



CS 639: Foundation Models **Architectures I**

Fred Sala

University of Wisconsin-Madison

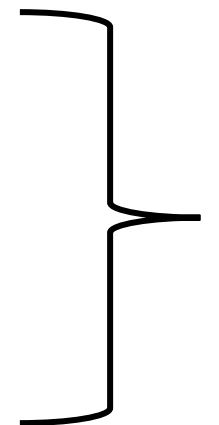
Feb. 17, 2026



Announcements

- Midterm: **March 11, 5:40 pm - 7:20 pm**
 - Location: **Ingraham Hall, Room B10**
- Homework 1: out!
- **Resources**
 - <https://jalammar.github.io/illustrated-transformer/>
- **Class roadmap:**

Tuesday Feb. 17	Architectures: Encoder-Only
Thursday Feb. 19	Architectures: Others
Tuesday Feb. 24	Attention Variants
Thursday Feb. 26	Multimodal Architectures



LLMs, FMs and Arch

Outline

- **Finish up last time: full Transformer architecture**
 - Encoder layer, decoder layer, full original Transformer architecture
- **Encoder-only models**
 - Example: BERT, architecture, multitask training, fine-tuning
- **Preview of decoder-only models**
 - GPT architecture, examples

Outline

- **Finish up last time: full Transformer architecture**

- Encoder layer, decoder layer, full original Transformer architecture

- **Encoder-only models**

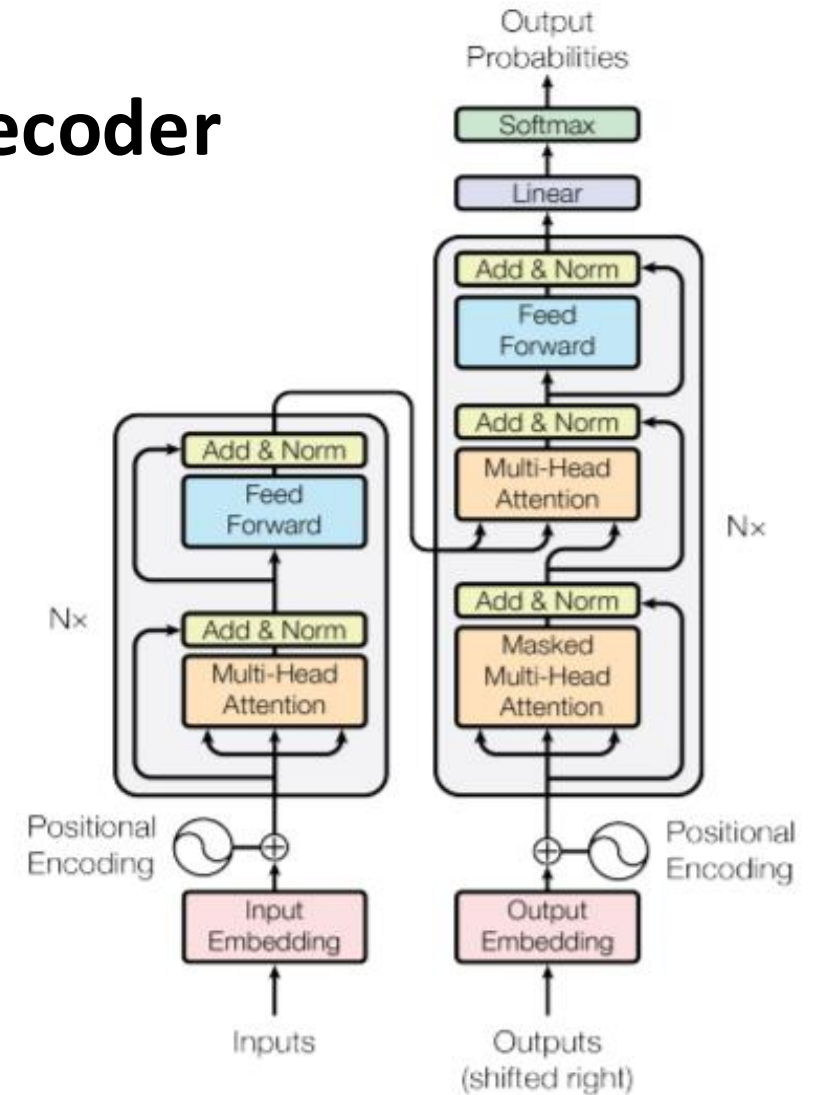
- Example: BERT, architecture, multitask training, fine-tuning

- **Preview of decoder-only models**

- GPT architecture, examples

Transformers: Model Architecture

- Initial goal for an architecture: **encoder-decoder**
 - Get **rid of recurrence**
 - Replace with **self-attention**
- Architecture
 - You may have seen this picture
 - Centered on self-attention blocks



Interlude: Encoder-Decoder Models

- Translation tasks: natural encoder-decoder architecture
- Intuition:

