
LECTURE 3-1

Transformer Review

Part 2: Generating Text, Model Families, and Beyond Text



Parisa Rashidi · Herbert Wertheim College of Engineering, University of Florida



Two Lectures, One Step at a Time

Today we turn the encoder into a model that can write.

Part 1 (last week)

- 1 Tokens and embeddings
- 2 The problem: context
- 3 Attention, step by step
- 4 Queries, keys, and values
- 5 Building the encoder block

Part 2 (this week)

- 6 Predicting the next token
- 7 Masking and the decoder
- 8 Encoder-decoder models
- 9 Model families
- 10 Beyond text: images and other data

Recap: The Encoder

Last week: five steps take a sentence to one contextual vector per token.



$$\text{Attention}(Q, K, V) = \text{softmax}(Q K^T / \sqrt{d_k}) V$$

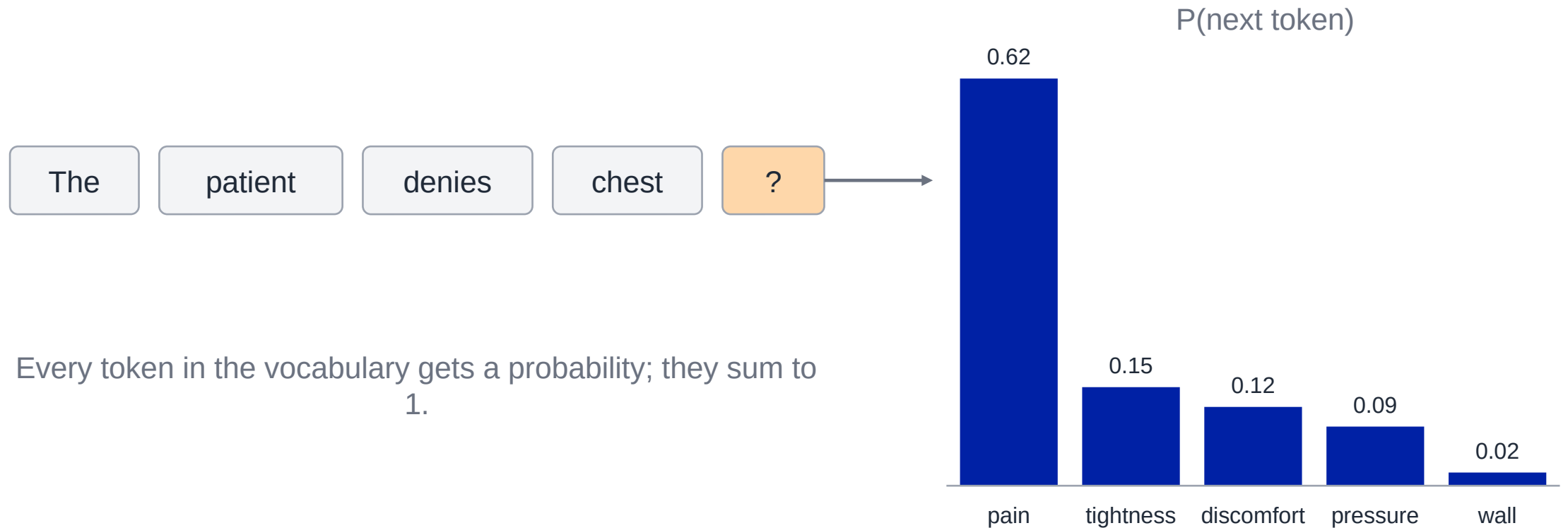
Open question: the encoder labels text. How does a model write new text?



Predicting the Next Token

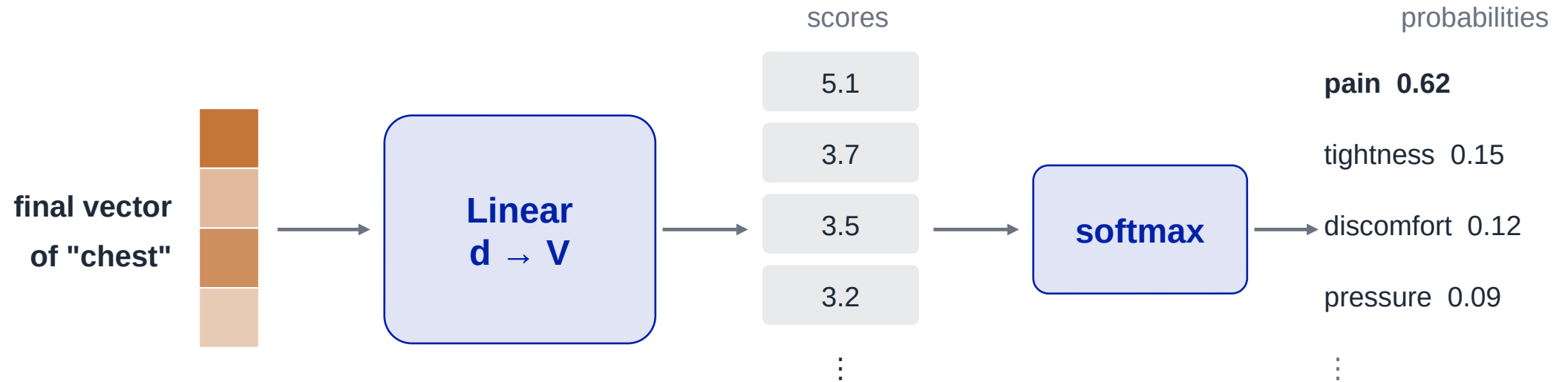
A New Task: What Comes Next?

