Discussion Mini Lecture 11

More on Transformers

The Complete (Original) Transformer Architecture

CS 189/289A, Fall 2025 @ UC Berkeley
Sara Pohland

Concepts Covered

1. Transformer Encoder
2. Transformer Decoder
3. Training a Transformer
4. Using a Transformer for Inference

A Running Example: Machine Translation

Task: Translate sentence from English to Spanish

Input: Where is the library?

Target: ¿Dónde está la biblioteca?

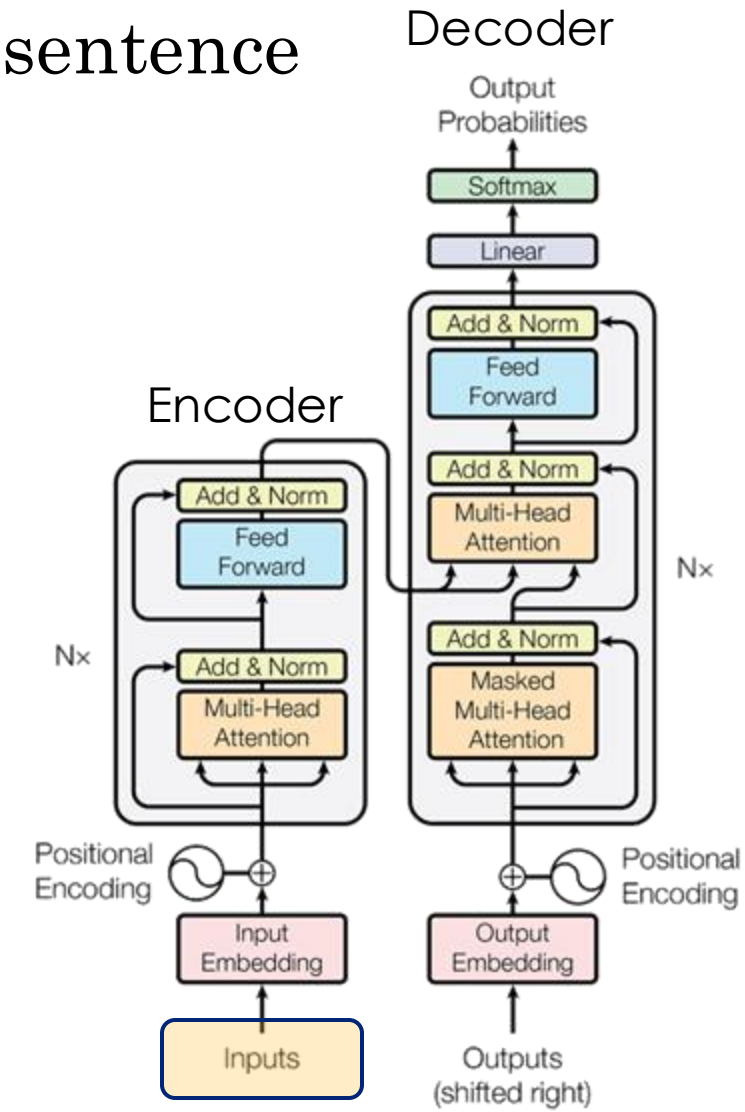
The Encoder of a Transformer

1. Transformer Encoder
2. Transformer Decoder
3. Training a Transformer
4. Using a Transformer for Inference

The Transformer Encoder

Goal: Learn compact representation of English sentence

Input: Where is the library?



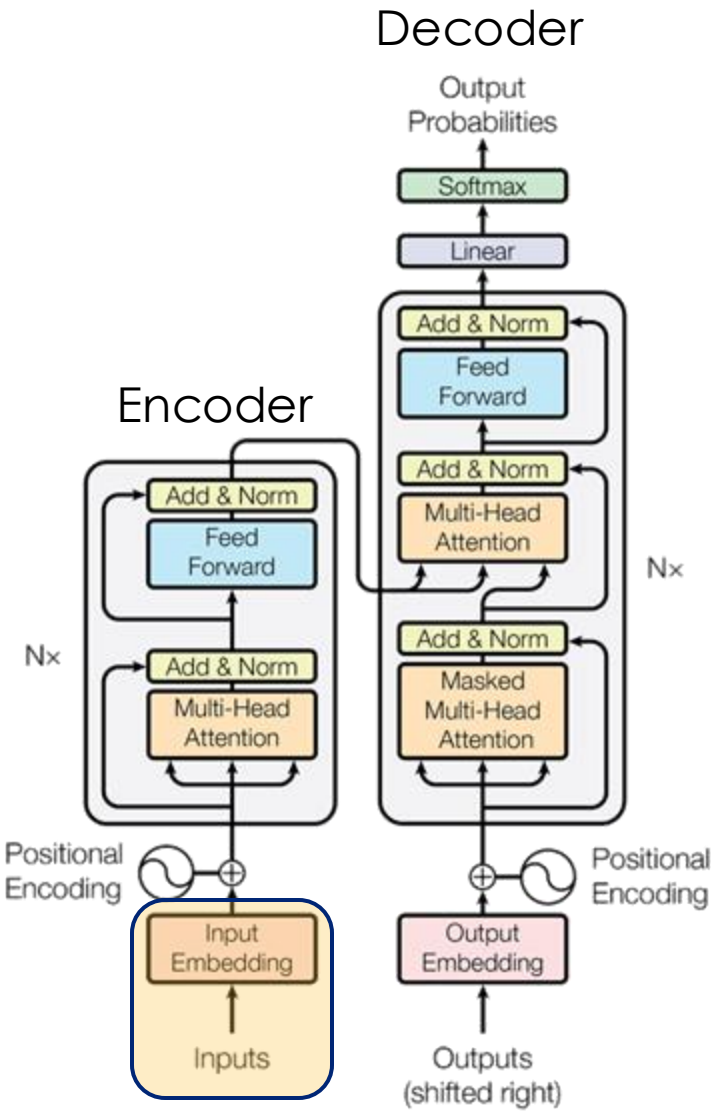
1) Obtain Input Embeddings

Input: Where is the library?

121	95	55	67	23
⋮	⋮	⋮	⋮	⋮
205	11	216	128	102
196	212	43	58	205
⋮	⋮	⋮	⋮	⋮
25	38	18	111	95

$x_1 \in \mathbb{R}^D$ ... $x_5 \in \mathbb{R}^D$

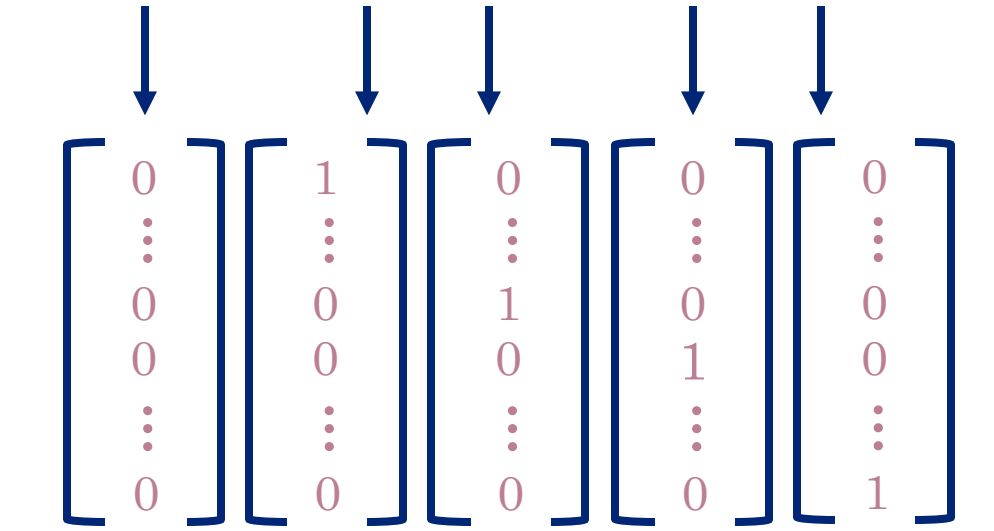
N = 5 input embeddings capturing word (part)



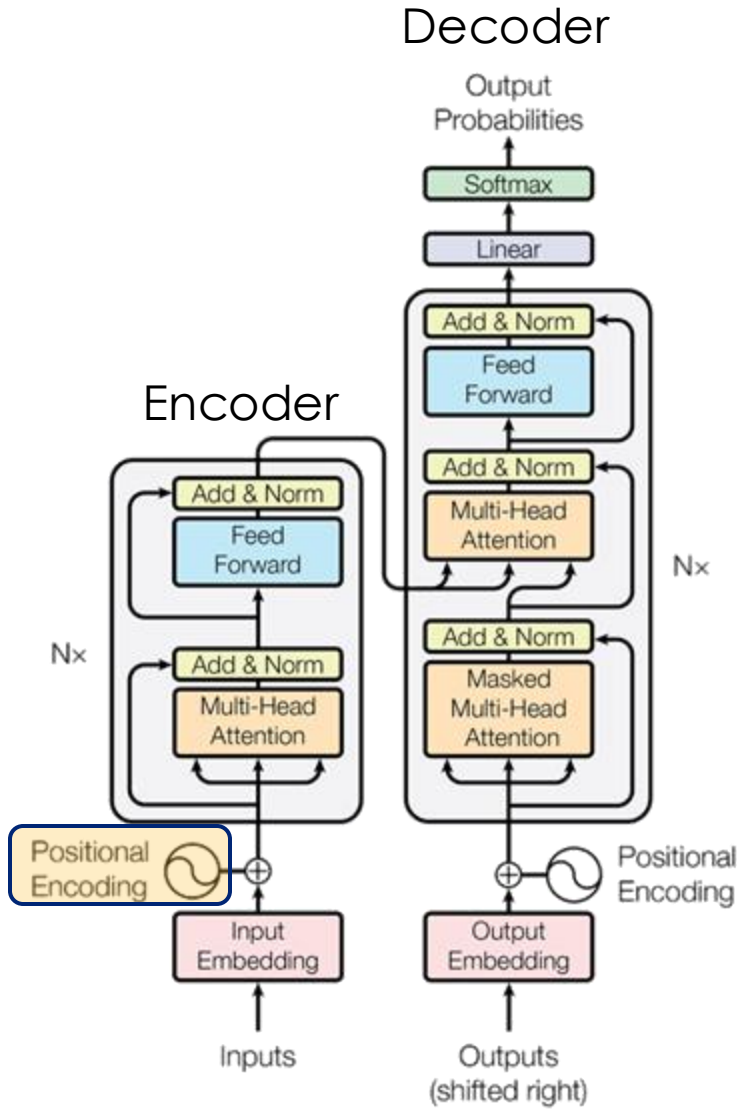
2) Grab Positional Encodings

We'll discuss this in Problem 1 on the Disc. 11 worksheet.

