

Ashish Vaswani\*  
Google Brain  
avaswani@google.com

Noam Shazeer\*  
Google Brain  
noam@google.com

Niki Parmar\*  
Google Research  
nikip@google.com

Jakob Uszkoreit\*  
Google Research  
usz@google.com

Llion Jones\*  
Google Research  
llion@google.com

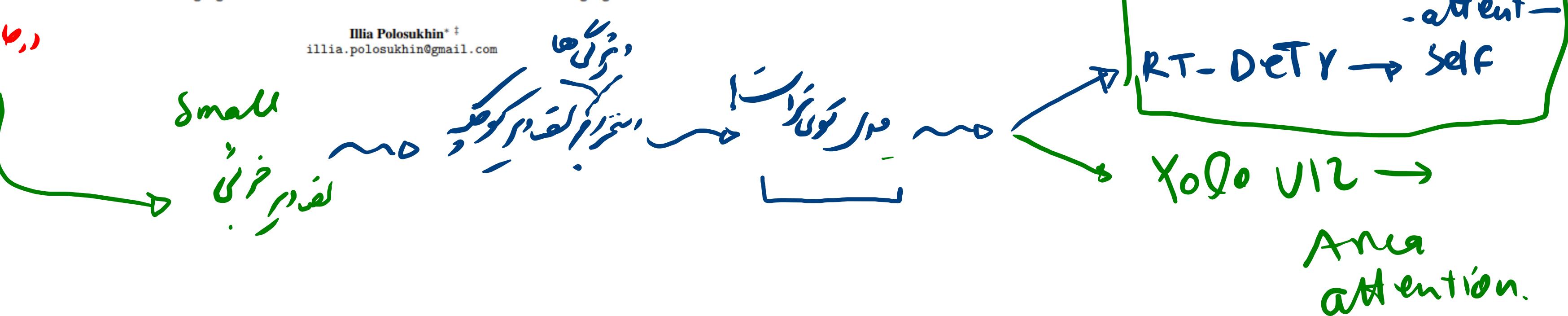
Aidan N. Gomez\* †  
University of Toronto  
aidan@cs.toronto.edu

Lukasz Kaiser\*  
Google Brain  
lukasz.kaiser@google.com

Illia Polosukhin\* ‡  
illia.polosukhin@gmail.com

## نقطه عطف

در کار با سیستم های بلند و سری های زمانی و تصاویر و ...



# RNN vs Trans.

تمام توکن‌ها را به صورت هم‌زمان پردازش می‌کند (موازی) → سریع‌تر روی GPU و TPU سرعت آموزش بسیار بالاتر

پردازش موازی (Parallel Processing)

در Transformer، هر کلمه می‌تواند مستقیماً به همه کلمات دیگر دسترسی مستقیم از طریق attention داشته باشد. درک بهتر ساختار معنایی در متن‌های طولانی

درک وابستگی‌های بلندمدت

امکان تحلیل، خطایابی، و visualization ساده‌تر

تفسیرپذیری بالاتر (Interpretability)

پیاده‌سازی و درک ساده‌تر در معماری کلی	شامل توابع پیچیده: forget gate, update gate, input gate	LSTM/GRU
	ساختاری ساده‌تر با recurrence	Transformer

