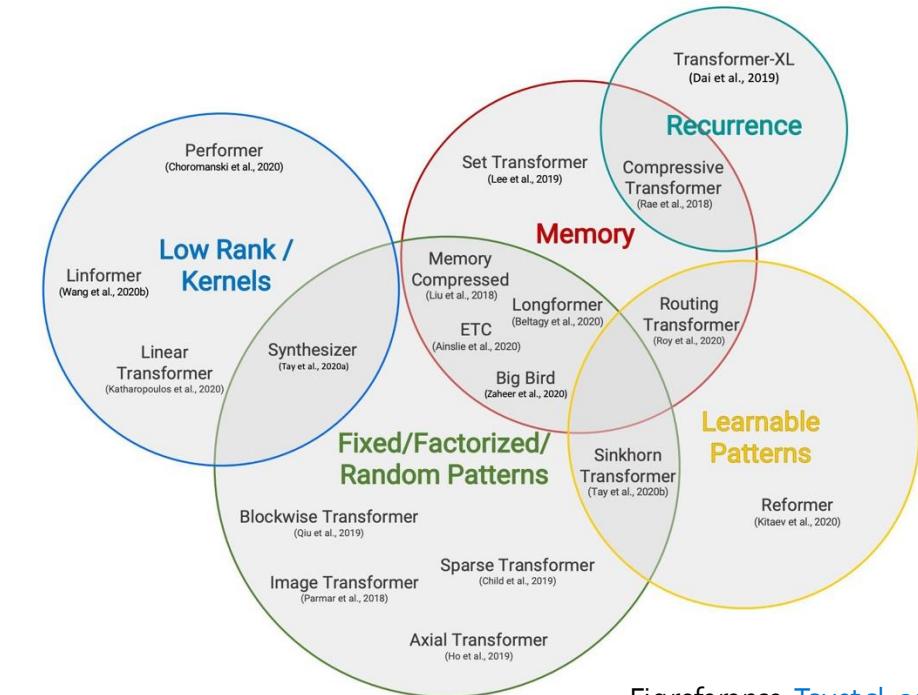


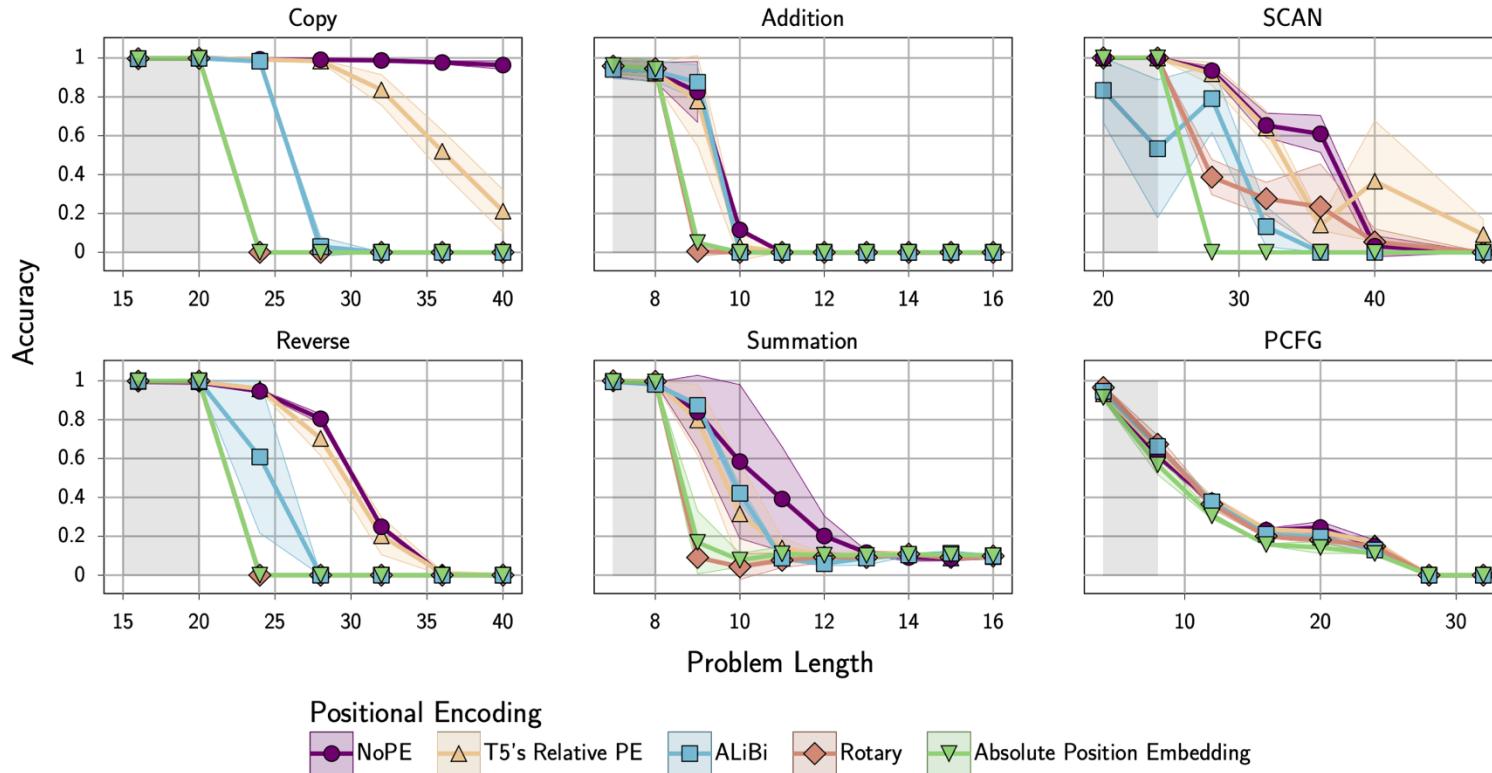
Fig reference: [Tay et al., 2020](#)