Input: Where is the library?



N = 5 positional encodings capturing position in sequence



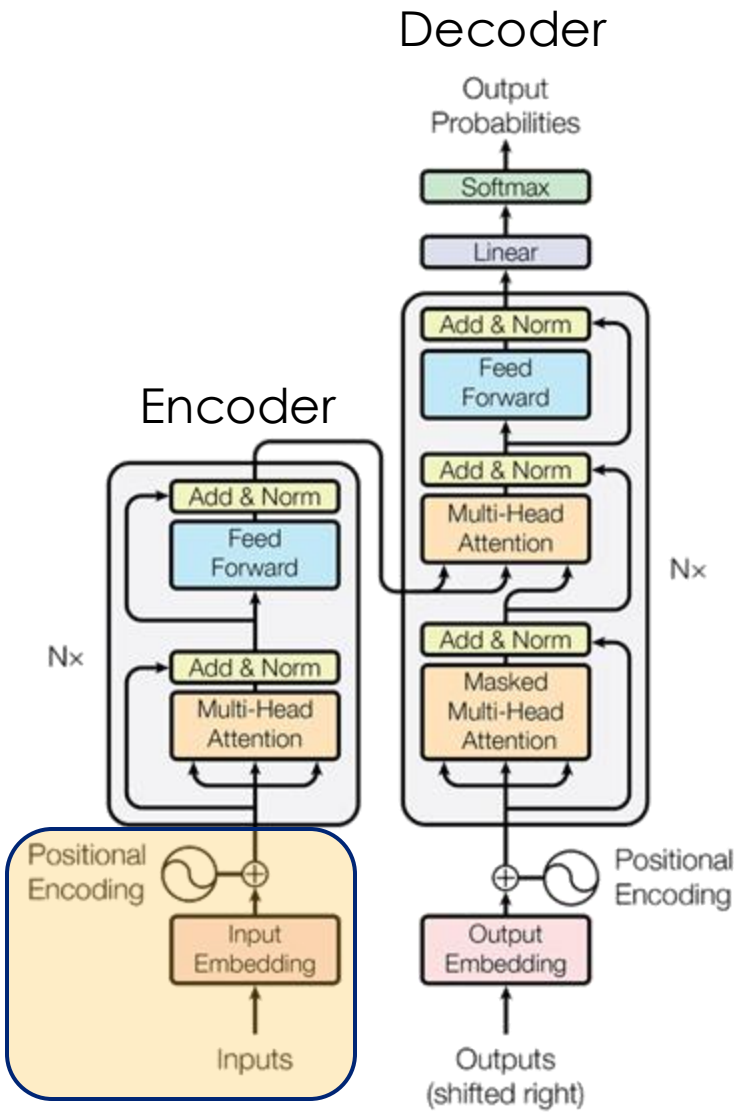
3) Combine Embeddings + Positions

Input: Where is the library?

$$\begin{bmatrix} 121 \\ \vdots \\ 205 \\ 196 \\ \vdots \\ 25 \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 121 \\ \vdots \\ 205 \\ 196 \\ \vdots \\ 25 \end{bmatrix}$$

$x_1 \in \mathbb{R}^D$ $r_1 \in \mathbb{R}^D$ $\tilde{x}_1 \in \mathbb{R}^D$

N = 5 encoder inputs
(word + position)



4) Generate Queries, Keys, and Values

Input: Where is the library?

Query: Question token asks rest of sequence

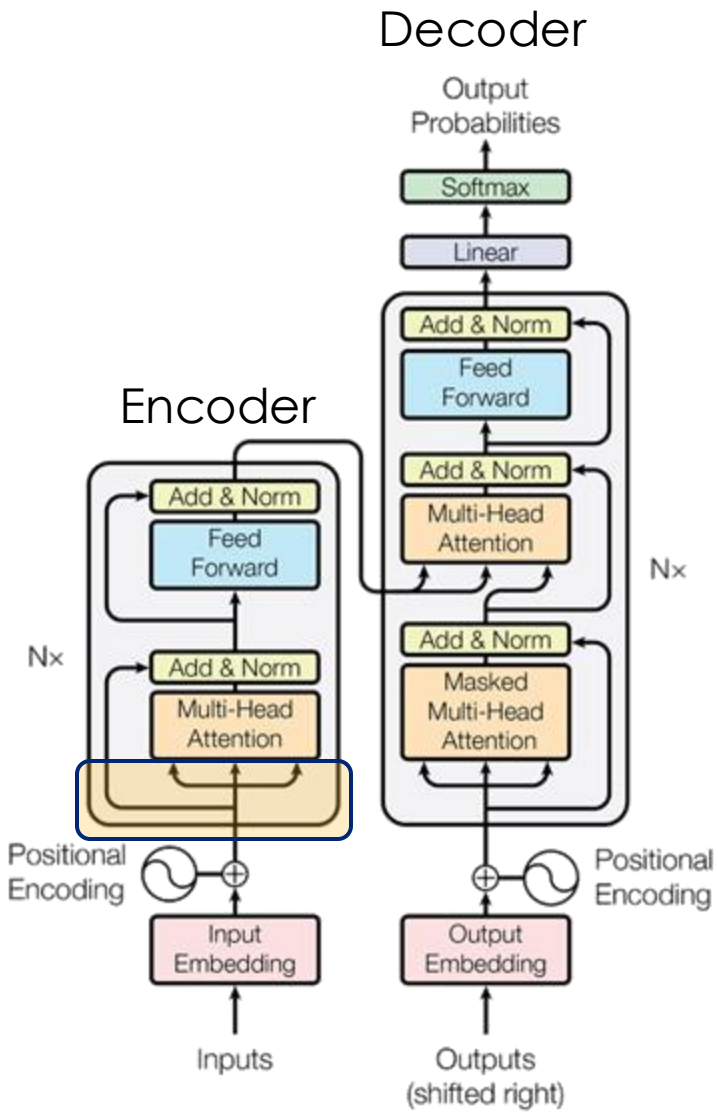
$$\boldsymbol{q}_1 = (\mathbf{W}^{(q)})^\top \tilde{\boldsymbol{x}}_1 \in \mathbb{R}^{D_k}$$

Key: Attributes token exposes to be searched

$$\boldsymbol{k}_1 = (\mathbf{W}^{(k)})^\top \tilde{\boldsymbol{x}}_1 \in \mathbb{R}^{D_k}$$

Value: Info token provides if deemed relevant

$$\boldsymbol{v}_1 = (\mathbf{W}^{(v)})^\top \tilde{\boldsymbol{x}}_1 \in \mathbb{R}^{D_v}$$



5) Get Contextualized Embeddings

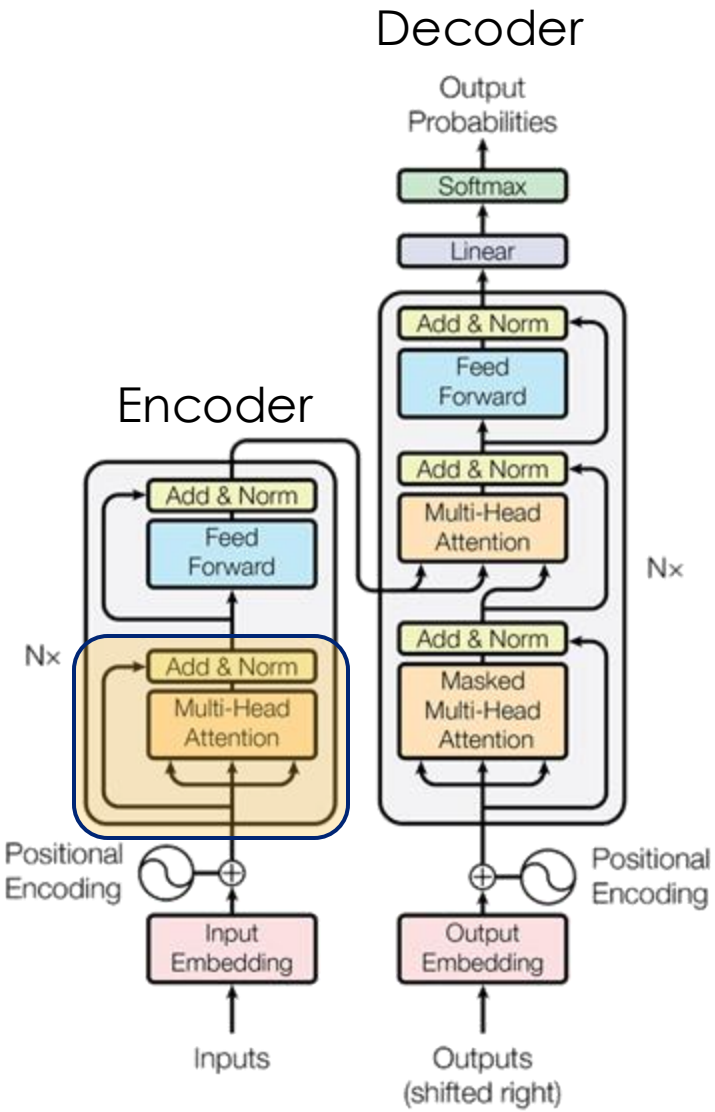
Input: Where is the library?