A language model gives a probability to every token in the vocabulary as the next token.



The Prediction Head

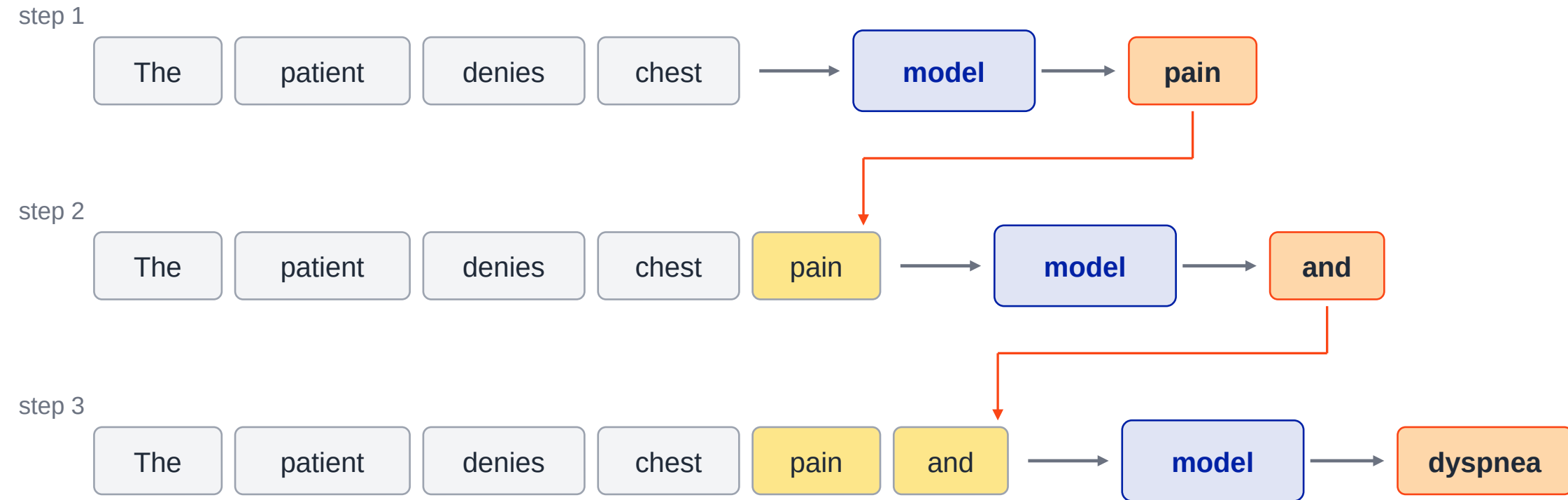
The last token's final vector goes through a linear layer (one score per vocabulary entry) and a softmax.



V = vocabulary size (about 50,000 for GPT-2; 128,000 for Llama 3)

Generate One Token at a Time

The chosen token is appended to the input, and the model runs again. This is autoregressive generation.



Greedy: take the most likely token. Sampling: draw at random (temperature sets how random).

Animation: Generating the Next Token

Animation clip: week03-clip1-next-token-generation.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

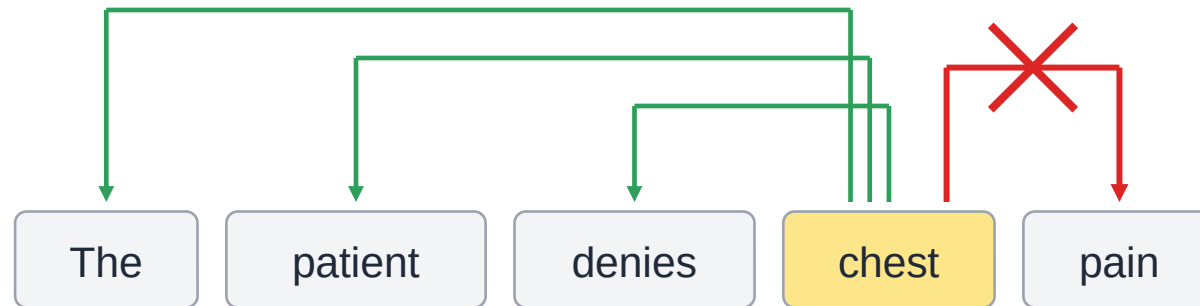


Masking and the Decoder

Training Without Cheating

In training, the whole sentence is given at once. A token must not see the tokens it is supposed to predict.