Length Generalization

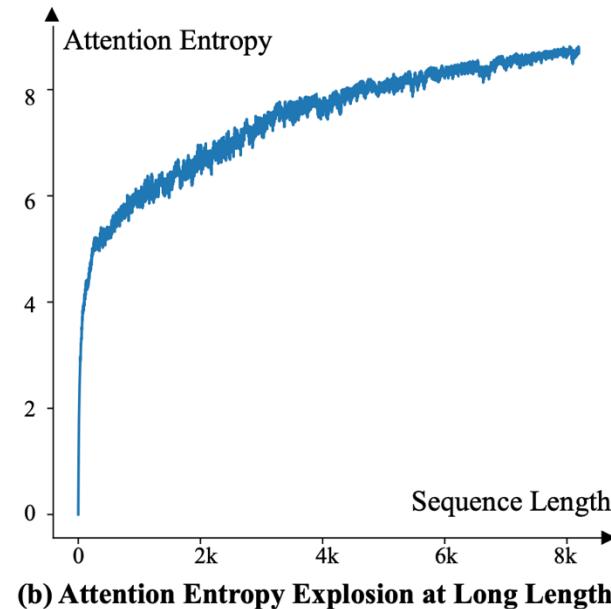
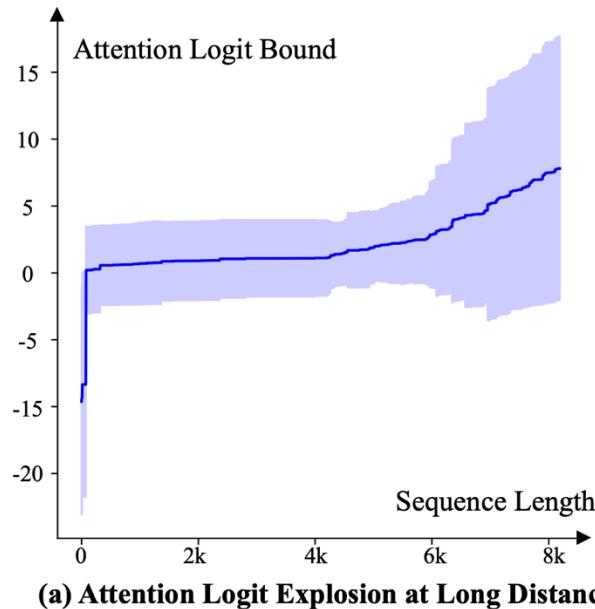
Table 1: Examples of the input and output of the tasks.

Task	Input Example	Output Example
Primitive Tasks		
Copy	Copy the following words: <w1> <w2> <w3> <w4> <w5>	<w1> <w2> <w3> <w4> <w5>
Reverse	Reverse the following words: <w1> <w2> <w3> <w4> <w5>	<w5> <w4> <w3> <w2> <w1>
Mathematical and Algorithmic Tasks		
Addition	Compute: 5 3 7 2 6 + 1 9 1 7 ?	The answer is 5 5 6 4 3.
Polynomial Eval.	Evaluate x = 3 in (3 x ** 0 + 1 x ** 1 + 1 x ** 2) % 10 ?	The answer is 5.
Sorting	Sort the following numbers: 3 1 4 1 5 ?	The answer is 1 1 3 4 5.
Summation	Compute: (1 + 2 + 3 + 4 + 7) % 10 ?	The answer is 7.
Parity	Is the number of 1's even in [1 0 0 1 1] ?	The answer is No.
LEGO	If a = -1; b = -a; c = +b; d = +c. Then what is c?	The answer is +1.
Classical Length Generalization Datasets		
SCAN	jump twice and run left	JUMP JUMP TURN_LEFT RUN
PCFG	shift prepend K10 R1 K12 , E12 F16	F16 K10 R1 K12 E12

Length Generalization



Why Transformers Fail to Generalize?



Unseen distances between tokens cause attention logits to explode

Attention distribution becomes less informative (high entropy).