Contextualized embedding: Meaning of token in context of all other input tokens

$$y_1 = \tilde{x}_1 + \sum_{i=1}^{N} \alpha_{1i} v_i$$

attention (relevance of token i to token 1)

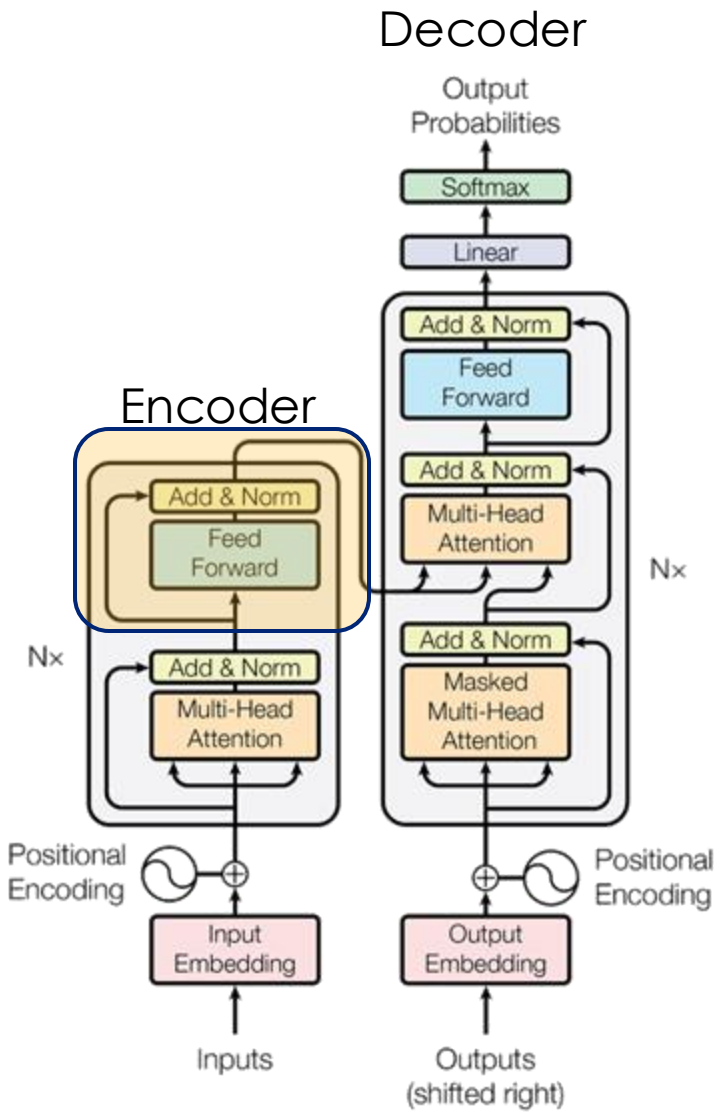
Review Disc. 10 on attention. We'll also continue with this in Prob. 2 of Disc. 11.



6) Obtain Representation of Input

Input: Where is the library?

Output of encoder: Compact representation of the input (English) sentence



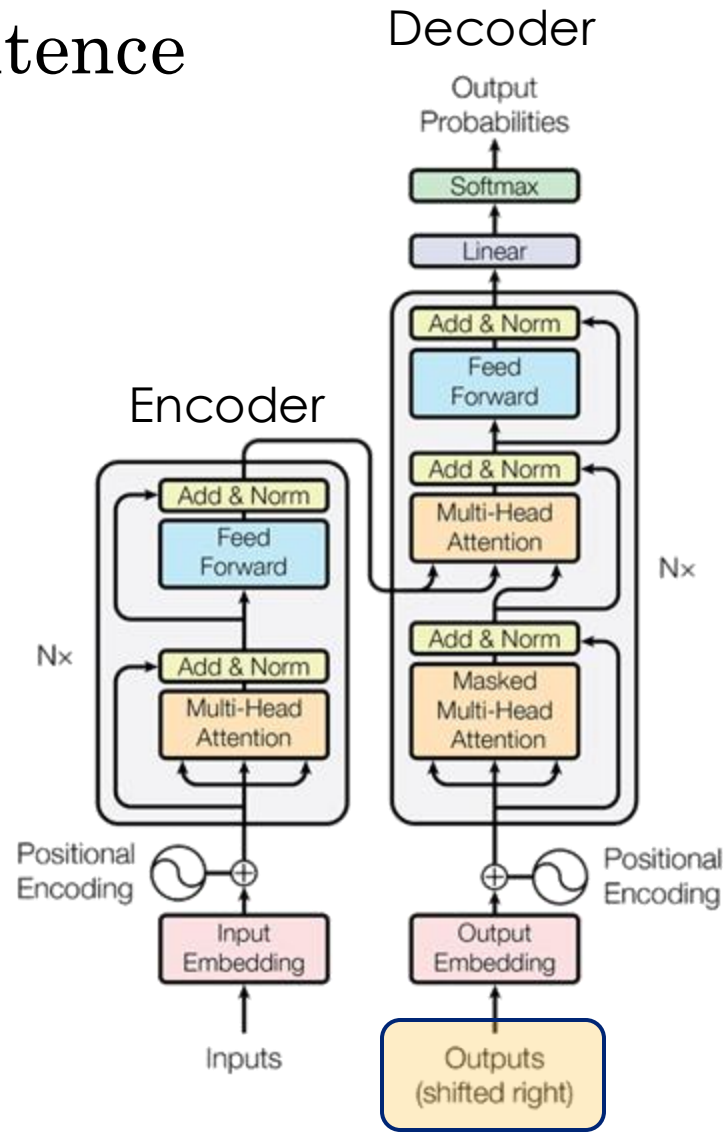
The Decoder of a Transformer

1. Transformer Encoder
2. Transformer Decoder
3. Training a Transformer
4. Using a Transformer for Inference

The Transformer Decoder

Goal: Learn to predict next word in Spanish sentence

Output: ¿Dónde está la biblioteca?



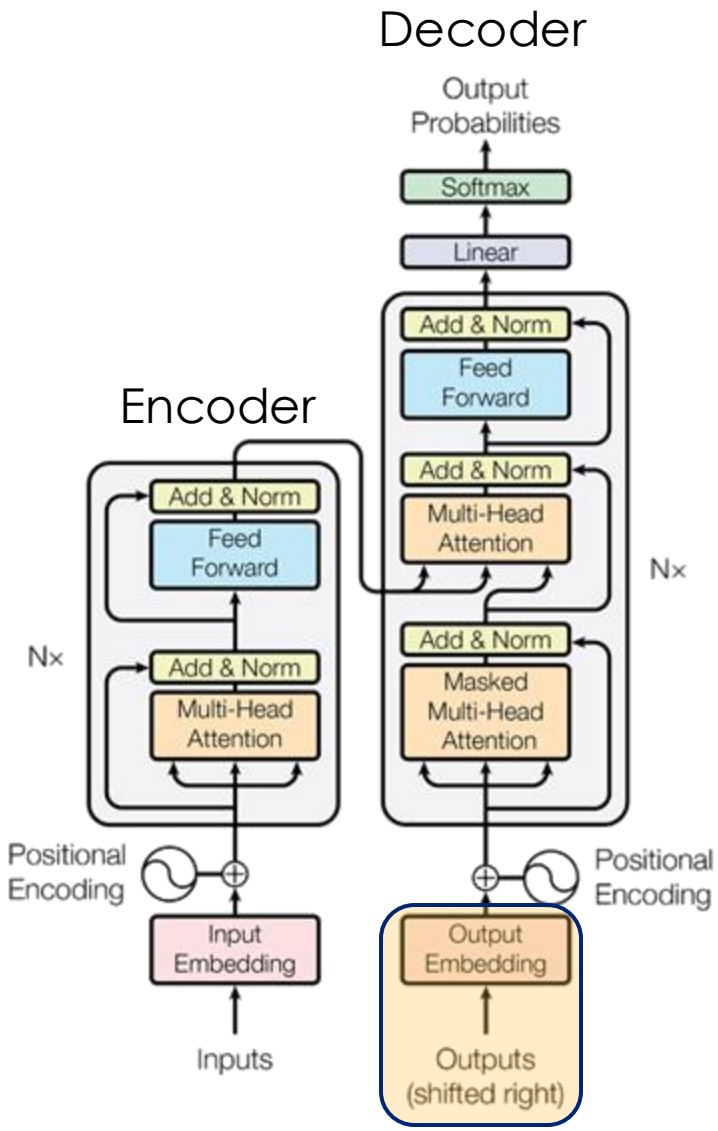
1) Obtain Input Embeddings

Output: [¿Dónde] [está] [la] [biblioteca] [?]

101	22	11	158	143	26
⋮	⋮	⋮	⋮	⋮	⋮
55	176	202	111	86	67
124	213	54	25	75	122
⋮	⋮	⋮	⋮	⋮	⋮
16	87	123	64	202	118

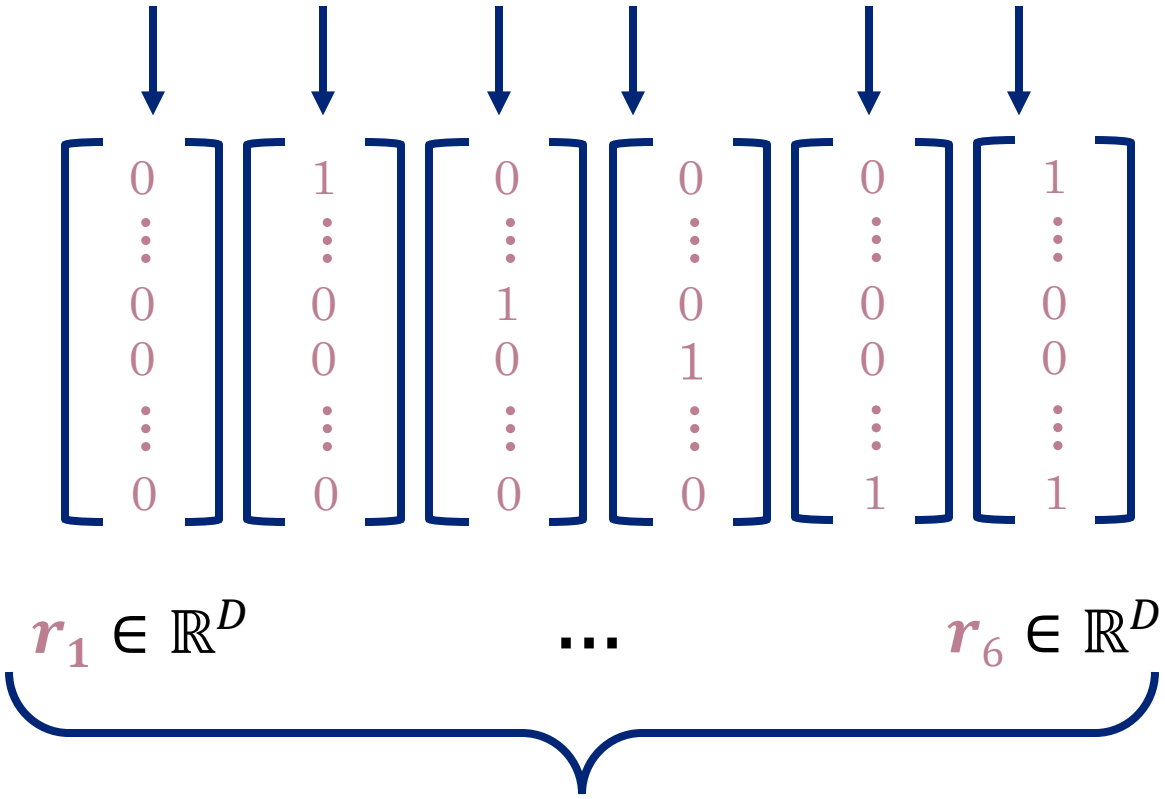
$x_1 \in \mathbb{R}^D$ ... $x_6 \in \mathbb{R}^D$