To predict the word after "chest", the model may use "chest" and the words before it ...



... but not "pain", which is the answer.

The Causal Mask

Before softmax, set every score above the diagonal to $-\infty$. Each token can attend only to itself and earlier tokens.

queries ↓	The	patient	denies	chest	pain
The	1.2	$-\infty$	$-\infty$	$-\infty$	$-\infty$
patient	0.6	1.4	$-\infty$	$-\infty$	$-\infty$
denies	0.3	1.1	1.0	$-\infty$	$-\infty$
chest	0.2	0.5	0.9	1.2	$-\infty$
pain	0.1	0.4	1.3	1.4	1.0
keys →					

Kept: itself and earlier
tokens (on and below
the diagonal)

Masked: later tokens
score = $-\infty$

After Softmax, Future Weights Are Zero

$\exp(-\infty) = 0$, so masked positions get zero weight. Each row still sums to 1 over the allowed tokens.

	The	patient	denies	chest	pain
The	1.00	0	0	0	0
patient	0.31	0.69	0	0	0
denies	0.19	0.43	0.38	0	0
chest	0.14	0.19	0.28	0.39	0
pain	0.08	0.11	0.28	0.32	0.21

"The" sees only itself:
weight 1.00

"pain" sees all five tokens
(the last row)

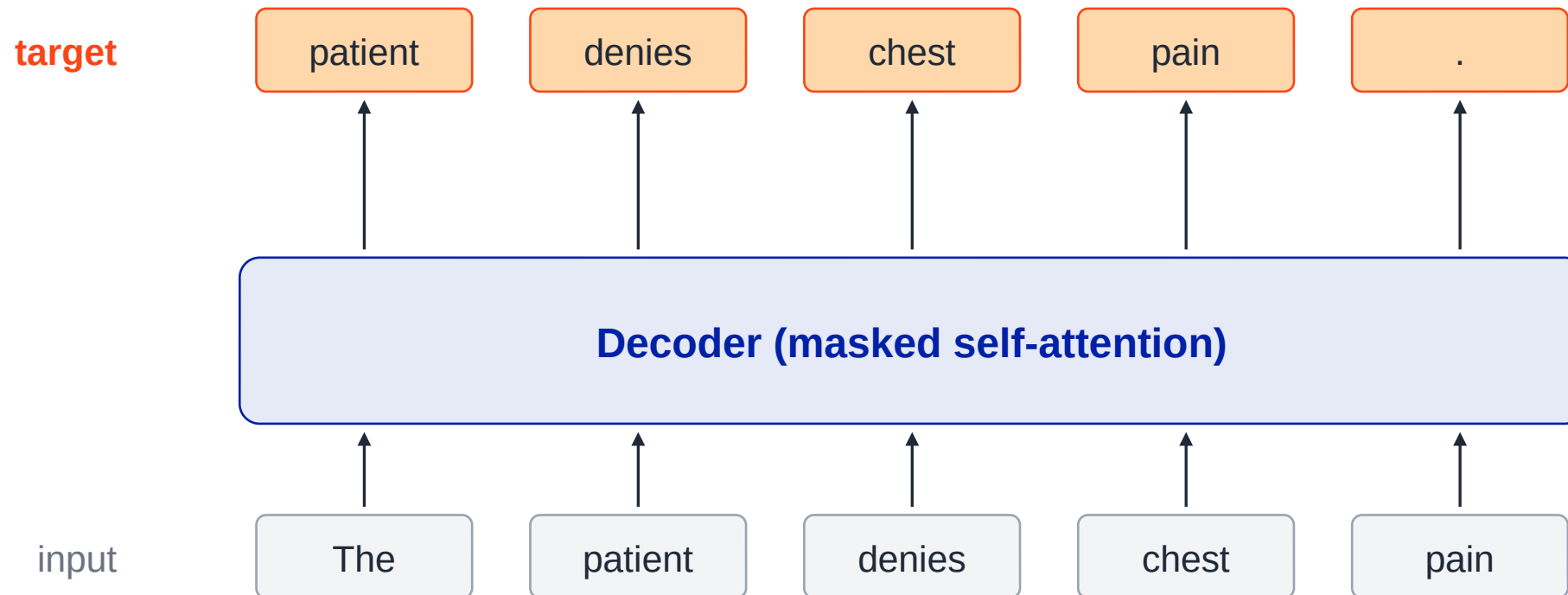
Animation: The Causal Mask

Animation clip: week03-clip2-causal-mask.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