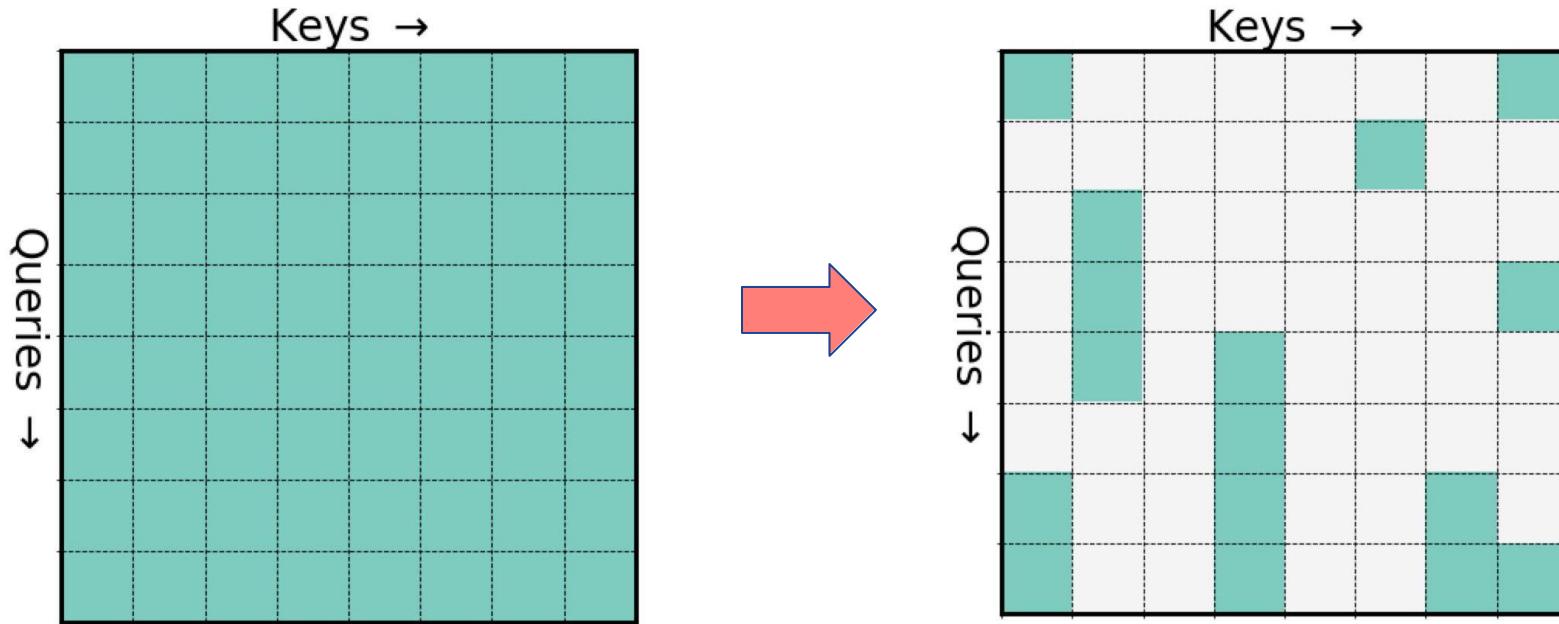


Efficiency considerations



Sparse Attention Patterns

- The idea is to make the attention operation sparse

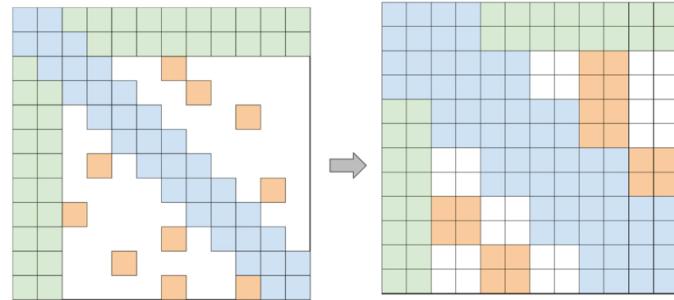


Sparse Attention Patterns: Challenge

- Ok sparsity is great, but how to efficiently implement this?
- **Challenge:** Arbitrary sparse matrix multiplication is not supported in DL libraries
- **A solution:** Perform computations in blocks

Pre-specified Sparsity Patterns: Computations

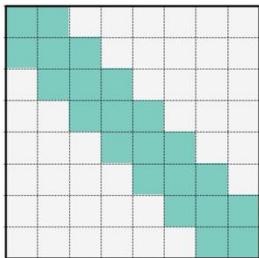
- Efficient blockified implementation
- There are libraries for implementing blockified sparse matrix multiplication.
 - Can be hardware specific
 - Block Sparse ([Gray et al., 2017](#))
 - TVM toolkit ([Chen et al., 2018](#))
 - cuSPARSE



Pre-specified Sparsity Patterns

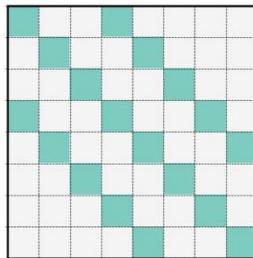
- A variety of patterns has been explored in the past work
 - Longformer ([Beltagy et al., 2020](#)), Sparse Transformer ([Child et al., 2019](#)), ...