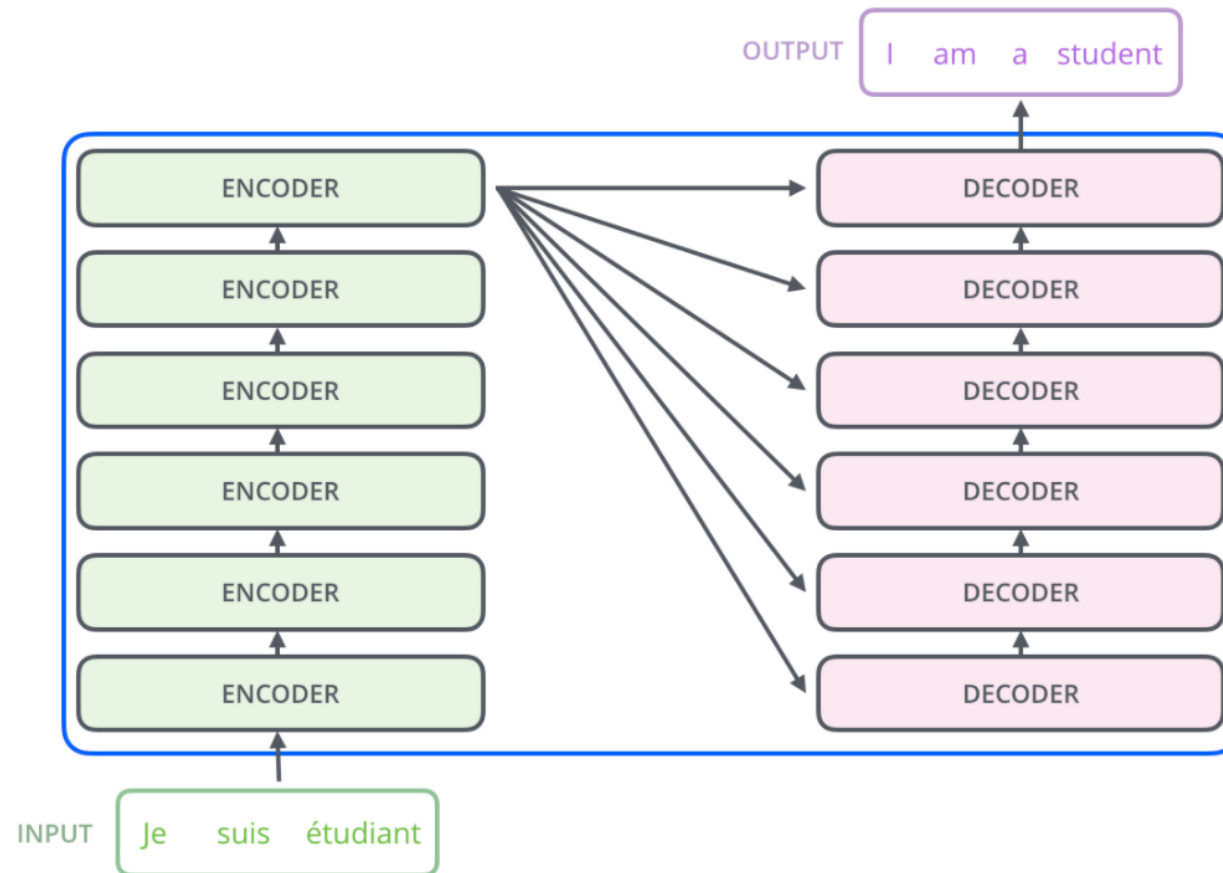
...again, Mummie, there wouldn't be any railway late and we shouldn't
oms. Oh, *do* let us go in a caravan." "I know it sounds lovely, darling; but
Mrs. Russell shook her head. "I know it sounds lovely, darling; but
we to get a caravan? It would cost at least fifty pounds to buy one,
en if we had one, Daddy couldn't get away this summer. No, we
ke up our minds to do without a holiday this year; but I'll tell you what
ll do: we'll all go to Southend for the day, as we did last year, and
r lunch and tea with us and have a splendid picnic."
"Then we can bathe again," said Bob; "but, oh! I do wish I could ha
ny and ride," he added unexpectedly. "You don't know how I long
ny," he continued, sighing deeply as he remembered the blissful holi
en a friend let him share his little Dartmoor pony and ride occasional
"Southend is nothing but houses and people," cried Phyllis; "it's no b
an this place; and oh! Mummie, I do so *long* for fields and flowers
imals," she added piteously; and she shook her long brown hair forw
hide the tears in her eyes.
"Never mind, darling, you shall have them one day," answered
assell with easy vagueness.
This really was not very comforting, and it was the most fortunate thing
st at that moment a car stopped at the door.
"Uncle Edward!" shouted Bob, rushing from the room. Phyllis br
e tears so hastily from her eyes that she arrived at the front door almo
on as he did, and both flung themselves on the tall, kindly-looking man st
g beside the car.
"Uncle Edward! Uncle Edward!" they cried. "You've come at
e've been longing to see you. Oh, how glad we are you're here!"
Now the delightful thing was that their uncle seemed just as pleased to
em as they were to see him, and returned their hugs and greetings with
most cordiality. They were just on the point of dragging him into
use, hanging one on each arm, when he said: "Stop, not so fast. There
me things to fetch in from the car."
So saying he began diving into the back of it and bringing out, not on
itcase, but various parcels, which he handed out one by one.
"That's the pair of chickens I've brought for your mother," said he, br



...à 4 heures du matin, en descendant
après avoir été sur les quais. Elle était couronné
côté, dans le coin de sa cage, les pattes
Comme si elle courait dans son sommeil.
La dissection montre que mes prédictions
justes. Comparé à un cerveau normal, celui d'Alg
a diminué de poids et montre un effacement
des circonvolutions cérébrales ainsi qu'un
et un élargissement des scissures.
C'est épouvantable de penser que la même
m'arrive peut-être à moi, en ce moment. L'avoir
produire chez Algernon rend cette menace réelle.
la première fois, je suis effrayé de l'avenir.
J'ai mis le corps d'Algernon dans une petite bo
métal et je l'ai emporté à la maison avec moi. Je n
pas les laisser le jeter dans l'incinérateur. C'est
sentimental mais tard hier soir, je l'ai enterrée
cour de derrière. J'ai pleuré en mettant un bouquet
fleurs sauvages sur la tombe.

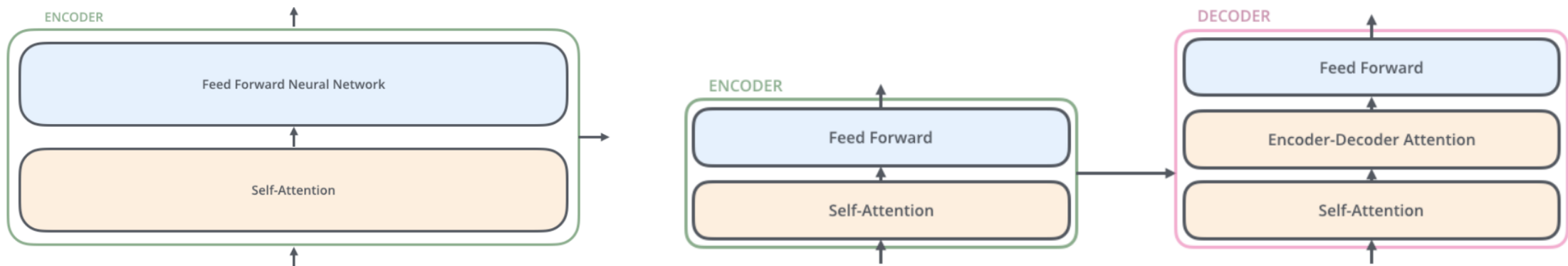
Transformers: Architecture

- **Sequence-sequence** model with **stacked** encoders/decoders:
 - For example, for French-English translation:



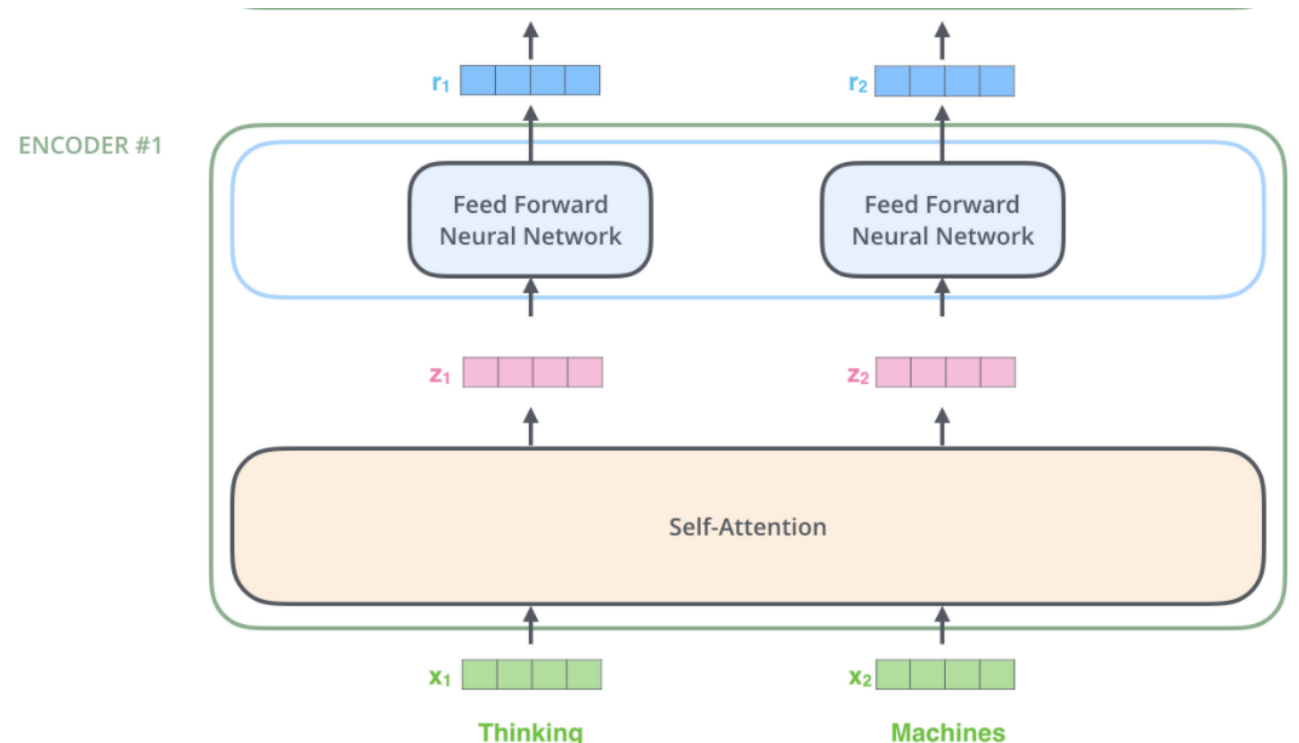
Transformers: Architecture

- Sequence-sequence model with **stacked** encoders/decoders:
 - What's inside each encoder/decoder unit?
- Focus encoder first: **pretty simple!** 2 components:
 - Self-attention block
 - Fully-connected layers (i.e., an MLP)



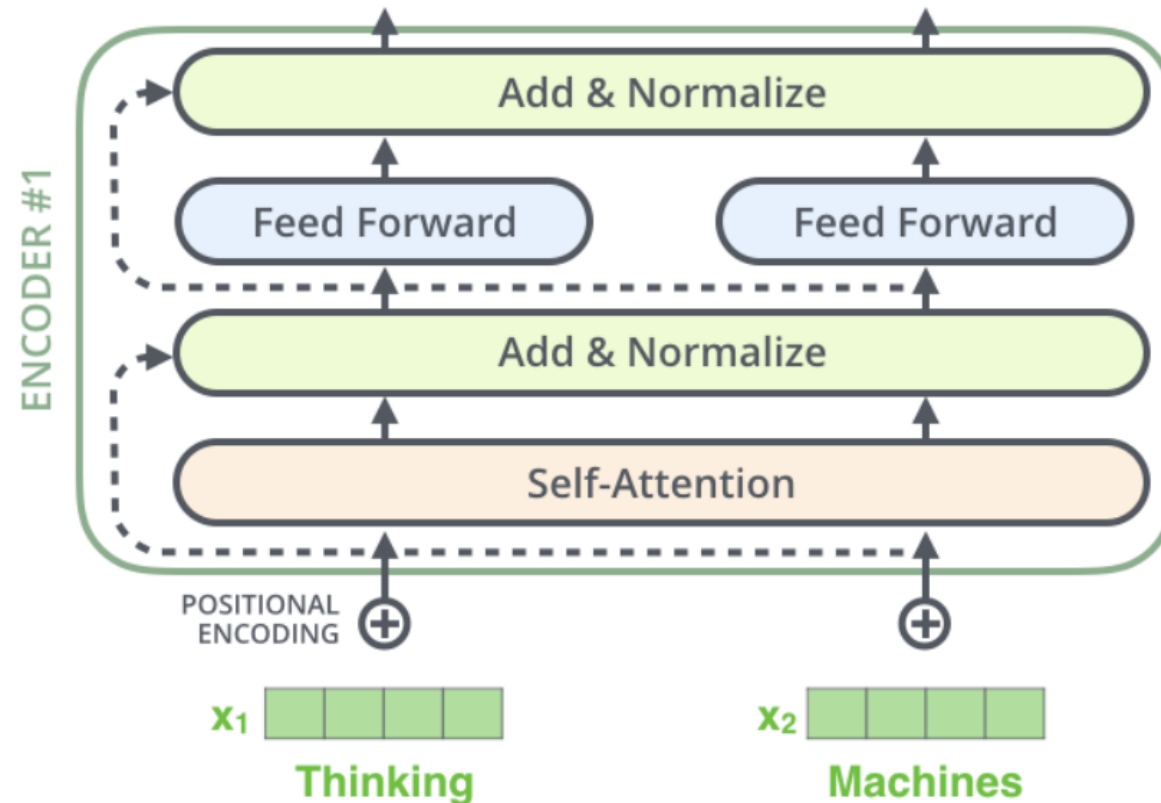
Transformers: Inside an Encoder

- Let's take a look at the encoder. Two components:
 - 1. **Self-attention** layer (covered this)
 - 2. “Independent” **feedforward nets**
 - **Note:** same MLP (often 2-layer) at every position