Every Position Learns at Once

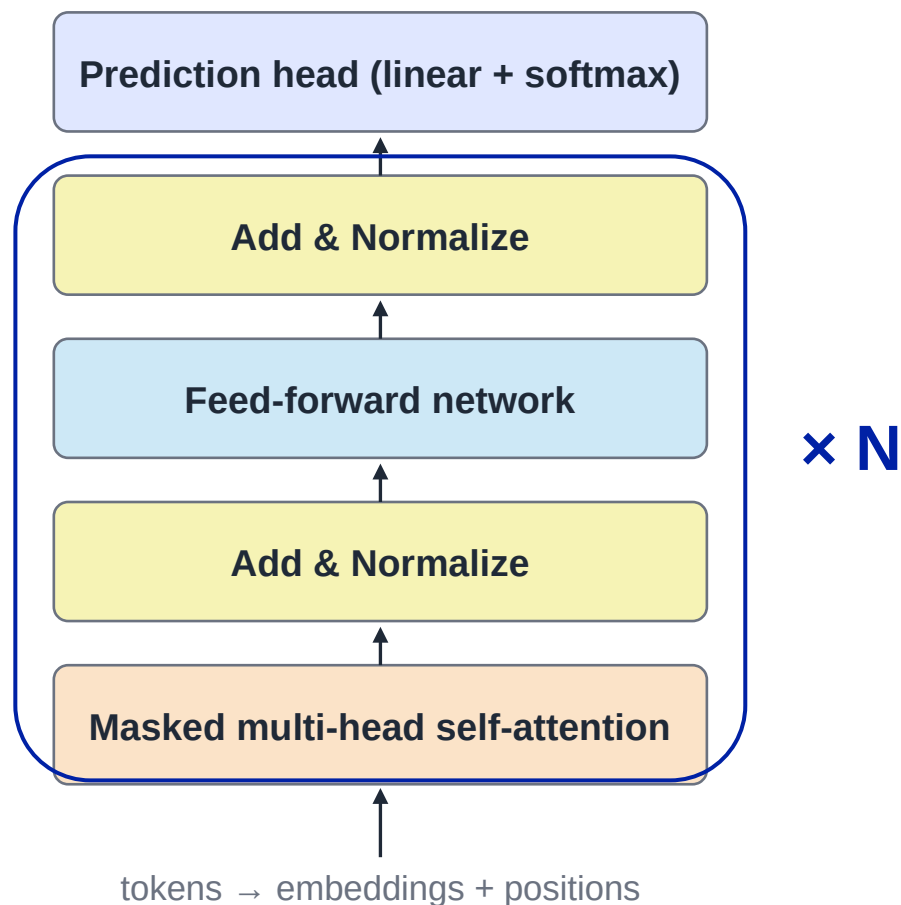
With the mask, one forward pass trains every position to predict its next token.



Target = input shifted by one position. The loss is averaged over all positions.

The Decoder Block

A decoder block is an encoder block with the causal mask added. GPT-style models stack only these blocks.