Sliding window



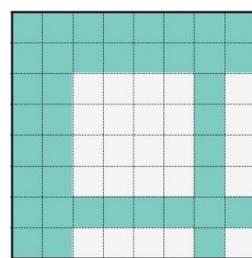
Sparse Transformer
Longformer

Dilated



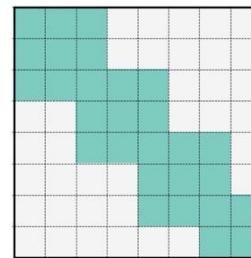
Longformer

Global



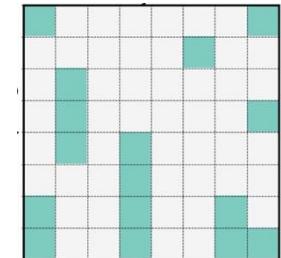
Big Bird

Blocked



Big Bird
Sinkhorn

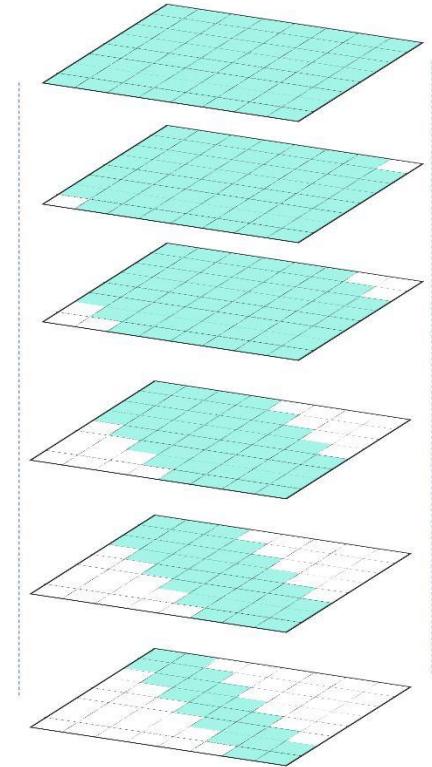
Random



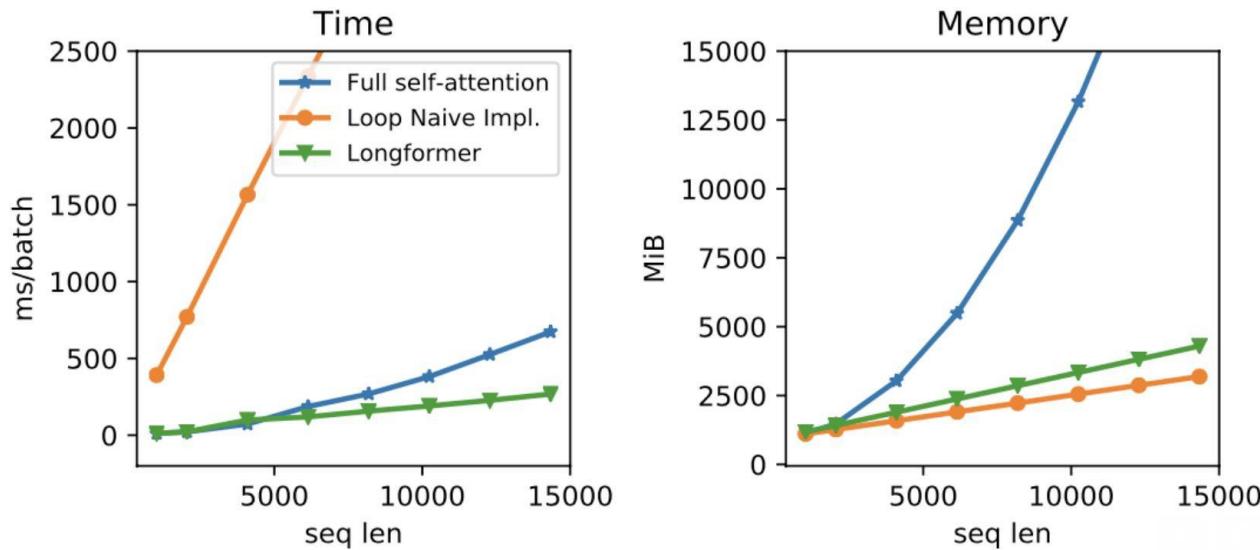
Big Bird

Pre-specified Sparsity Patterns

- Different layers and attention heads can follow different patterns
- A common setup is to have earlier layers with sparser attention pattern.
 - Longformer ([Beltagy et al., 2020](#))



Pre-specified Sparsity Patterns: Computations



A Notable Adoption: GPT-3

- Sparse patterns also used in GPT-3 ([Brown et al., 2020](#))

2.1 Model and Architectures

We use the same model and architecture as GPT-2 [RWC⁺19], including the modified initialization, pre-normalization, and reversible tokenization described therein, with the exception that we use alternating dense and locally banded sparse attention patterns in the layers of the transformer, similar to the Sparse Transformer [CGRS19]. To study the dependence of ML performance on model size, we train 8 different sizes of model, ranging over three orders of magnitude from 125 million parameters to 175 billion parameters, with the last being the model we call GPT-3. Previous work [KMH⁺20]

Summary

- How well do Transformers work on long sequences? Not so well.
- How can we make them more efficient? Induce sparsity.
- Many open questions we did not get into:
 - Limitations of sportifying Transformers
 - Alternatives to Transformers (e.g., state-space models)
 - ...

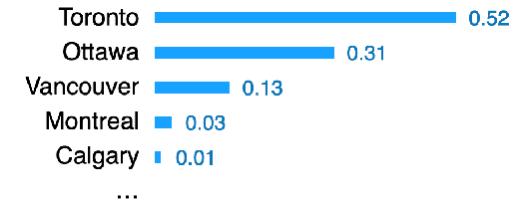
Retrieval-Augmented LMs

Retrieval-based Language Models

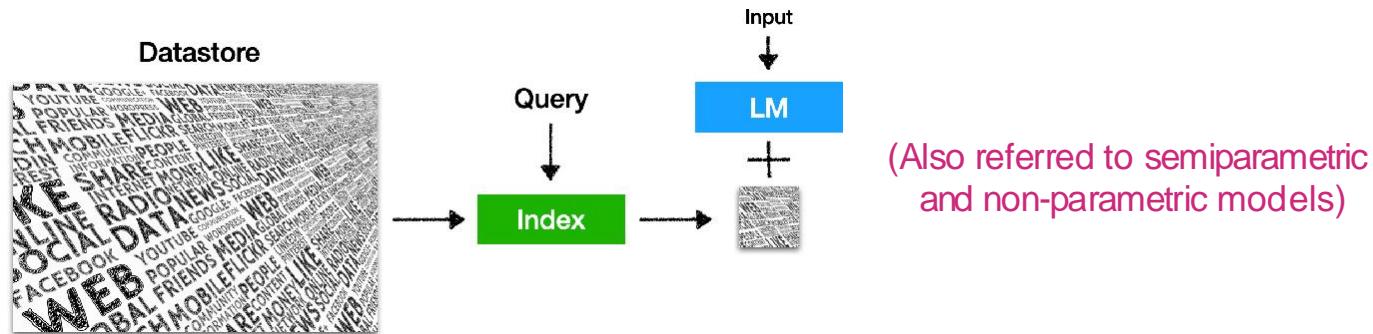
- It is a **language model** $P(x_n | x_1, x_2, \dots, x_{n-1})$

The capital city of Ontario is __

(can be broadly extended to masked language models or encoder-decoder models)



- It retrieves from an **external datastore** (at least during inference time)



Why retrieval-based LMs?

LLMs can't memorize all (long-tail) knowledge in their parameters



List 5 important papers authored by Geoffrey Hinton

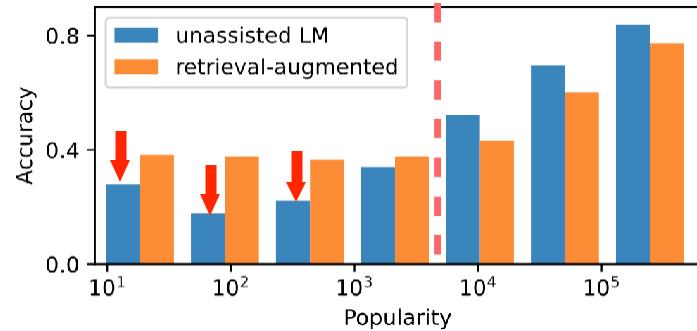


Geoffrey Hinton is a renowned computer scientist ...

Here are five important papers authored by him:

1. **"Learning Internal Representations by Error Propagation"** (with D. E. Rumelhart and R. J. Williams) - This paper, published in **1986**, ... 
2. **"Deep Boltzmann Machines"** (with R. Salakhutdinov) - Published in **2009**, ... 
3. ...
4. **"Deep Learning"** (with Y. Bengio and A. Courville) - Published as a book in **2016**, ... 
5. **"Attention Is All You Need"** (with V. Vaswani, N. Shazeer, et al.) - Published in **2017**, this paper introduced the Transformer model, ... 

What is Kathy Saltzman's occupation?



(Mallen et al., 2023)

GPT-3 davinci-003: 20%-30% accuracy

Why retrieval-based LMs?