N = 6 input embeddings capturing word (part)

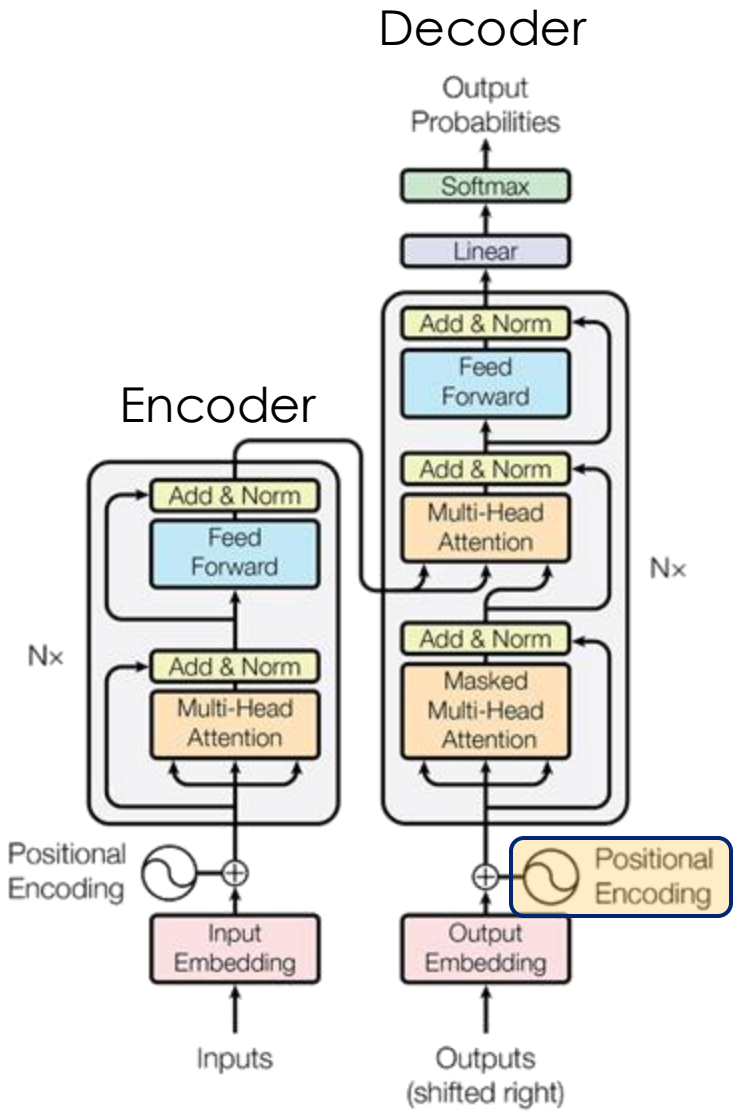


2) Grab Positional Encodings

Output: ¿Dónde está la biblioteca?



N = 6 positional encodings
capturing position in sequence



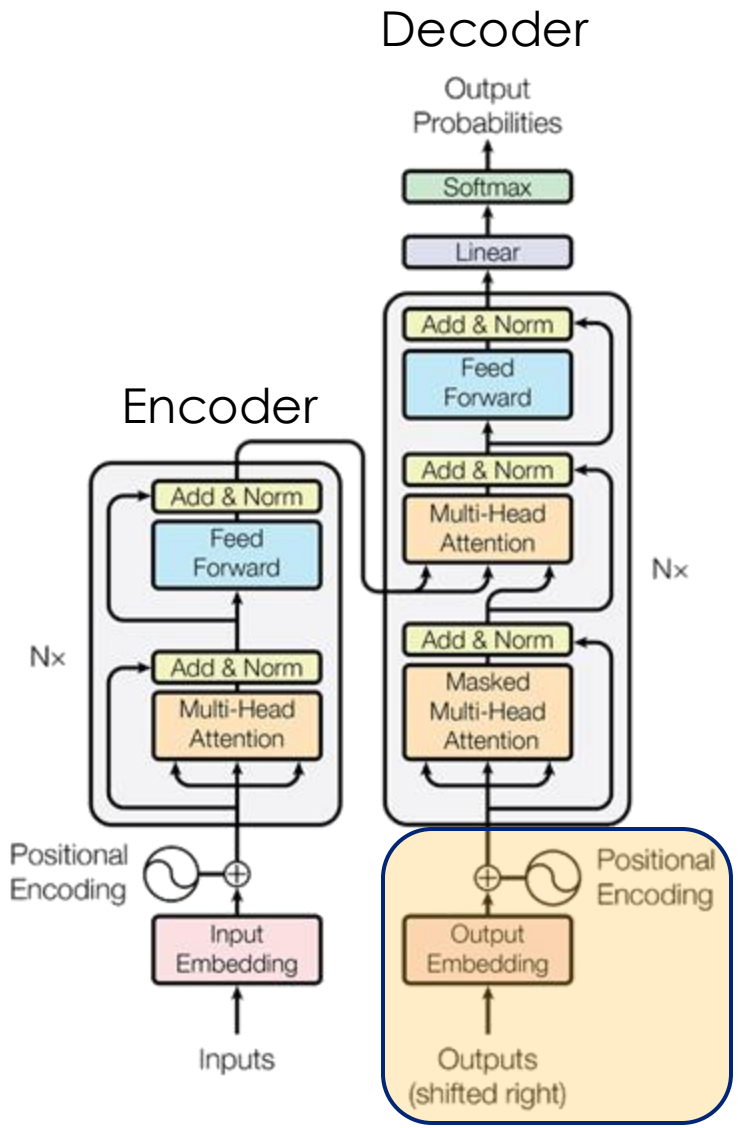
3) Combine Embeddings + Positions

Output: ¿Dónde está la biblioteca?

$$\begin{bmatrix} 101 \\ \vdots \\ 55 \\ 124 \\ \vdots \\ 16 \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 101 \\ \vdots \\ 55 \\ 124 \\ \vdots \\ 16 \end{bmatrix}$$

$x_1 \in \mathbb{R}^D \quad r_1 \in \mathbb{R}^D \quad \tilde{x}_1 \in \mathbb{R}^D$

N = 6 decoder inputs
(word + position)



4) Generate Queries, Keys, and Values

Output: ¿Dónde está la biblioteca?

Query: Question token asks rest of sequence

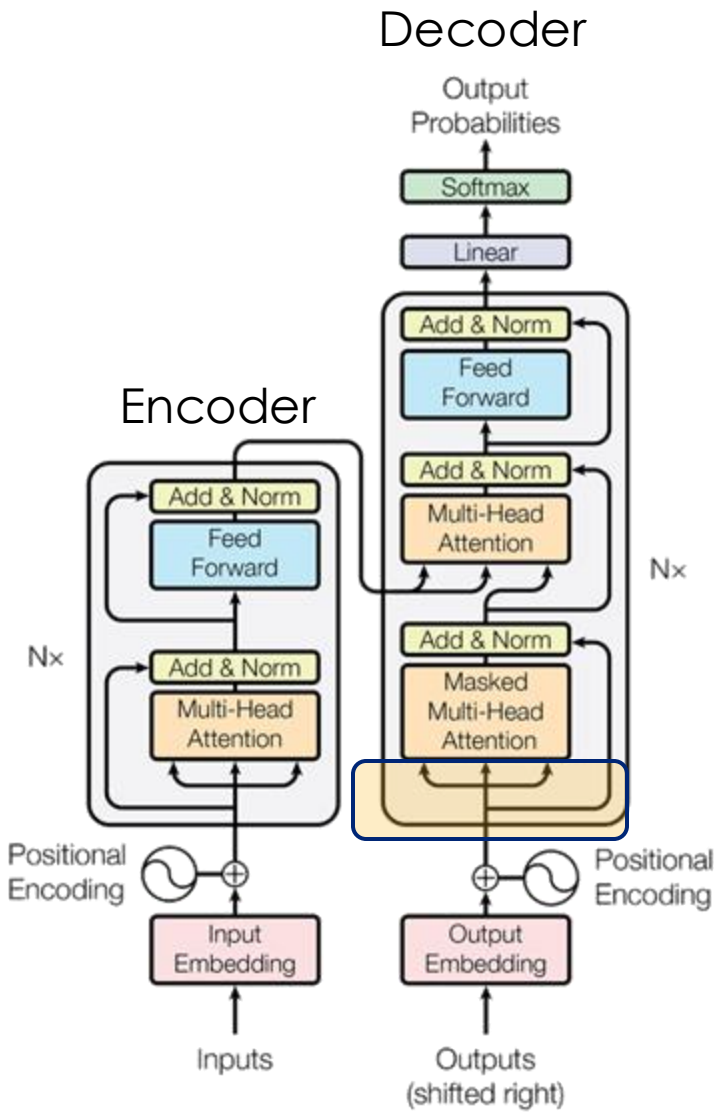
$$\boldsymbol{q}_1 = (\mathbf{W}^{(q)})^\top \tilde{\boldsymbol{x}}_1 \in \mathbb{R}^{D_\mathrm{k}}$$