Same as the encoder block

attention, add & norm, feed-forward, add & norm

One change

the causal mask in self-attention

Examples

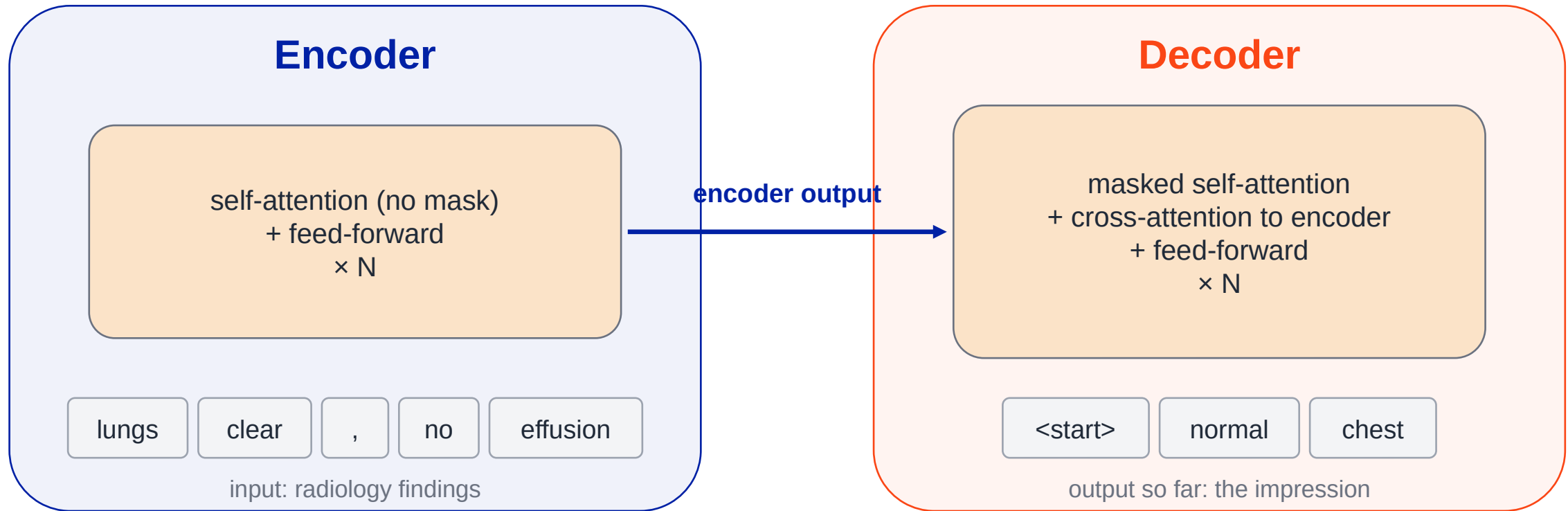
GPT-2 (12 to 48 blocks), Llama, BioGPT (24 blocks)



Encoder-Decoder Models

When the Input and Output Differ

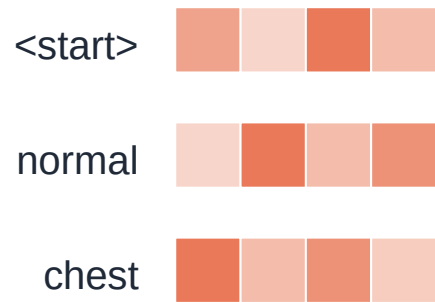
An encoder reads the whole input. A decoder writes the output, and looks back at the encoder while it writes.



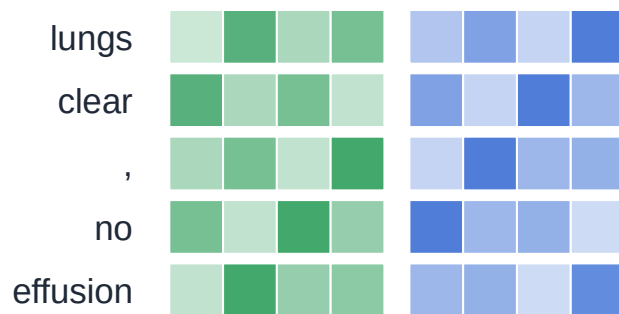
Cross-Attention

Queries come from the decoder. Keys and values come from the encoder output. This is how the input reaches the decoder.

queries from the decoder



keys and values from the encoder



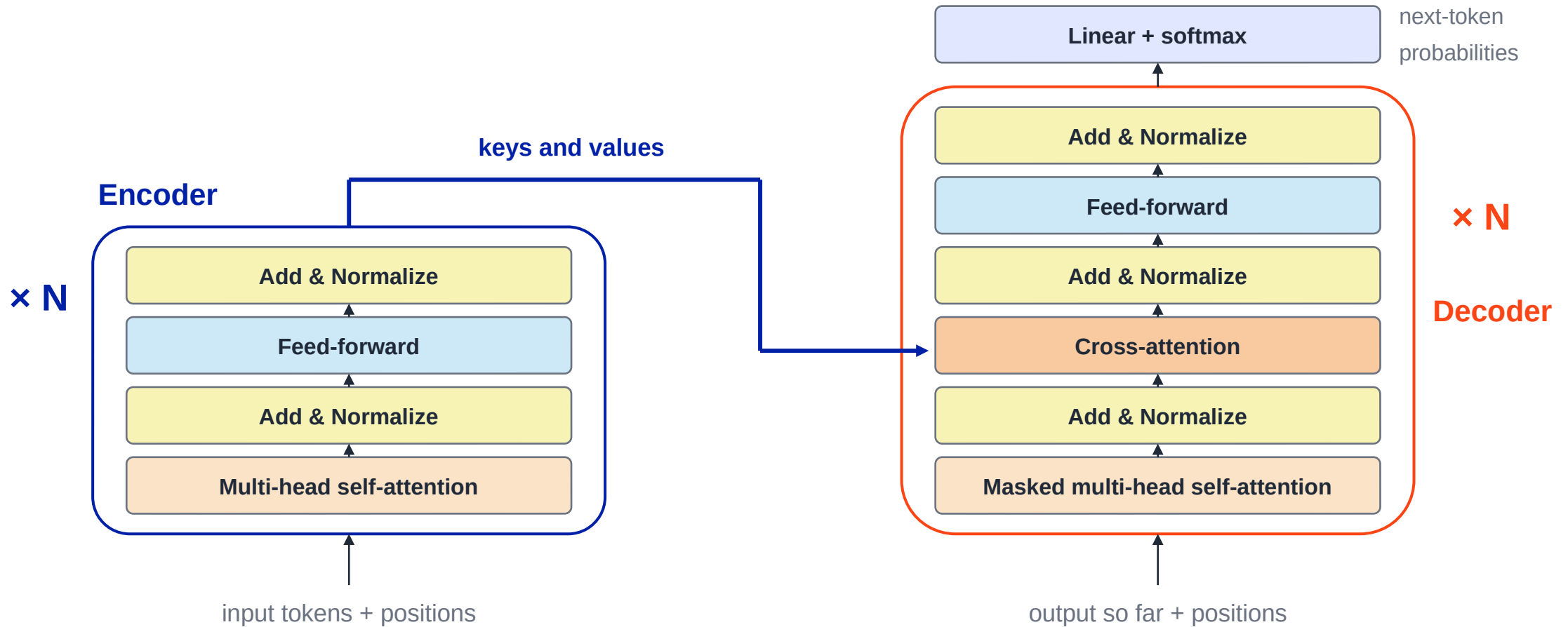
	lungs	clear	,	no	effusion
<start>	0.24	0.32	0.07	0.17	0.20
normal	0.19	0.57	0.03	0.12	0.09
chest	0.47	0.23	0.05	0.08	0.17