سادگی معماری و حذف تکرار

برتری‌های  
نسبت به مدل‌های قبلی

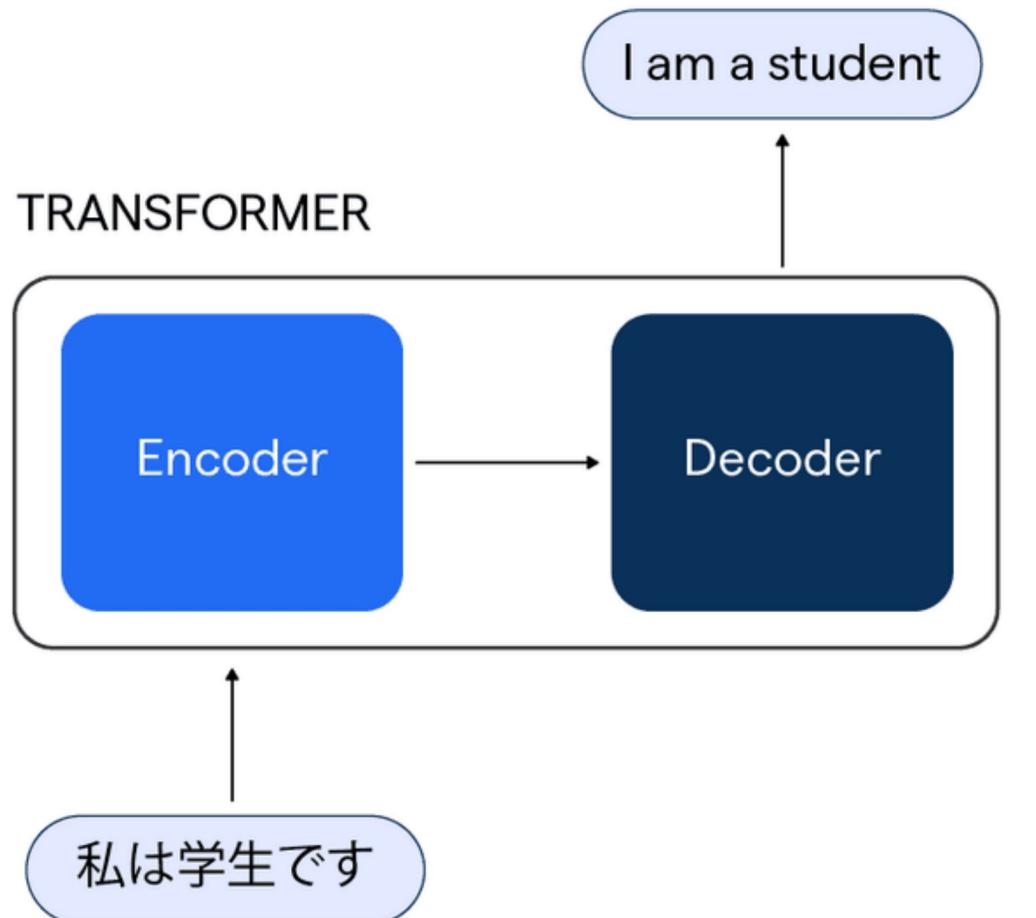
مدل‌های قبل فقط برای توالی‌های زمانی مناسب بودن (مثل متن یا صوت) تعمیم ترانسفورمر به بینایی ماشین (ViT)، موسیقی، کد برنامه‌نویسی، چندرشانه‌ای و ...

قابلیت تعمیم به دیگر حوزه‌ها

...BERT، GPT، T5، ViT، Whisper، DALL-E، CLIP با تغییر ساده در encoder/decoder/self-attention ساختارهای جدید قابل طراحی‌اند.