Key: Attributes token exposes to be searched

$$\boldsymbol{k}_1 = (\mathbf{W}^{(k)})^\top \tilde{\boldsymbol{x}}_1 \in \mathbb{R}^{D_\mathrm{k}}$$

Value: Info token provides if deemed relevant

$$\boldsymbol{v}_1 = (\mathbf{W}^{(v)})^\top \tilde{\boldsymbol{x}}_1 \in \mathbb{R}^{D_v}$$



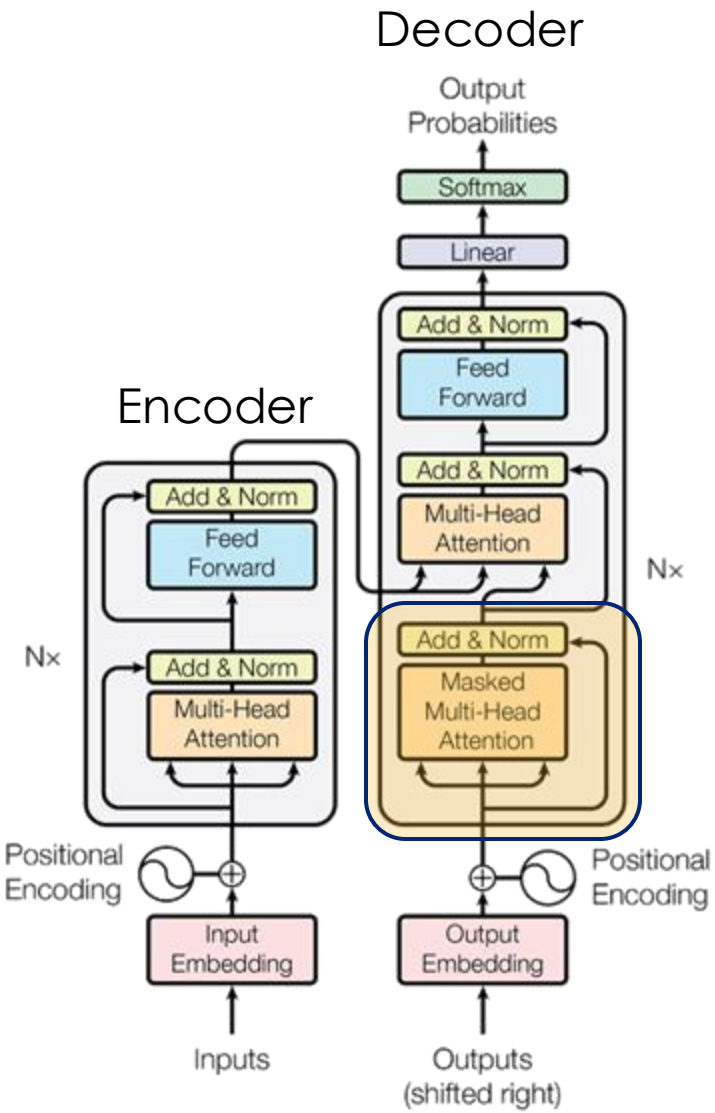
5) Get Contextualized Embeddings: Output Only

Output: ¿Dónde está la biblioteca?

Contextualized embedding: Meaning of token in context of all other output tokens

Key difference from encoder: Masked attention prevents earlier tokens from using context provided by later ones

(e.g., está can see Dónde but Dónde cannot see está during training)



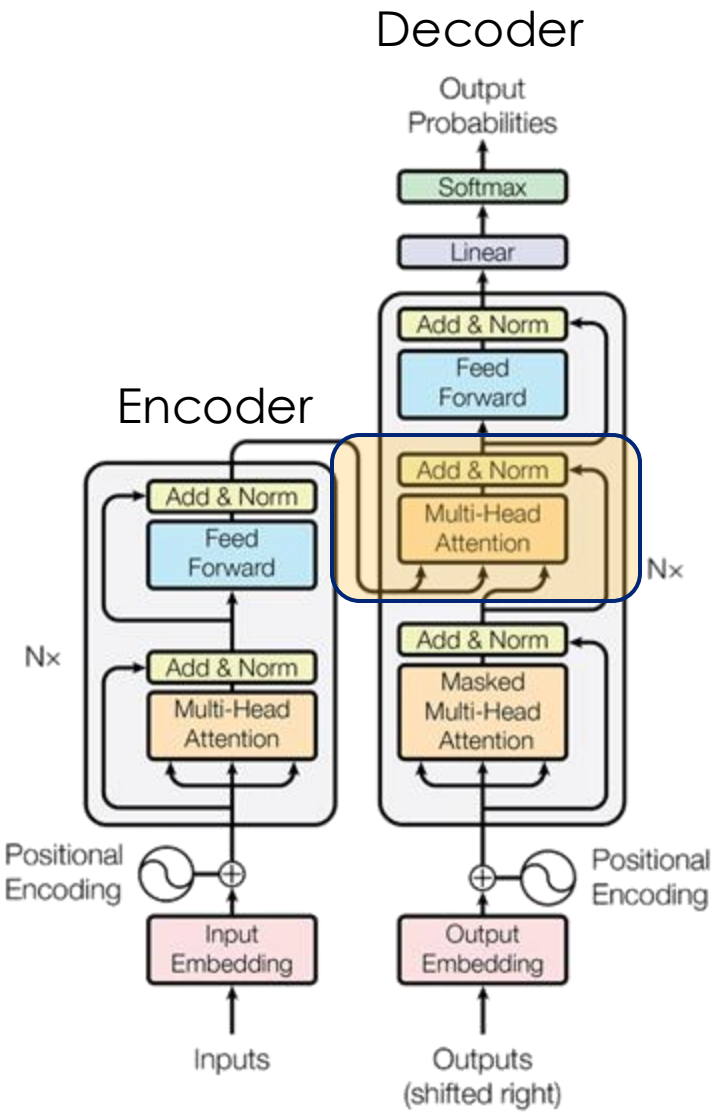
6) Get Contextualized Embeddings: With Input

Output: ¿Dónde está la biblioteca?

Contextualized embedding: Meaning of token in context of all other output tokens plus information provided by input

Key steps:

1. Extract input keys + values from encoder output (compact input representation)
2. Create queries from output representation
3. Compare output queries to the input keys to decide what to attend to in input
4. Produce contextualized embeddings of output

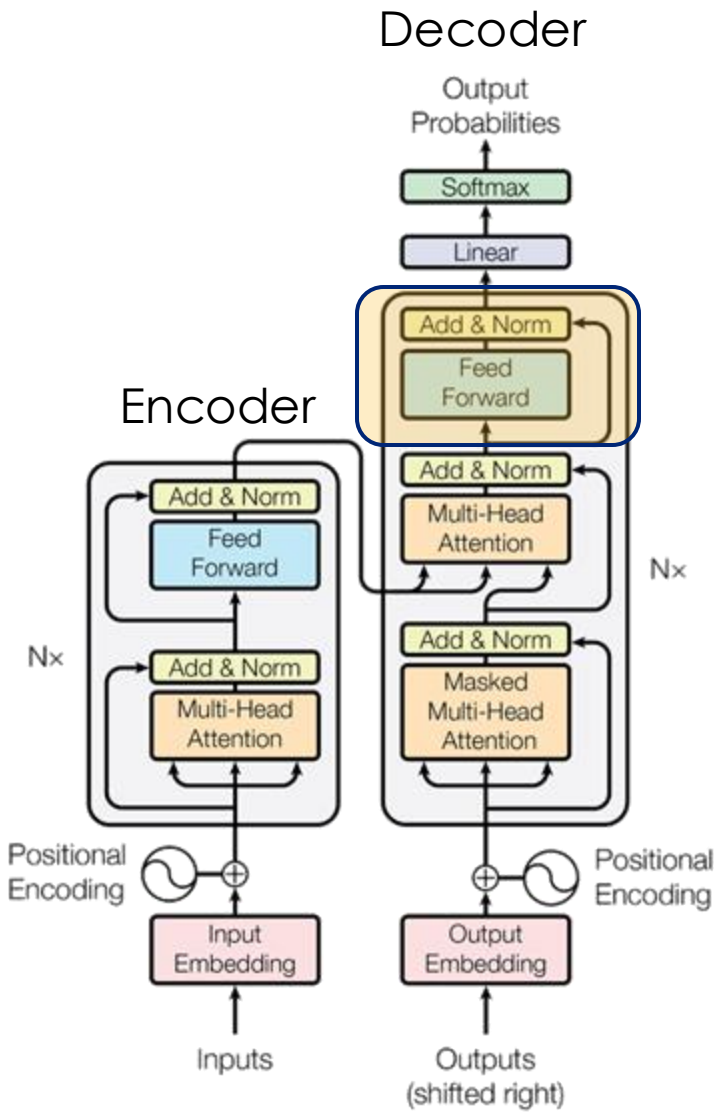


7) Obtain Representation of Output + Input

Input: Where is the library?

Output: ¿Dónde está la biblioteca?