LLMs' knowledge is easily outdated and hard to update



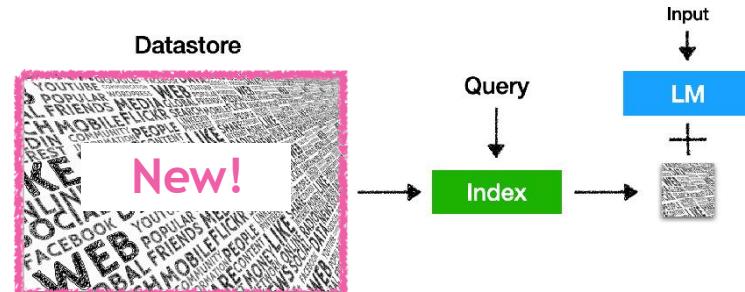
Who is the CEO of Twitter?



As of my **knowledge cutoff in September 2021**, the CEO of Twitter is **Jack Dorsey**....

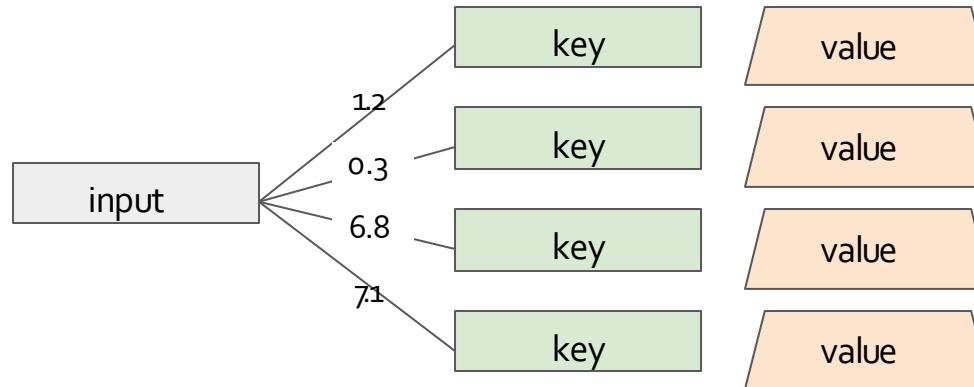
- Existing **knowledge editing** methods are still NOT scalable (**active research!**)
- The datastore can be easily **updated** and **expanded** - even without retraining!

A screenshot of a Google search results page. The search query "Who is the CEO of Twitter?" is entered in the search bar. Below the search bar, there are navigation links for All, News, Images, Shopping, Videos, and More, along with a Tools button. The search results section shows a snippet for "Twitter / CEO" followed by a card for "Linda Yaccarino" with a photo of her. The date "Jun 5, 2023–" is also visible.



Anatomy of a Neural Retriever

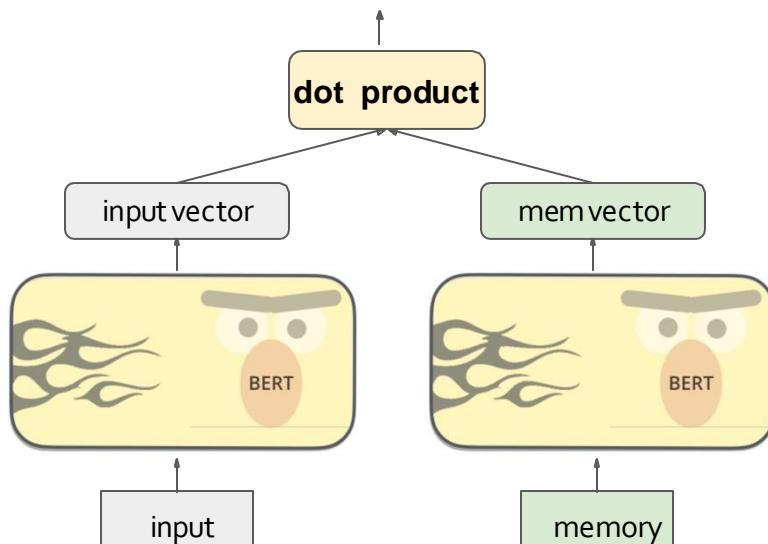
1. Score the input against each key.
2. Return the value for the highest scoring key.



A similarity function: $\text{sim}(\text{input}, \text{key}) \rightarrow \text{score}$

Similarity via Sentence Embeddings

$$\text{sim}(I, M) = \text{Encoder}(I) \times \text{Encoder}(M)$$



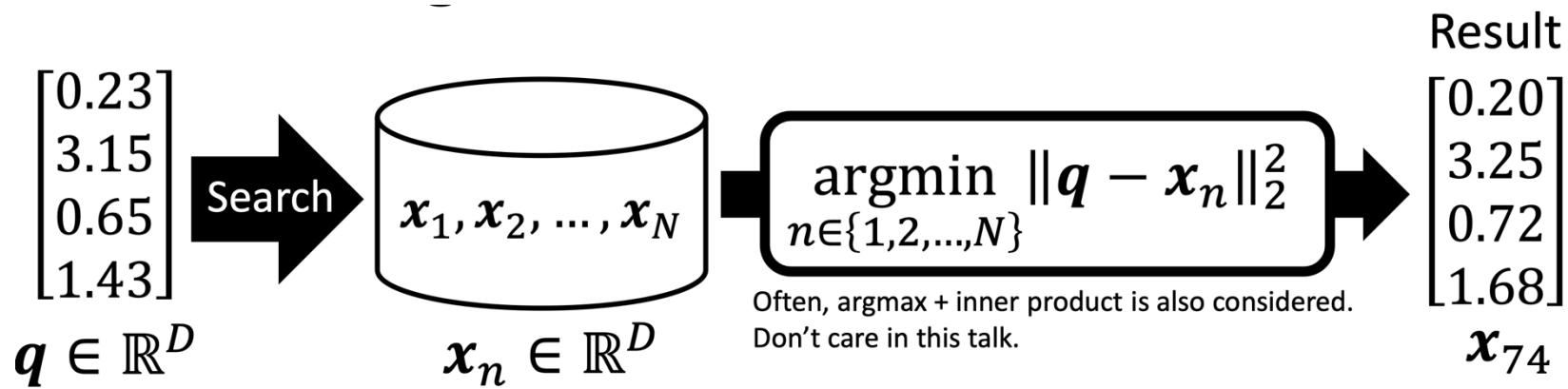
- **Advantages:**

- Differentiable -- can optimize with gradient descent.