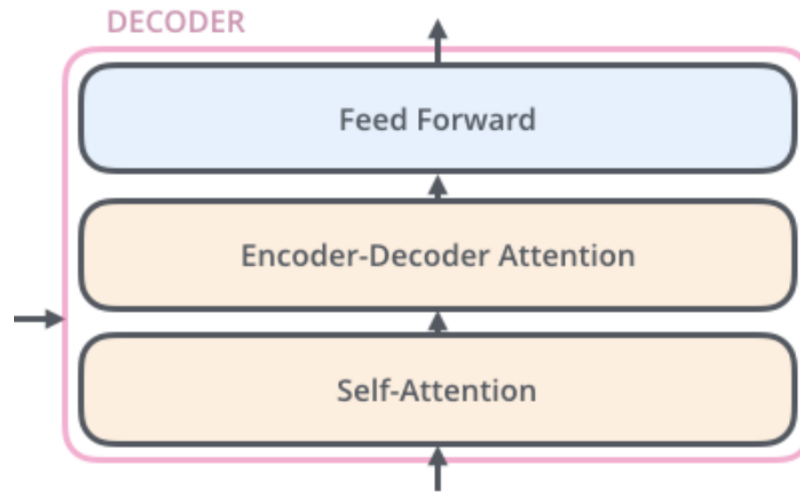
Transformers: More Tricks

- Recall a big innovation for ResNets: residual connections
 - And also layer normalizations
 - Apply to our encoder layers



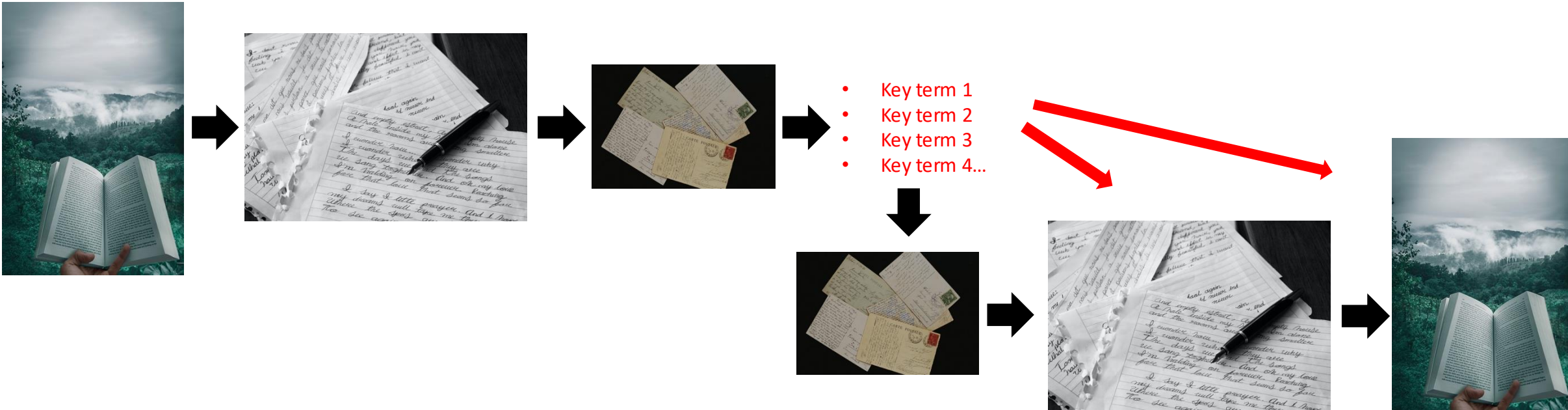
Transformers: Inside a Decoder

- Let's take a look at the decoder. Three components:
 - 1. **Self-attention** layer (covered this)
 - 2. Encoder-decoder attention (same, but K, V come from encoder)
 - 3. “Independent” **feedforward nets**



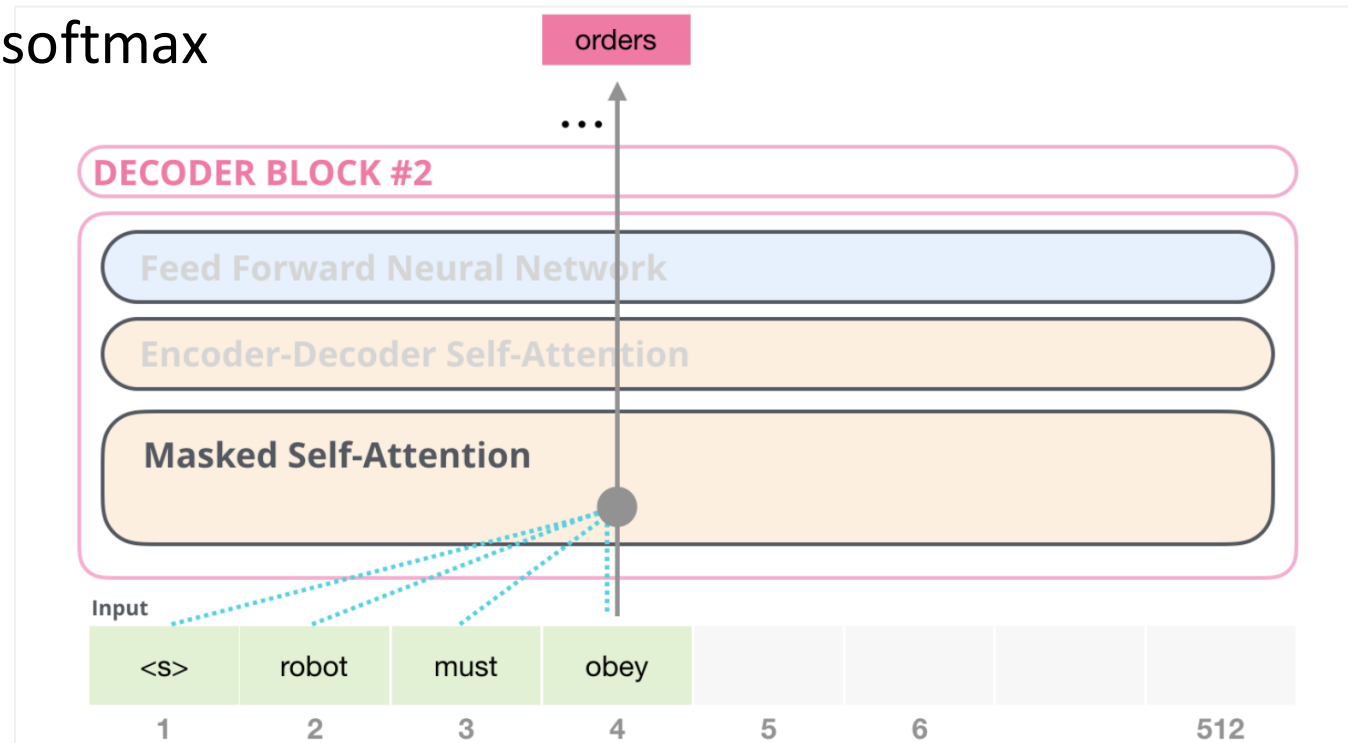
Transformers: Cross-Attention

- Why encoder-decoder attention ?
 - Recall: same as before, but K, V come from encoder
 - Actually more traditional, but... **intuition:**



Transformers: Decoder Masking

- One more interesting bit!
 - At the decoder level, self-attention changes a bit:
 - Masked instead: block *future* words from being attended to
 - **Important in training** (if we don't---model could look at future tokens during training, **but this is impossible in inference**)
 - **How to mask?** Add $-\infty$ before softmax



Transformers: Last Layers

- Next let's look at the end. Similar to a CNN,

- 1. Linear layer
- 2. Softmax

Get probabilities of words

Which word in our vocabulary is associated with this index?