Output of decoder attention blocks:
Compact representation of the output (Spanish) sentence and relevant parts of input (English) sentence



Training an Encoder-Decoder Transformer

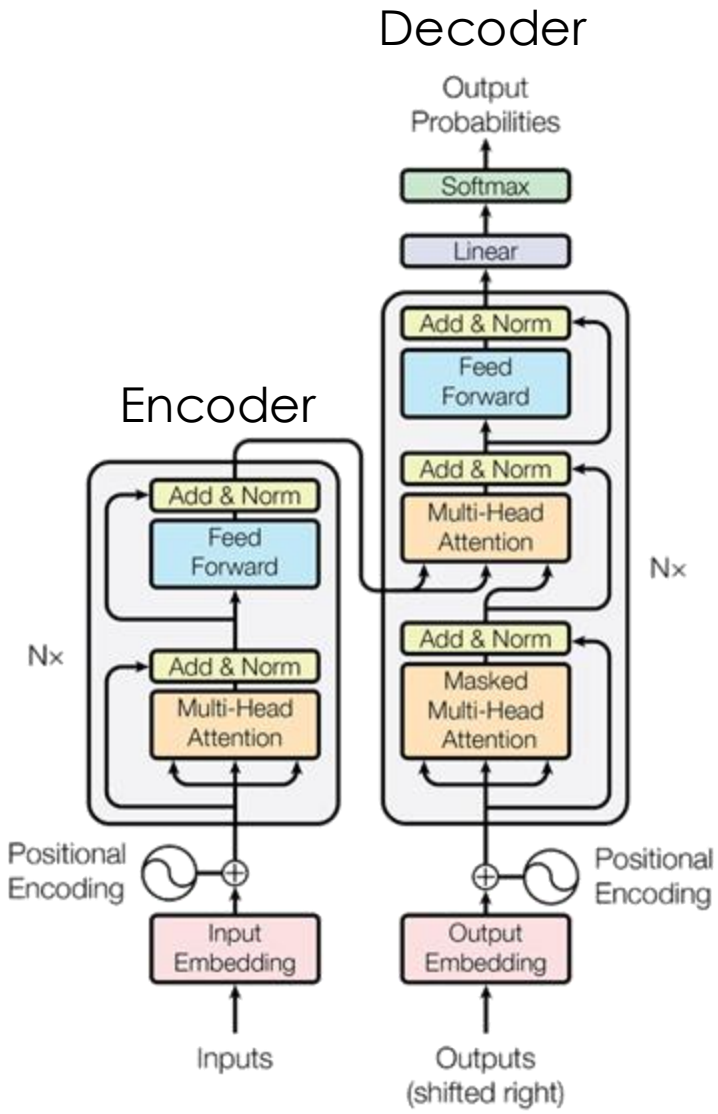
1. Transformer Encoder
2. Transformer Decoder
3. Training a Transformer
4. Using a Transformer for Inference

Data Sample for Training the Transformer

Inputs: <SOS> Where is the library <EOS>

Outputs: <SOS> Dónde está la biblioteca

Targets: Dónde está la biblioteca <EOS>

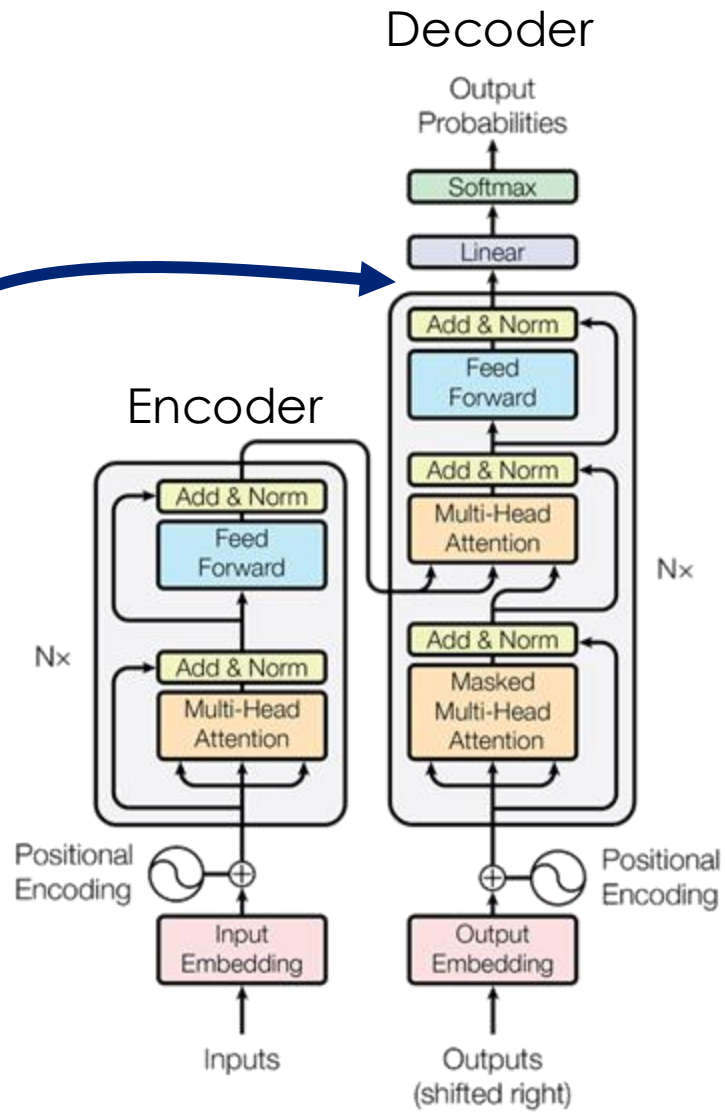


1) Obtain Representation of Input + Output

Inputs: <SOS> Where is the library <EOS>

Outputs: <SOS> Dónde está la biblioteca

Output of decoder attention blocks:
Compact representation of the output (Spanish) sentence and relevant parts of input (English) sentence



2) Map Representation to Word Probabilities

Inputs: <SOS> Where is the library <EOS>

Outputs: <SOS> Dónde está la biblioteca

Representation of inputs + outputs:

$$X \in \mathbb{R}^{N \times D_{model}}$$

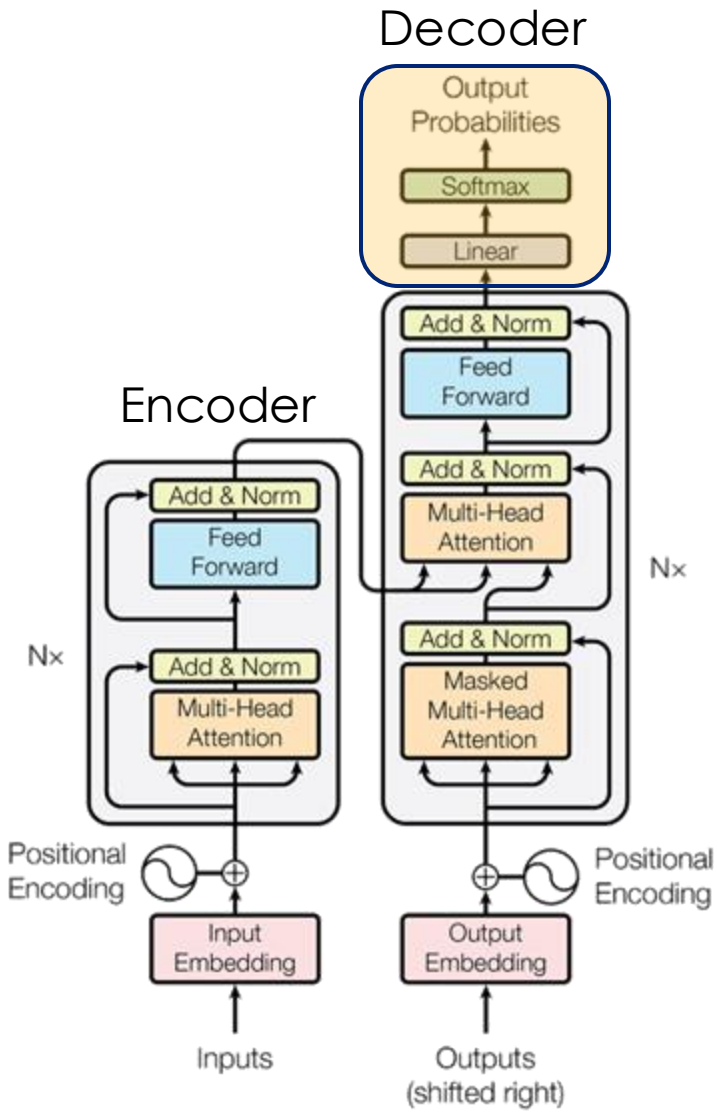
↓ **Linear:** $\mathbb{R}^{D_{model}} \mapsto \mathbb{R}^{D_{vocab}}$

$$Z \in \mathbb{R}^{N \times D_{vocab}}$$

↓ **Softmax:** $\mathbb{R}^{D_{vocab}} \mapsto \mathbb{R}^{D_{vocab}}$

Output word probabilities:

$$Y \in \mathbb{R}^{N \times D_{vocab}}$$



3) Compare Word Probabilities to Target

Inputs: <SOS> Where is the library <EOS>

Outputs: <SOS> Dónde está la biblioteca

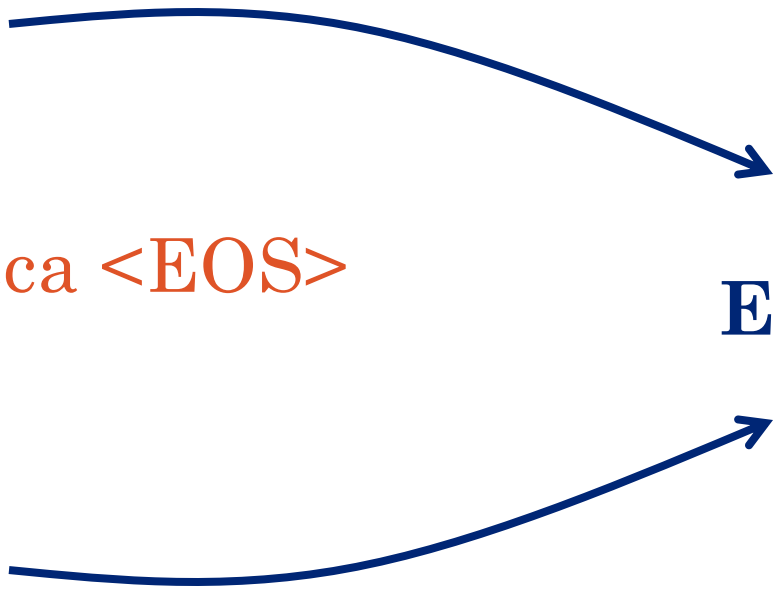
Output word probabilities:

$$Y \in \mathbb{R}^{N \times D_{vocab}}$$

Targets: Dónde está la biblioteca <EOS>

Target word probabilities:

$$T \in \mathbb{R}^{N \times D_{vocab}}$$

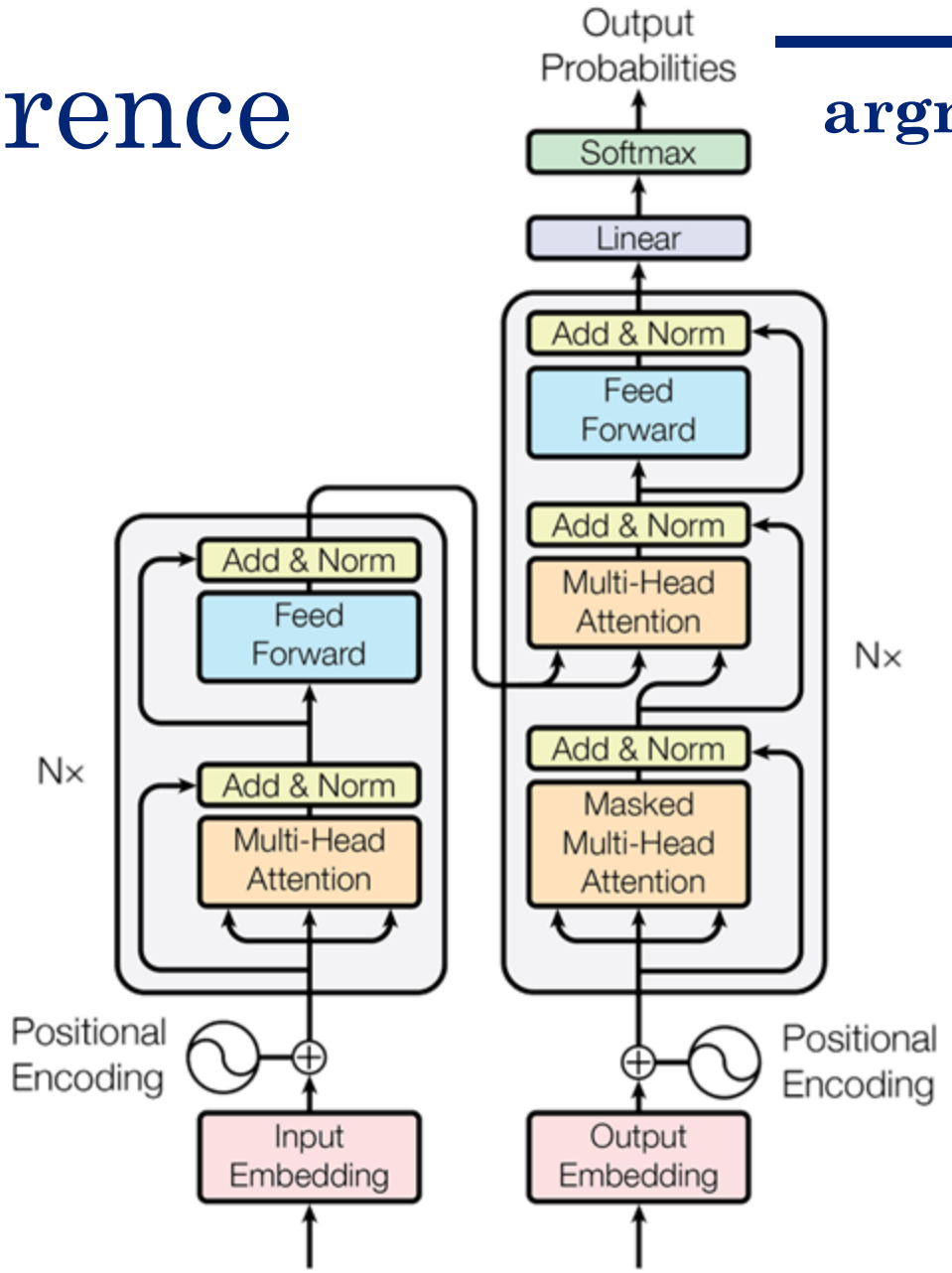


Cross Entropy Loss

Inference with Encoder-Decoder Transformer

1. Transformer Encoder
2. Transformer Decoder
3. Training a Transformer
4. Using a Transformer for Inference

Step 1 of Inference



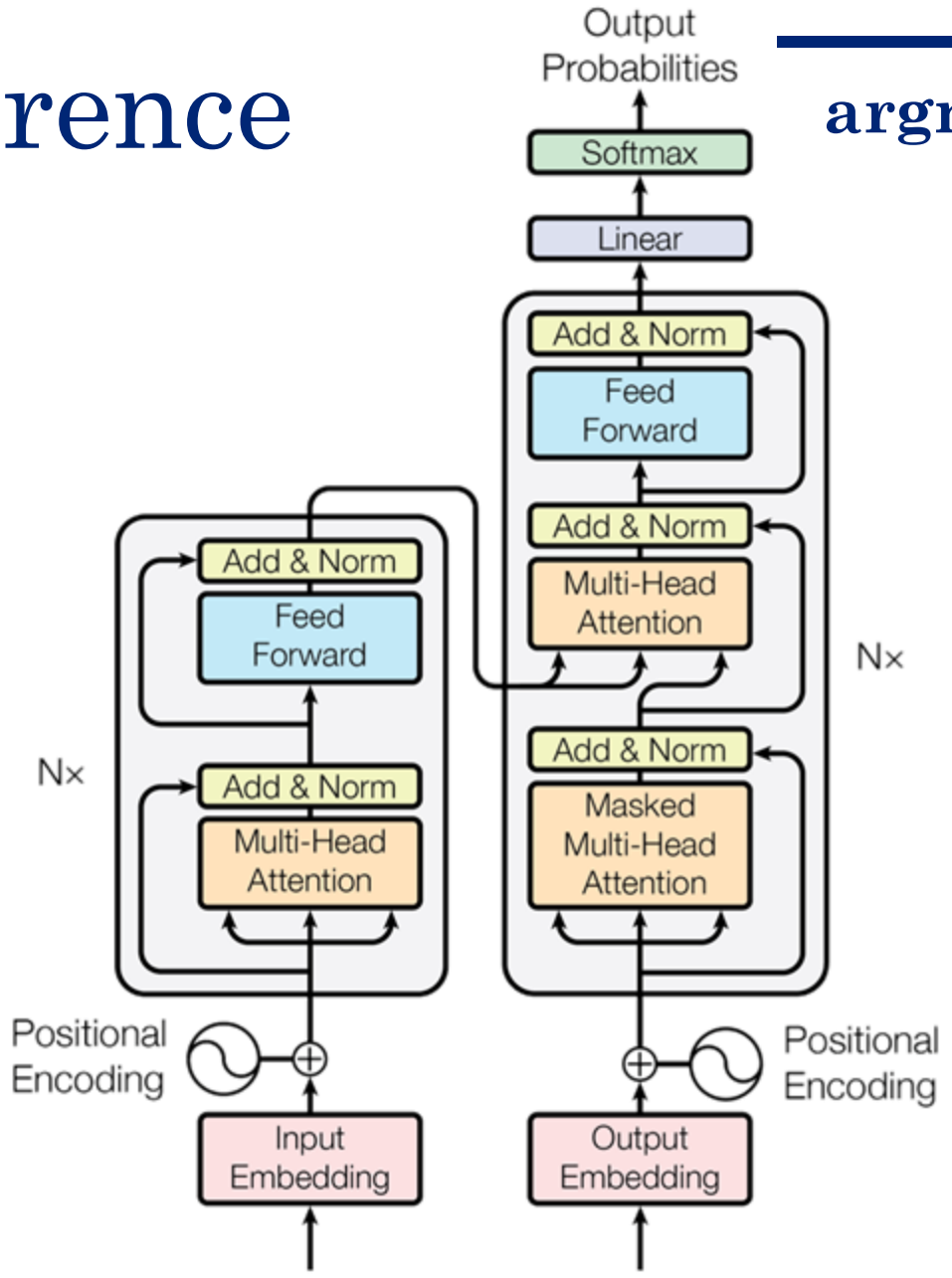
argmax

Dónde

<SOS> Where is the library <EOS>

<SOS>

Step 2 of Inference



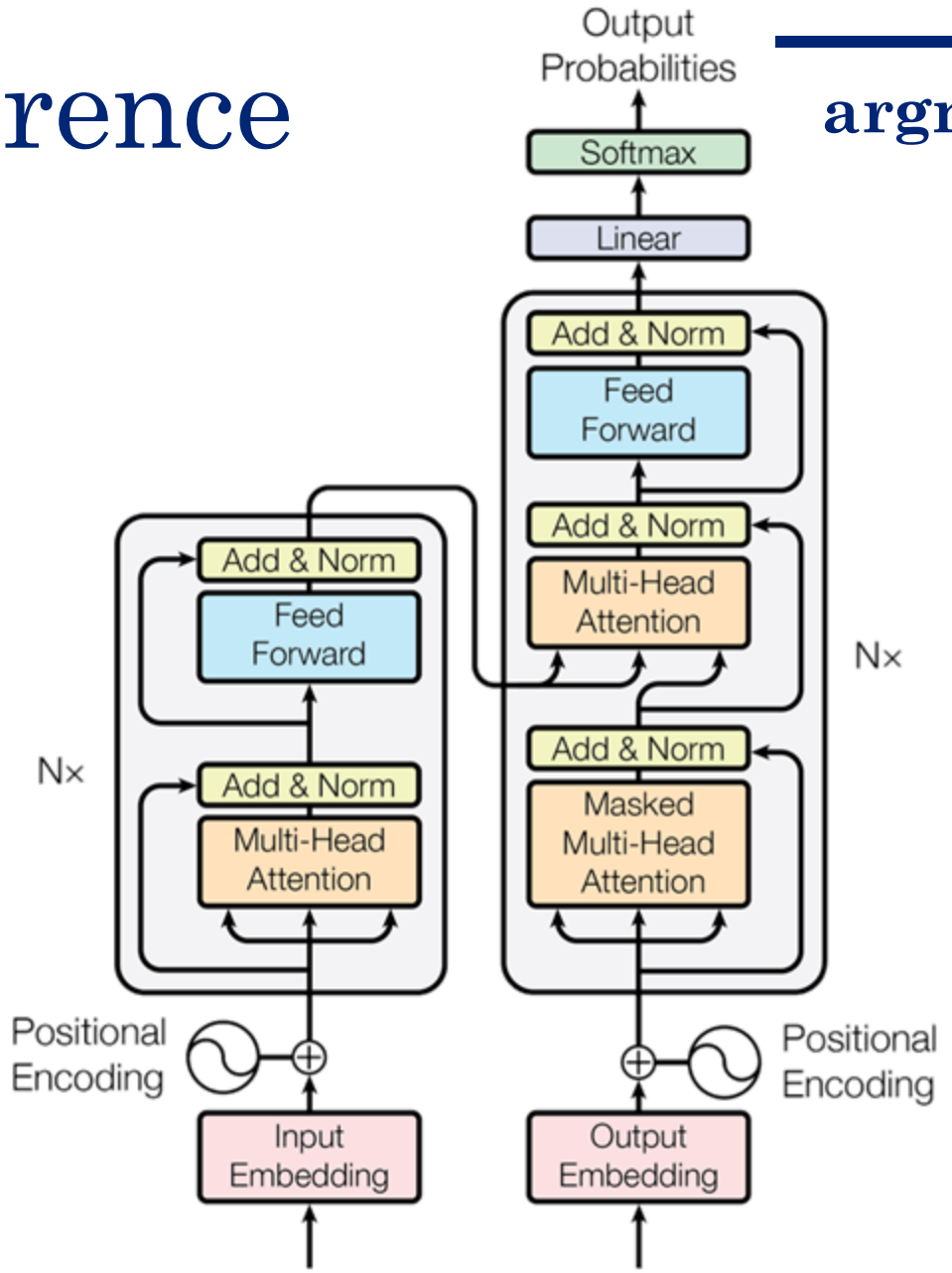
argmax

está

<SOS> Where is the library <EOS>

<SOS> Dónde

Step 3 of Inference



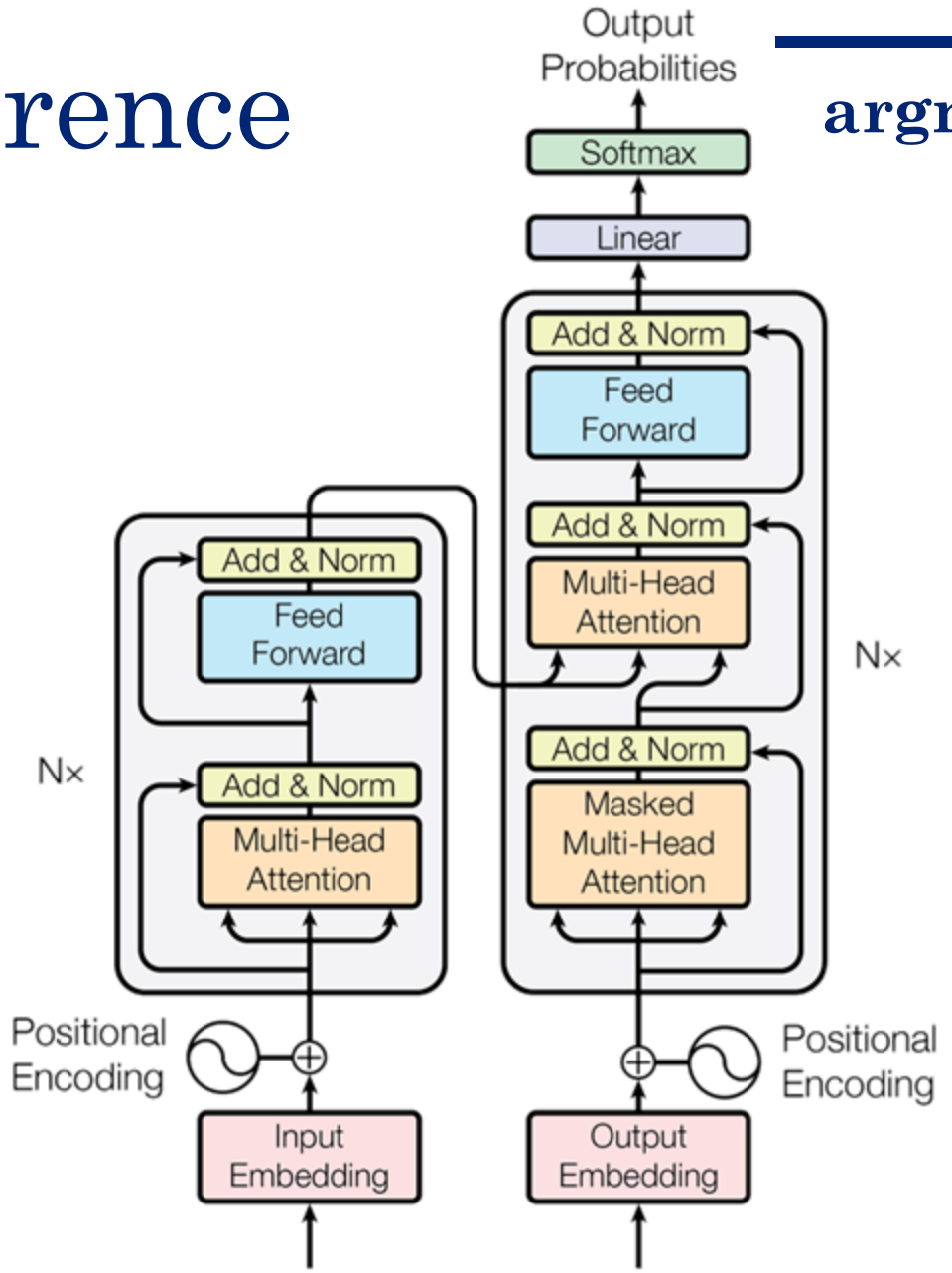
argmax

la

<SOS> Where is the library <EOS>

<SOS> Dónde está

Step 4 of Inference



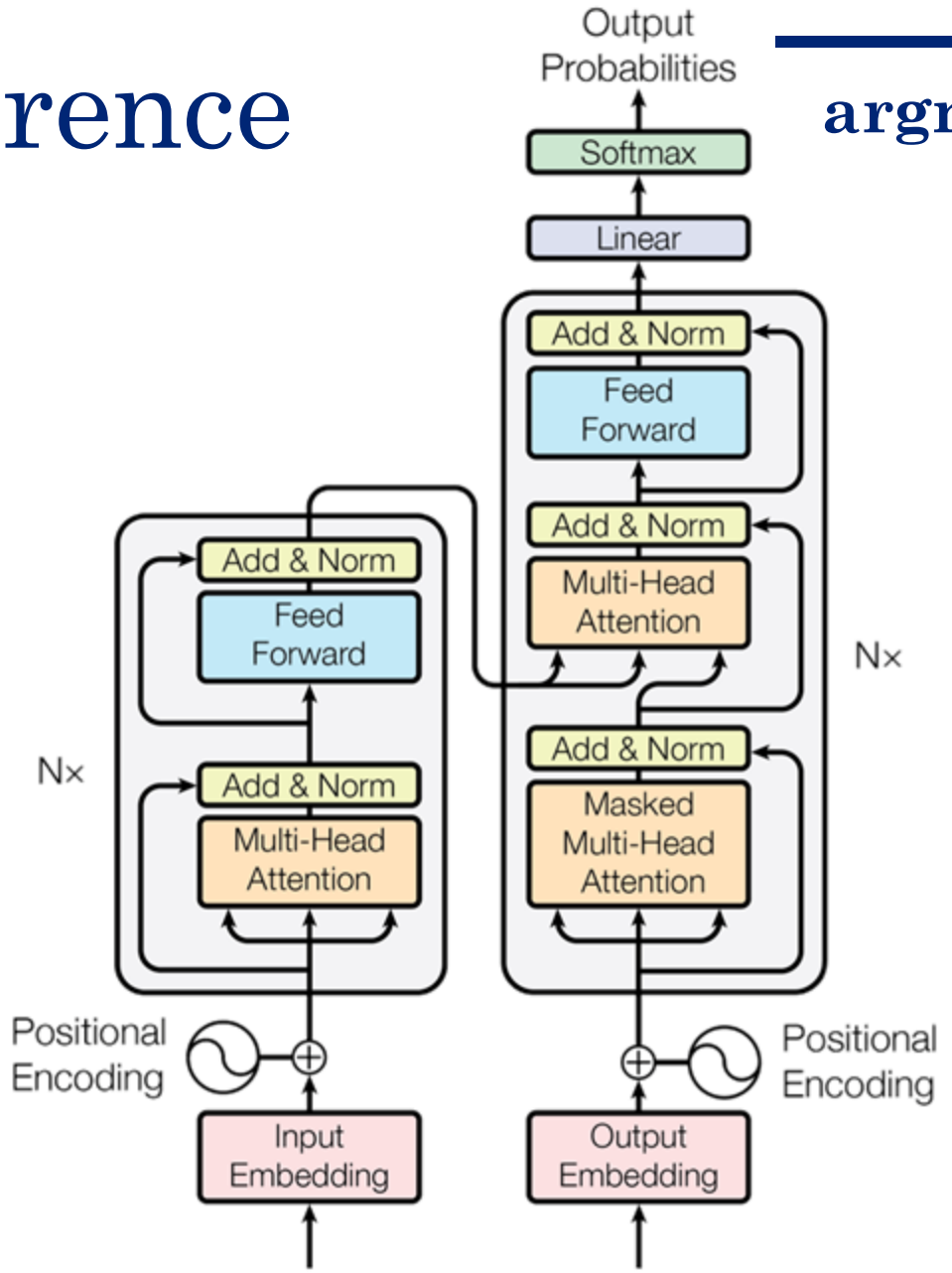
argmax

biblioteca

<SOS> Where is the library <EOS>

<SOS> Dónde está la

Step 5 of Inference



argmax

<EOS>

<SOS> Where is the library <EOS>

<SOS> Dónde está la biblioteca

KV Caching During Inference

During each step of inference, we predict the next token using context provided by all the previously generated tokens.

Therefore, we're continuously using the keys/values of earlier tokens.

To avoid recomputing the key and value matrices at each step of inference, we cache these matrices and re-use them.

We'll explore KV caching in Prob. 3 of Disc. 11.

Discussion Mini Lecture 11

More on Transformers

Contributors: Sara Pohland

Additional Resources

1. Transformers
   - Deep Learning Foundations and Concepts – Chapter 12.1
   - Umar Jamil – Attention is all you need (Transformer) - Model explanation (including math), Inference and Training