- **Disadvantages:**

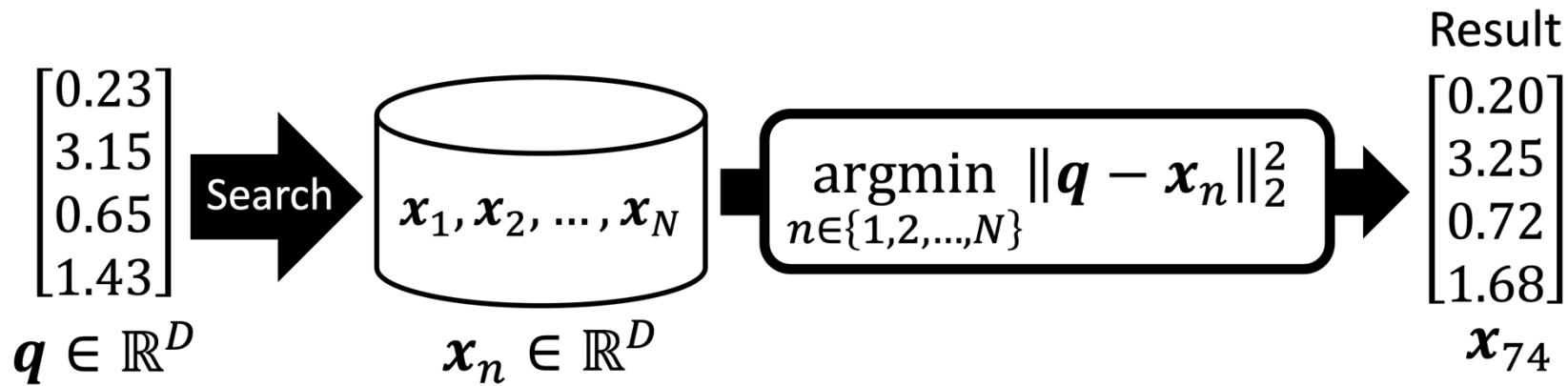
- Works well for data on which your LM is pre-trained on.

Finding Nearest Neighbors



- N D -dim database vectors: $\{x_n\}_{n=1}^N$
- Given a query q , find the closest vector from the database
- One of the fundamental problems in computer science
- Solution: linear scan, $O(ND)$, slow ☹

Approximate Finding Nearest Neighbors



- Faster search
- Don't necessarily have to be exact neighbors
- Trade off: runtime, accuracy, and memory-consumption

Approximate NNs: Algorithms, Libraries, Services

Algorithm

- Scientific paper
- Math
- Often, by researchers

Product Quantization +
Inverted Index (PQ, IVFPQ)
[Jégou+, TPAMI 2011]

Hierarchical Navigable
Small World (HNSW)
[Malkov+, TPAMI 2019]

ScaNN (4-bit PQ)
[Guo+, ICML 2020]

Library

- Implementations of algorithms
- Usually, a search function only
- By researchers, developers, etc

faiss

NMSLIB

hnswlib

ScaNN

Pinecone

Qdrant

Milvus

jina

Vald

Vertex AI
Matching Engine

Weaviate

Three levels of technology

Product Quantization +
Inverted Index (PQ, IVFPQ)
[Jégou+, TPAMI 2011]

Hierarchical Navigable
Small World (HNSW)
[Malkov+, TPAMI 2019]

faiss

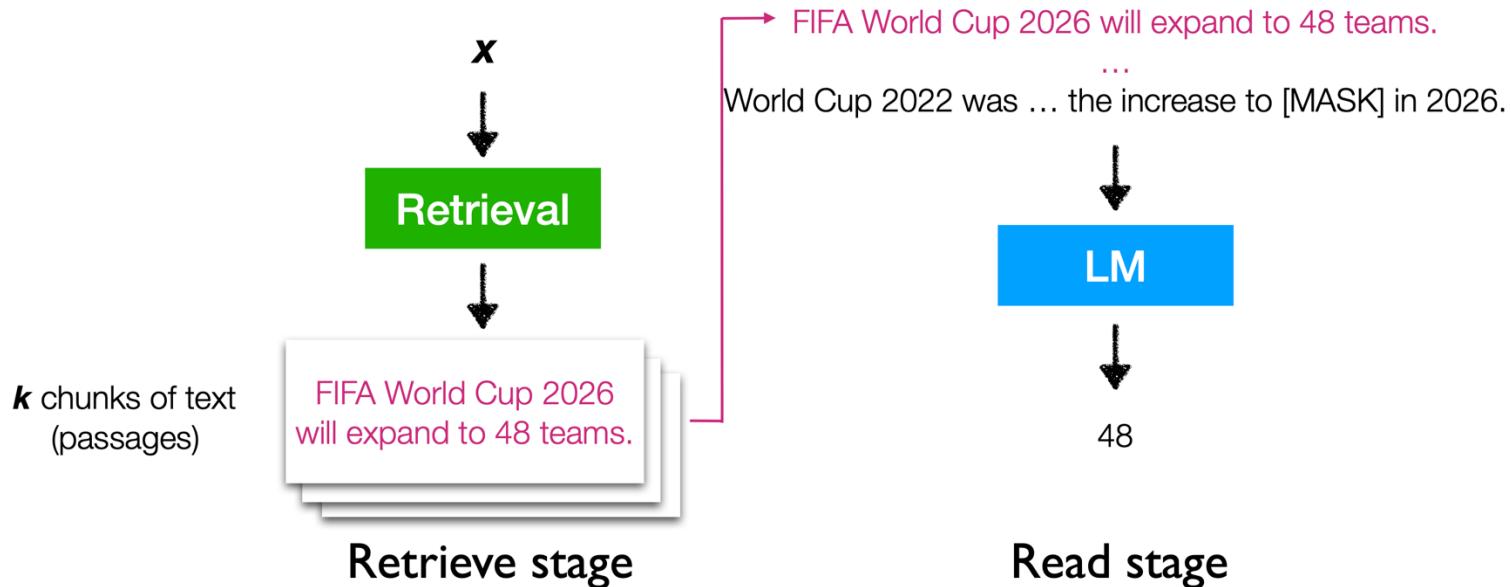
One library may implement
multiple algorithms

- “I benchmarked faiss”
- “I benchmarked PQ in faiss”

Let's assume that we have our retrieval engine and data ready

Retrieval-Augmented LM

- x = World Cup 2022 was the last before the increase to [MASK] in the 2026 tournament.