Get the index of the cell with the highest value (argmax)

am

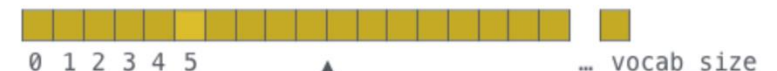
5

log_probs



Softmax

logits



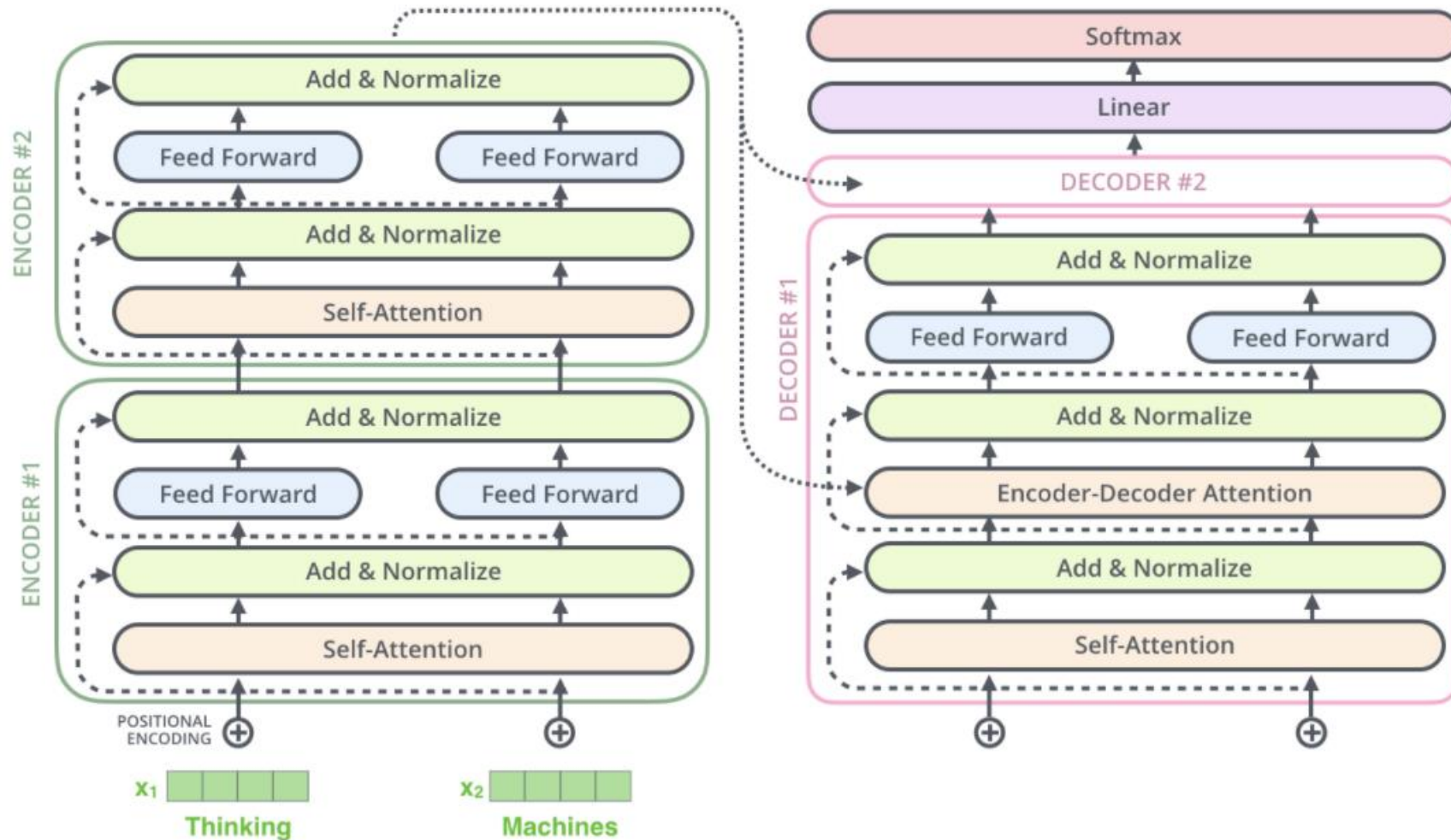
Linear

Decoder stack output



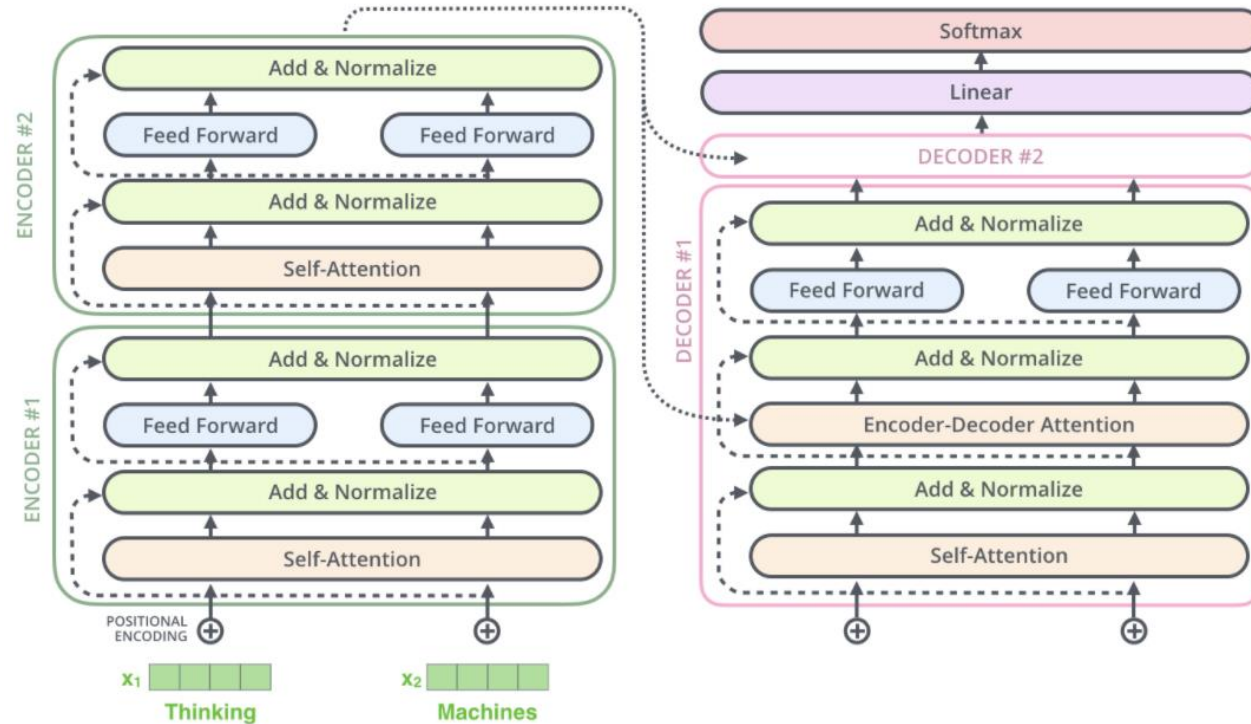
Transformers: Putting it All Together

- What does the full architecture look like?



Transformers: Training

- Data: standard datasets (WMT English-German)
 - Note: **supervised task**. Soon: switch to self-supervised
 - ~5 million pairs of sentences for this dataset
 - Training procedure not special: cross-entropy loss, Adam optimizer



Outline

- **Finish up last time: full Transformer architecture**
 - Encoder layer, decoder layer, full original Transformer architecture
- **Encoder-only models**
 - Example: BERT, architecture, multitask training, fine-tuning
- **Preview of decoder-only models**
 - GPT architecture, examples

Why Encoder-Decoder?

Wanted two things for translation:

- 1) **Outputs** in natural language
- 2) Tight alignment with **input**

What happens if we relax these?

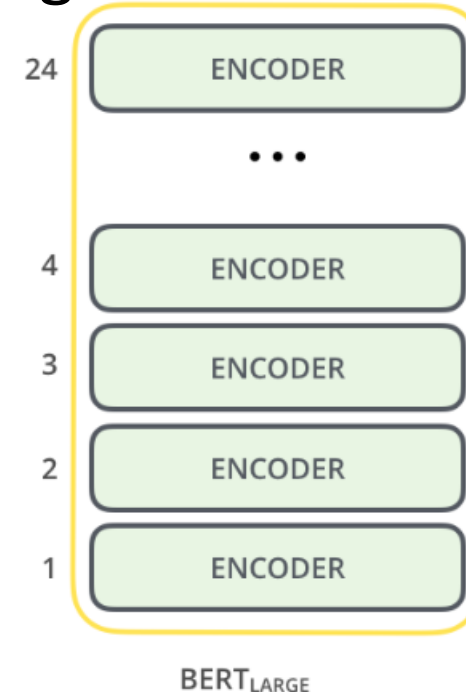
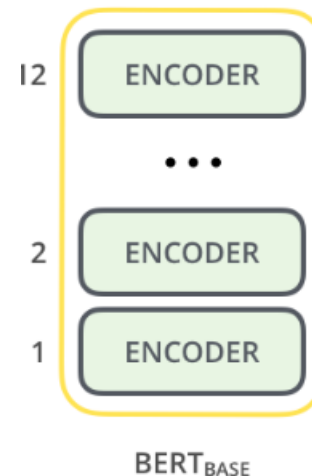
1. Encoder-only models
2. Decoder-only models



Encoder-Only Models: BERT

Let's get rid of the first part