weights: 3 output tokens $\times$ 5 input tokens

$$\text{softmax}(Q_{\text{dec}} K_{\text{enc}}^T / \sqrt{d_k}) V_{\text{enc}}$$

Now the Original Diagram Makes Sense

Every box in the 2017 transformer is now familiar. The encoder is on the left, the decoder on the right.





Model Families

Three Families of Transformer Models

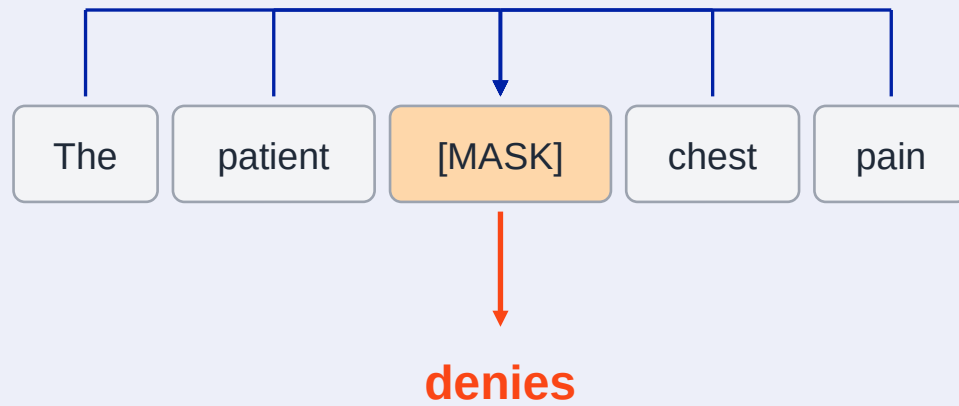
Models differ in which parts they keep, and that decides what they are good at.

	Encoder-only	Decoder-only	Encoder-decoder
Attention sees	both directions	only earlier tokens	input: both; output: earlier
Pretraining task	fill in masked tokens	predict the next token	reconstruct or translate
Good at	classification, extraction, embeddings	generation, chat, reasoning	summarization, translation
Examples	BERT, PubMedBERT, GatorTron	GPT, Llama, BioGPT	T5, BART

Two Ways to Pretrain

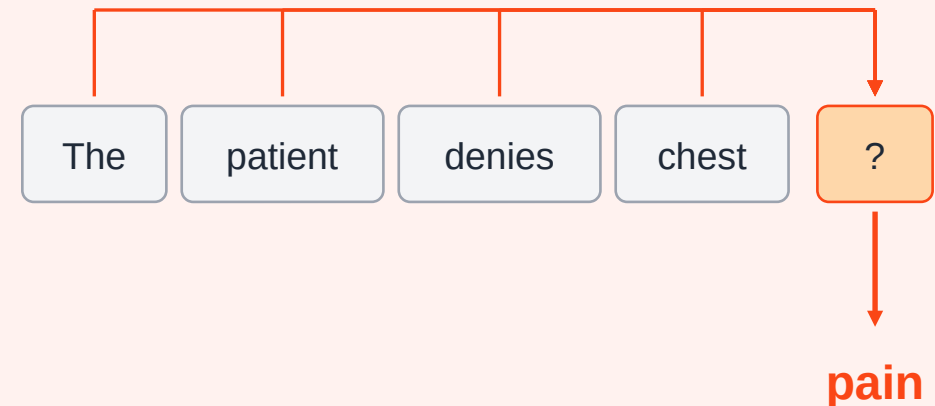
Encoders learn by filling in blanks. Decoders learn by predicting the next token. Both need only unlabeled text.