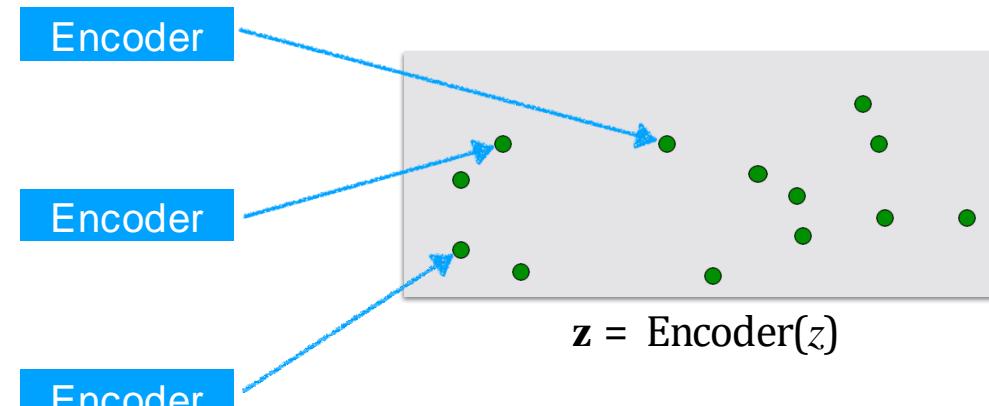


Retrieval-Augmented LM

FIFA World Cup 2026 will expand to 48 teams.

In 2022, the 32 national teams involved in the tournament.

Team USA celebrated after winning its match against Iran ...



Wikipedia

13M chunks (passages)
(called *documents* in the paper)

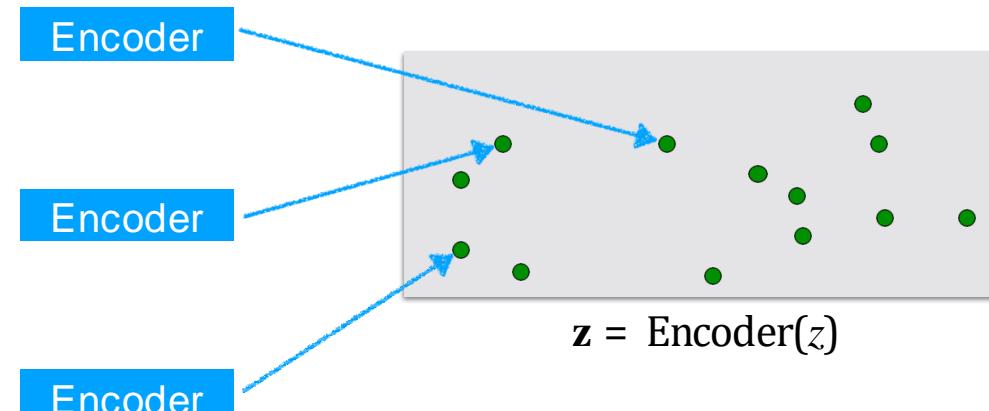
Retrieval-Augmented LM

x = World Cup 2022 was ... the increase to [MASK] in 2026.

FIFA World Cup 2026 will expand to 48 teams.

In 2022, the 32 national teams involved in the tournament.

Team USA celebrated after winning its match against Iran ...



Wikipedia

13M chunks (passages)
(called *documents* in the paper)

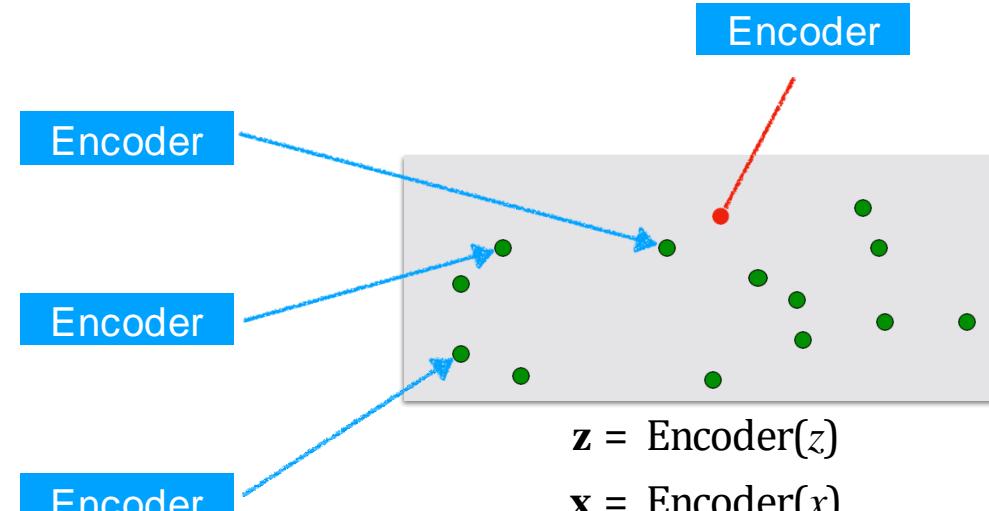
Retrieval-Augmented LM

x = World Cup 2022 was ... the increase to [MASK] in 2026.

FIFA World Cup 2026 will expand to 48 teams.

In 2022, the 32 national teams involved in the tournament.

Team USA celebrated after winning its match against Iran ...