- 1) **Outputs** in natural language
- 2) Tight alignment with **input**
- So **not** a generative model → get representations
 - Like we talked about in self-supervised learning
- Rip away decoders
 - Just stack encoders



Interlude: Contextual Embeddings

Q: Why is it called “BERT”?

- A: In a sense, follows up ELMo

• Story:

- **2013**: “Dense” word embeddings (**Word2Vec**, **Glove**)
- Downside: fixed representations per word
 - “Bank”: building or riverside?
- Need: contextual representations
 - Using language model-like techniques
 - 2018: ELMo, BERT
 - ELMo: uses LSTMs, BERT uses transformers



Highlights

1. Nearest neighbors

The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words. Sometimes, the nearest neighbors according to this metric reveal rare but relevant words that lie outside an average human's vocabulary. For example, here are the closest words to the target word *frog*:

0. *frog*
1. *frogs*
2. *toad*
3. *litoria*
4. *leptodactylidae*
5. *rana*
6. *lizard*
7. *eleutherodactylus*



3. *litoria*



4. *leptodactylidae*



5. *rana*



7. *eleutherodactylus*

<https://nlp.stanford.edu/projects/glove/>

Interlude: Contextual Embeddings

Q: Why is it called “BERT”?

- A: In a sense, follows up ELMo

BERT acronym:

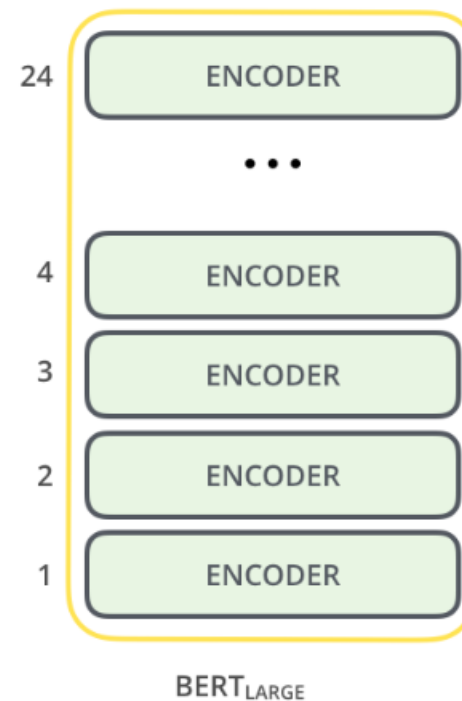
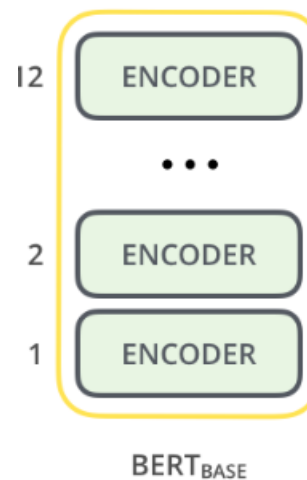
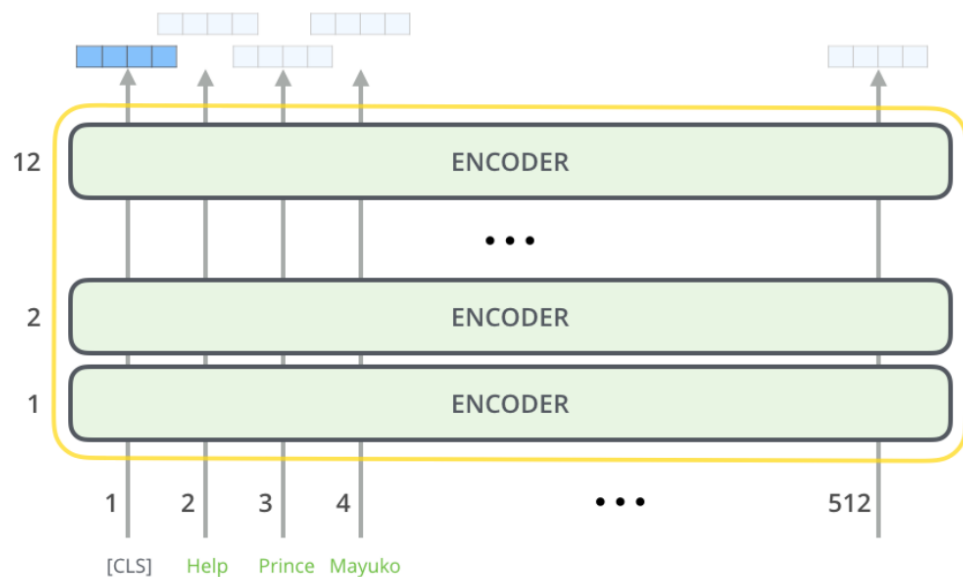
- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers.
- ERT should make sense,
- Bidirectional: no causal masks, look at both sides of a word!
- Captured in self-attention block



BERT: Forward Pass

BERT architecture

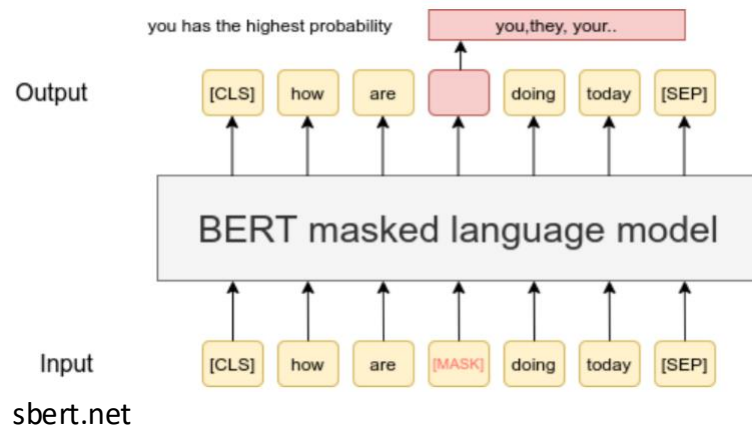
- Rip away decoders
 - Just stack encoders



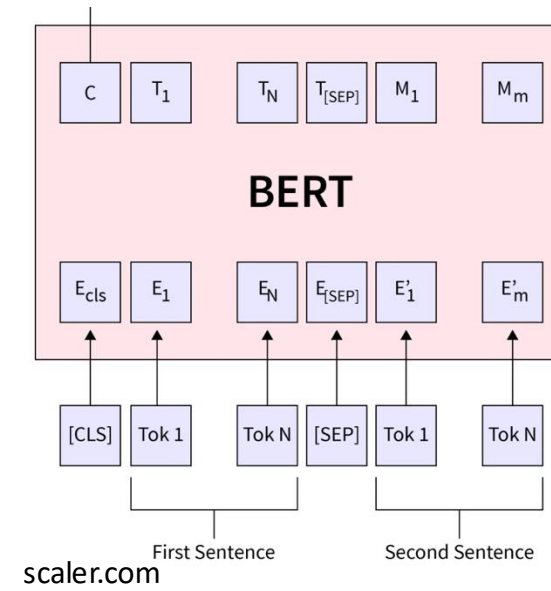
BERT: Training

Training is more interesting!

- Pretraining. Then fine-tuning on task of interest
- Back to **self-supervised learning**!
- Two tasks for **pretraining**.