انعطاف بالا در طراحی مدل‌ها

## شبکه ترانسفورمر (حالت کلی)

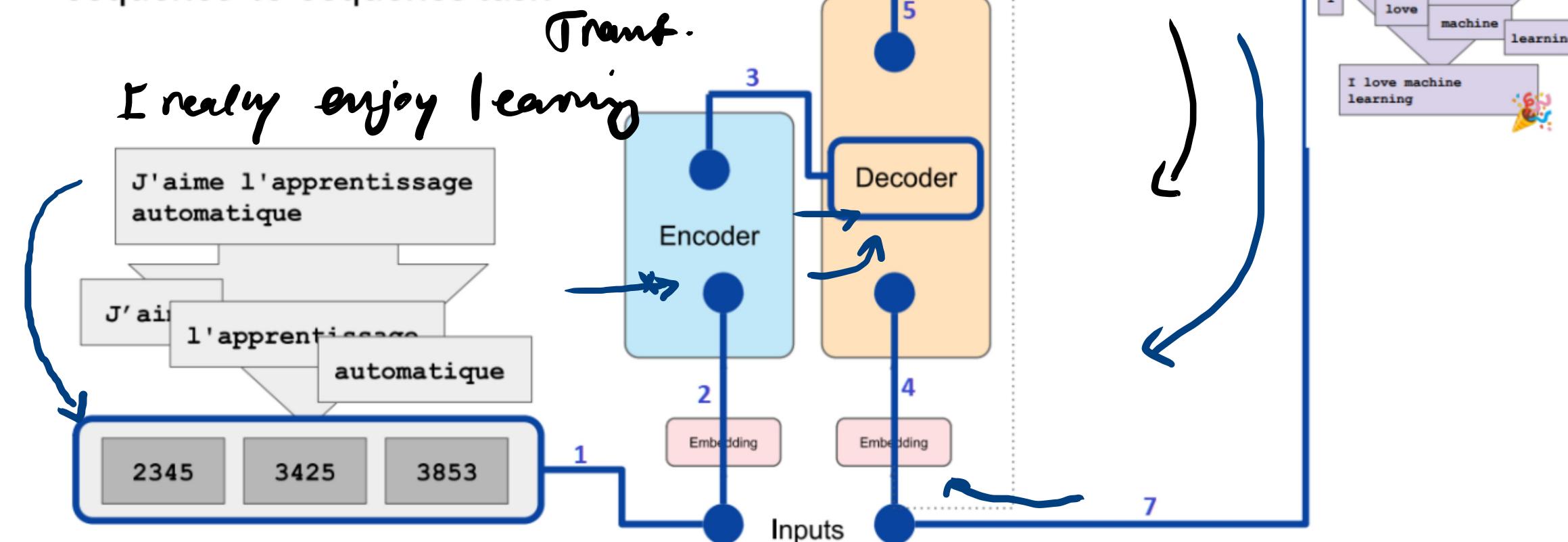


## شبکه ترانسفورمر (حالت کلی)