Masked language modeling (encoder)



Uses context on both sides of the blank.

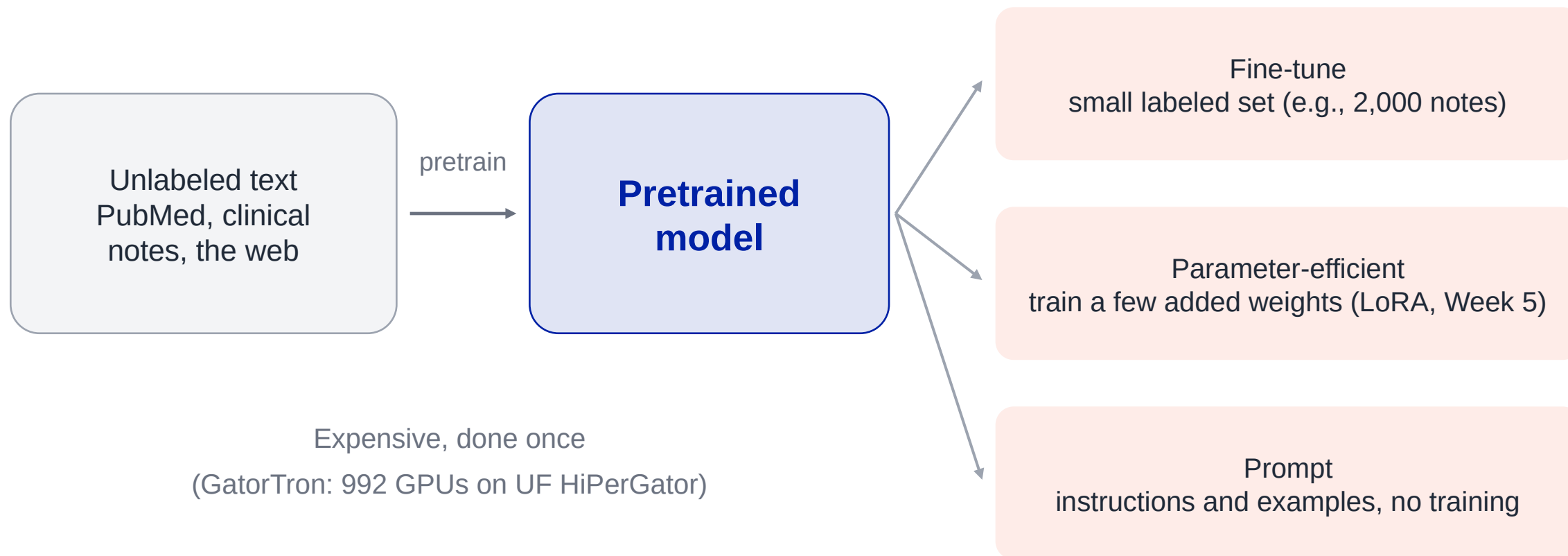
Next-token prediction (decoder)



Uses only the tokens to the left.

Pretrain Once, Adapt Many Times

Pretraining on large unlabeled text is expensive and done once. Adapting to a task needs far less data.