1. Masked Language Modeling

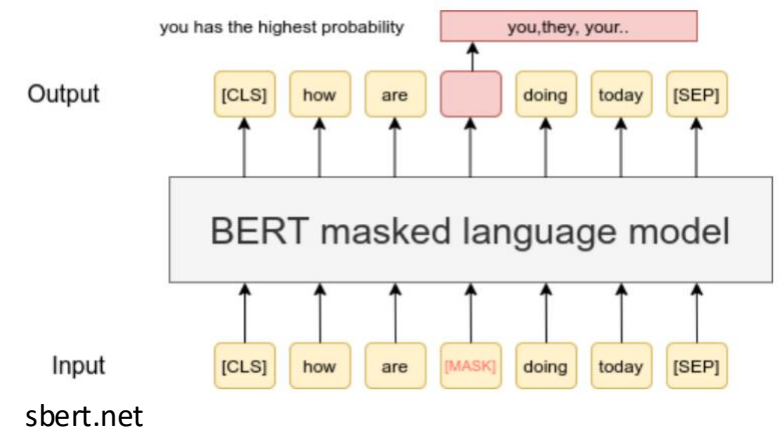


2. Next Sentence Prediction

BERT: Training Task 1

Masked Language Modeling Task

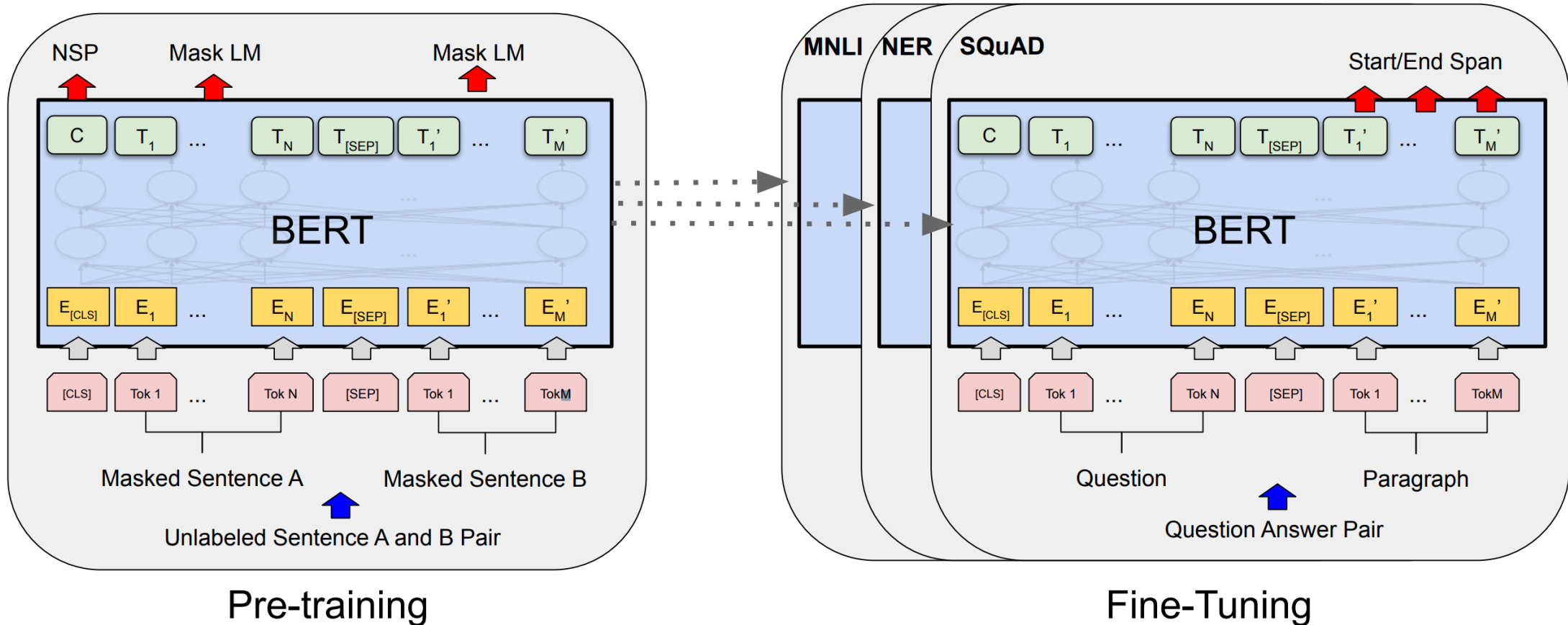
- Use [MASK] token for word to be predicted
- Which words to mask?
 - Original paper: 15% of words at random
 - But... of these
 - 10% of the time, no [MASK], flip word randomly
 - 10% of the time leave word unchanged



BERT: Training

Training is more interesting,

- Pretraining. Then fine-tuning on task of interest



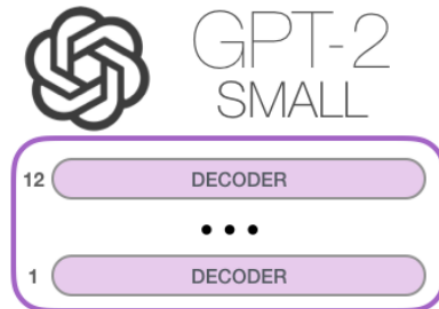
Outline

- **Finish up last time: full Transformer architecture**
 - Encoder layer, decoder layer, full original Transformer architecture
- **Encoder-only models**
 - Example: BERT, architecture, multitask training, fine-tuning
- **Preview of decoder-only models**
 - GPT architecture, examples

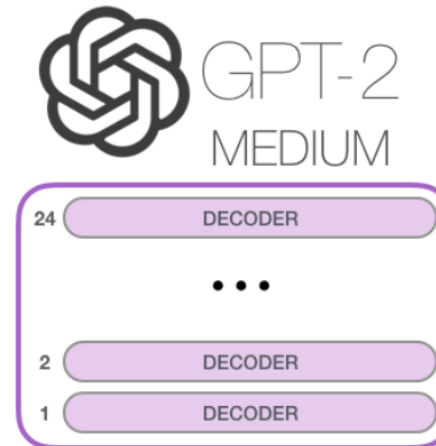
Decoder-Only Models: GPT

Let's get rid of the first part

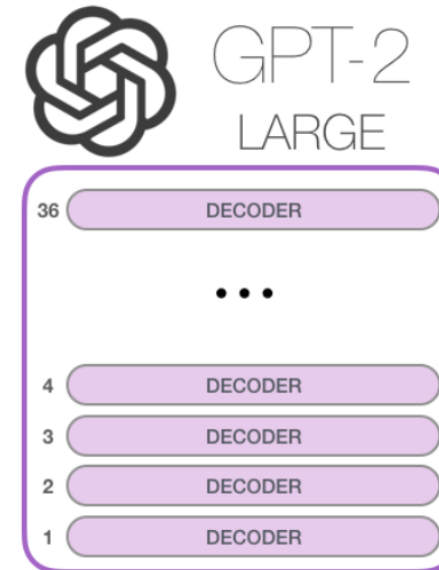
- 1) **Outputs** in natural language
 - 2) Tight alignment with **input**
-
- Rip away encoders
 - Just stack decoders



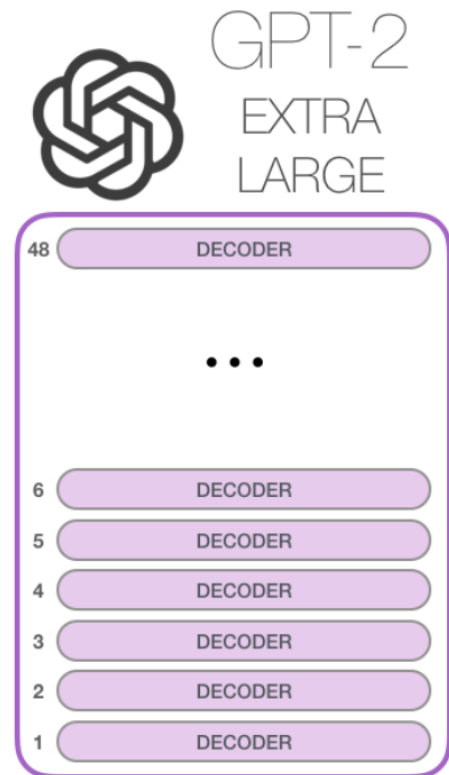
Model Dimensionality: 768



Model Dimensionality: 1024



Model Dimensionality: 1280



Model Dimensionality: 1600

Decoder-Only Models: GPT

Rip away encoders

- Just stack decoders
- Use causal masking! NB: not a *mask token* like in BERT
- Training: **next-token prediction**