Wikipedia

13M chunks (passages)
(called *documents* in the paper)

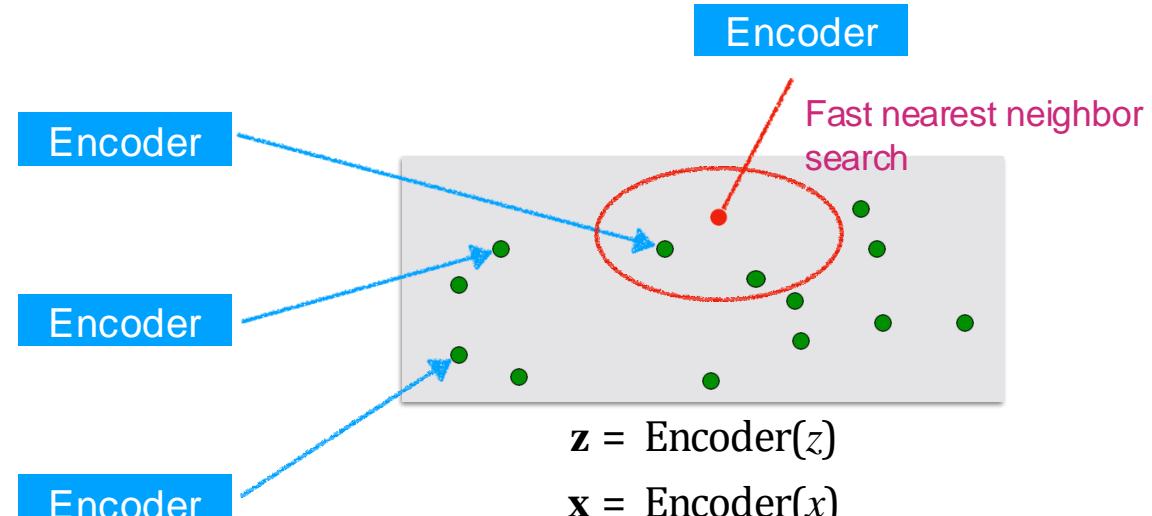
Retrieval-Augmented LM

x = World Cup 2022 was ... the increase to [MASK] in 2026.

FIFA World Cup 2026 will expand to 48 teams.

In 2022, the 32 national teams involved in the tournament.

Team USA celebrated after winning its match against Iran ...



Wikipedia

13M chunks (passages)
(called *documents* in the paper)

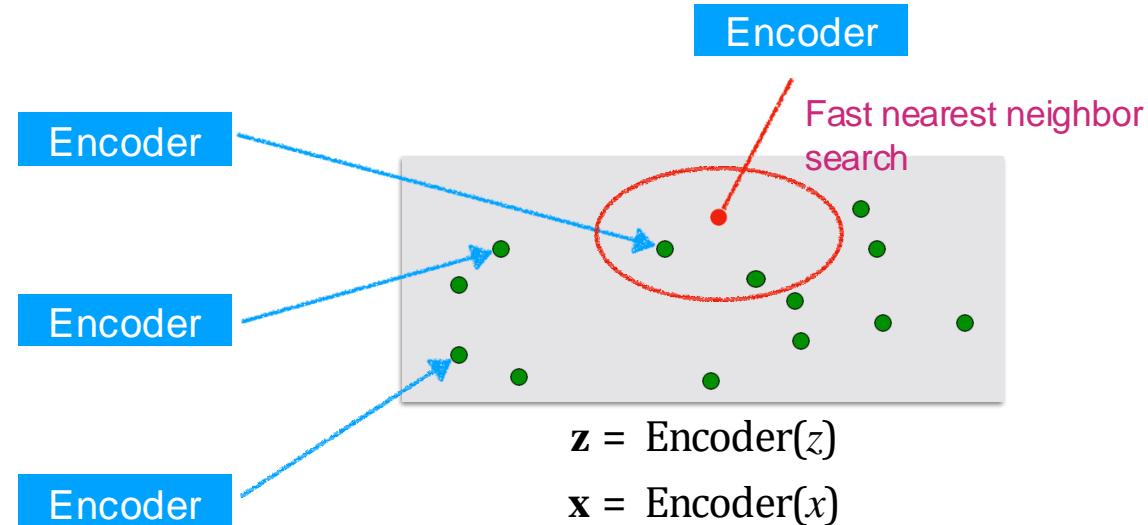
Retrieval-Augmented LM

x = World Cup 2022 was ... the increase to [MASK] in 2026.

FIFA World Cup 2026 will expand to 48 teams.

In 2022, the 32 national teams involved in the tournament.

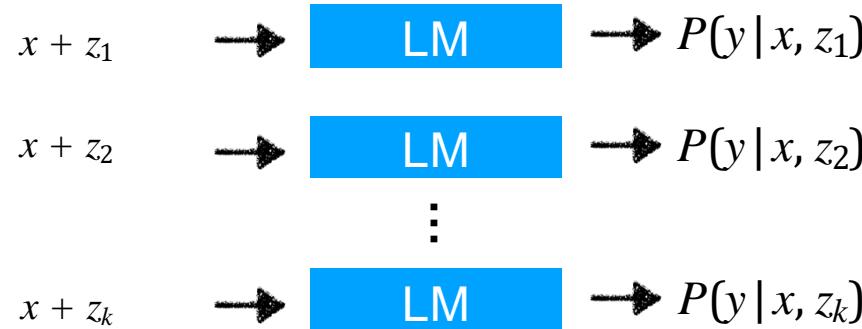
Team USA celebrated after winning its match against Iran ...



$$z_1, \dots, z_k = \text{argTop-}k (\mathbf{x} \cdot \mathbf{z})$$

k retrieved chunks

Retrieval-Augmented LM



$$\sum_{z \in \mathcal{D}} P(z | x) P(y | x, z)$$

Need to approximate
→ Consider top k chunks only

0 if not one of top k

\mathcal{D} from the retrieve stage
from the read stage

Summary

- How do we enable LMs to utilize external knowledge?
 - Retrieval-augmented language models
- A retriever is a function, $f(\text{input}, \text{memory}) \rightarrow \text{score}$
- What we did not discuss:
 - Attribution: Tracing decisions to the source knowledge
 - How to modify the knowledge
 - Conflicting knowledge
 - Editing knowledge
 - More efficient scaling
 -