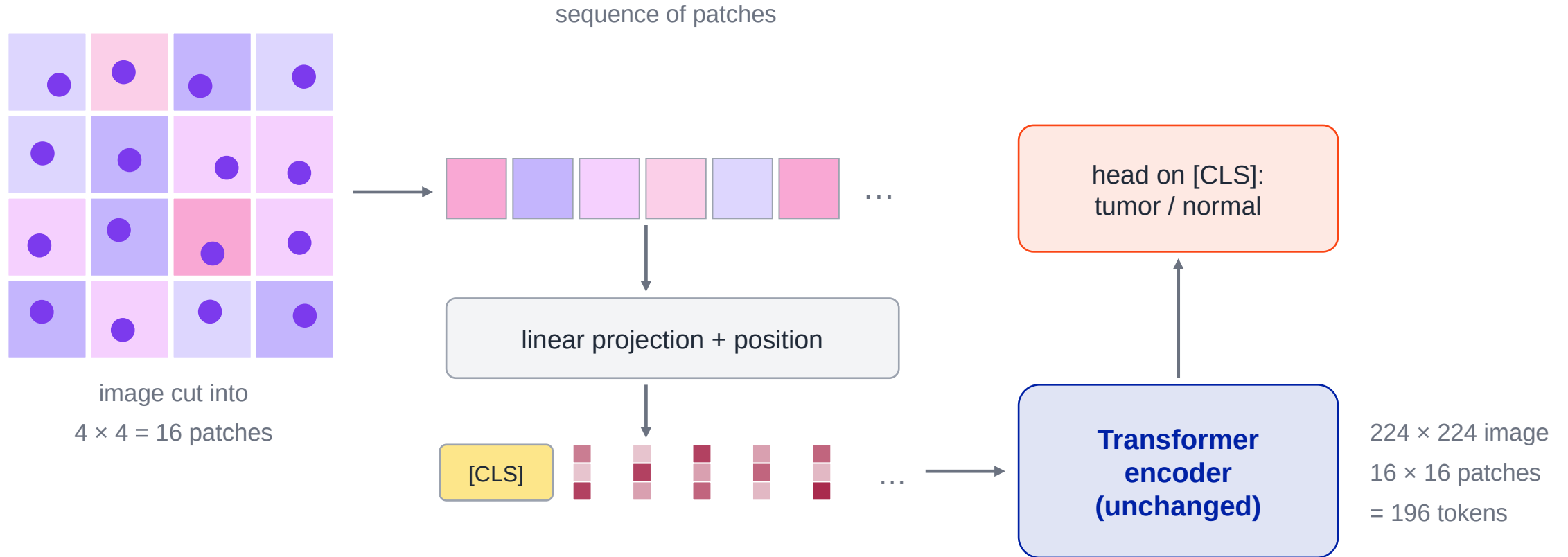


10

Beyond Text

Images as Tokens: The Vision Transformer

ViT cuts an image into patches, turns each patch into a vector, and feeds the sequence to a standard encoder.



Animation: An H&E Tile as Tokens

Animation clip: week03-clip3-vit-patches.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

Many Biomedical Data Types Can Be Tokens

Change the tokenizer and the same architecture models notes, images, records, signals, and molecules.

Clinical notes

token = word pieces

Weeks 2 and 3

Pathology and radiology

token = image patches

Week 6

Images with text

token = patches + word pieces

Week 7

EHR timelines

token = codes, labs, and events

Week 8

ECG and other signals

token = short time segments

Week 8

Proteins

token = amino acids

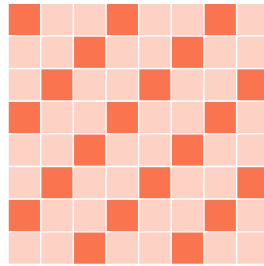
Week 11

The Cost of Attention

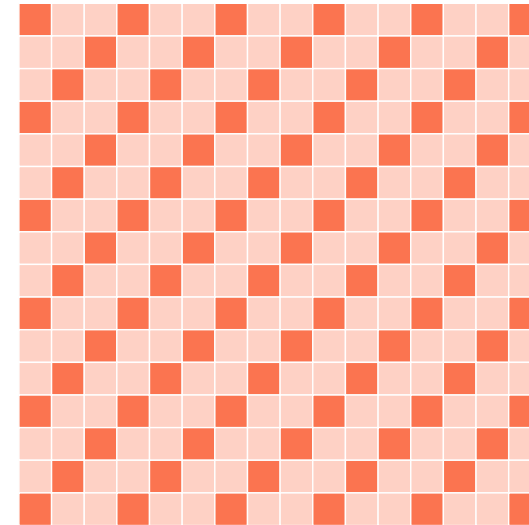
Every token attends to every token, so compute and memory grow with the square of the sequence length.



4 tokens
16 scores



8 tokens
64 scores



16 tokens
256 scores

n tokens
→
 n^2 scores

4,096-token note: about 16.8 million scores per head, per layer

Recap: Part 2

Three ideas extend last week's encoder.

Causal mask

Hide future tokens, so a decoder can be trained on whole sentences and generate one token at a time.

Cross-attention

Queries from the decoder, keys and values from the encoder: how an input text reaches the output.

Tokenizer

Change what a token is (patches, codes, signal segments) and the same encoder models new data.

Encoder-only (BERT) · decoder-only (GPT, BioGPT) · encoder-decoder (T5) · vision (ViT)

Before Thursday

Read

Luo et al., BioGPT (Briefings in Bioinformatics 2022) and Dosovitskiy et al., ViT (ICLR 2021).

Write

Reading responses on Canvas before Thursday's class.

Project

Teams are formed by Thursday, Jan 28.

Explore

Week 3 reference notebook (ungraded): masking, generation, and ViT patches in code.

Next week

Inside modern LLMs: the KV cache, GQA and MLA, sparse attention, and mixture-of-experts.