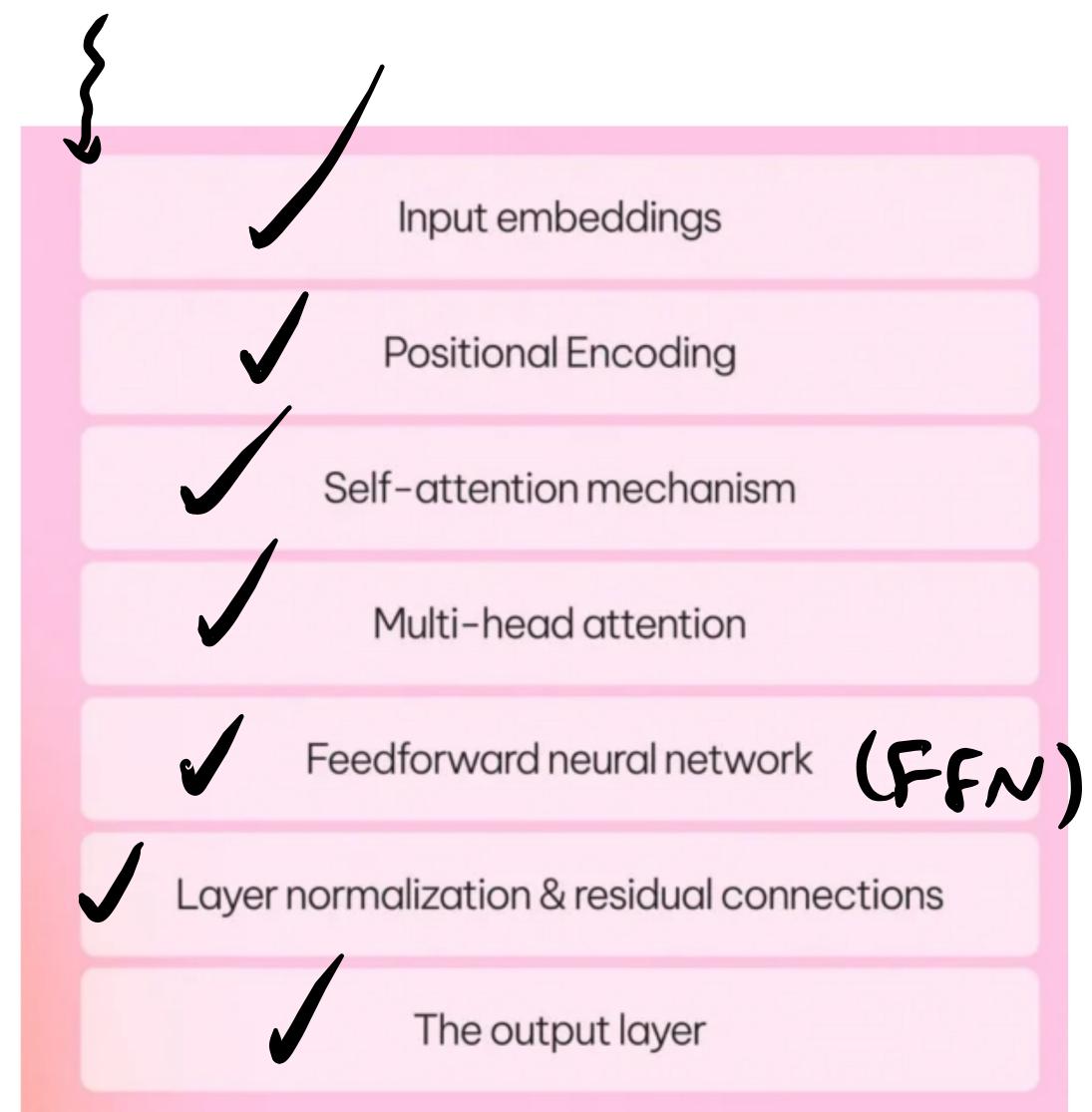
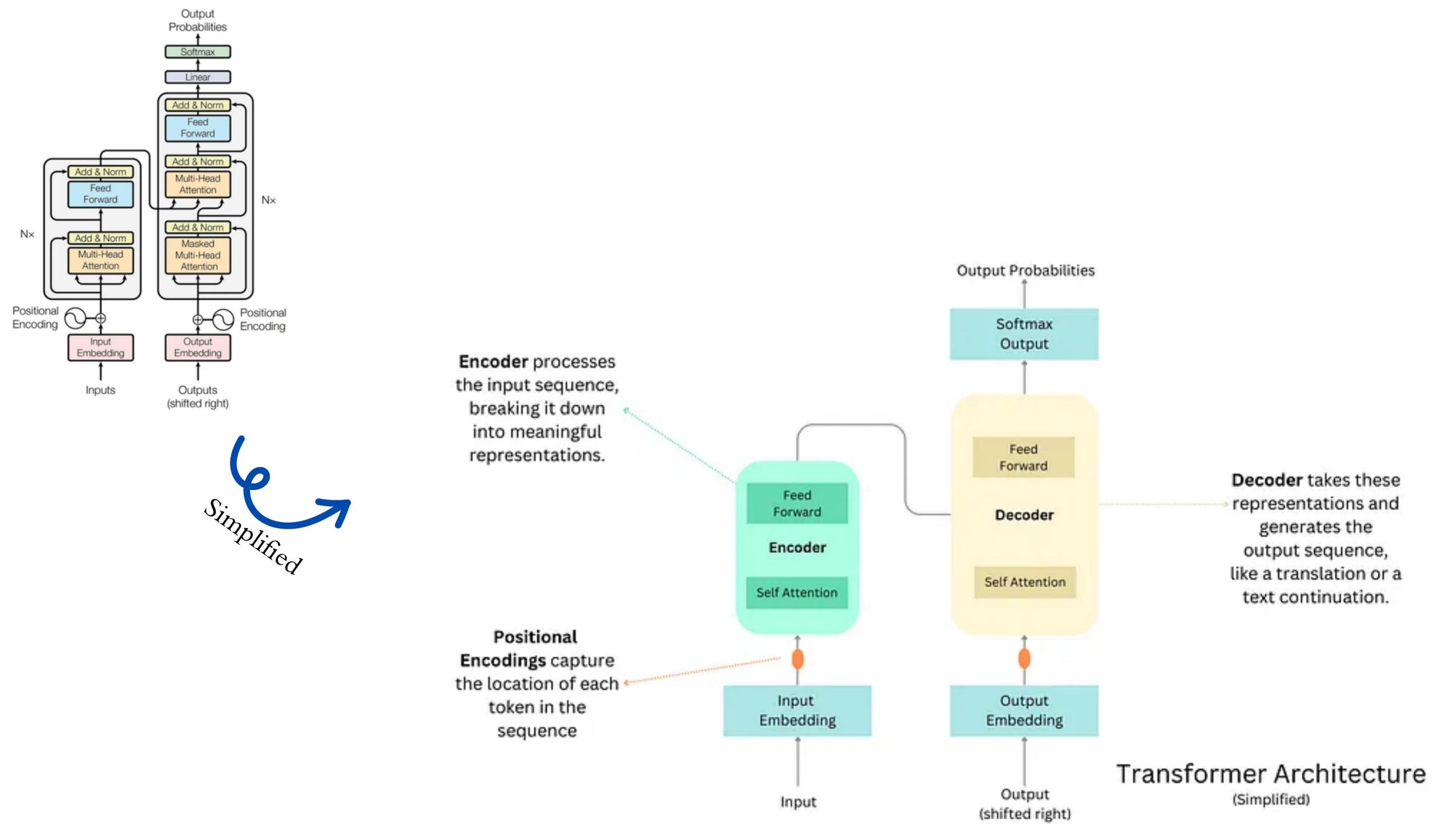
[ چن دانه ] نت‌ررم از بایری ترانسفورمر

## Transformers

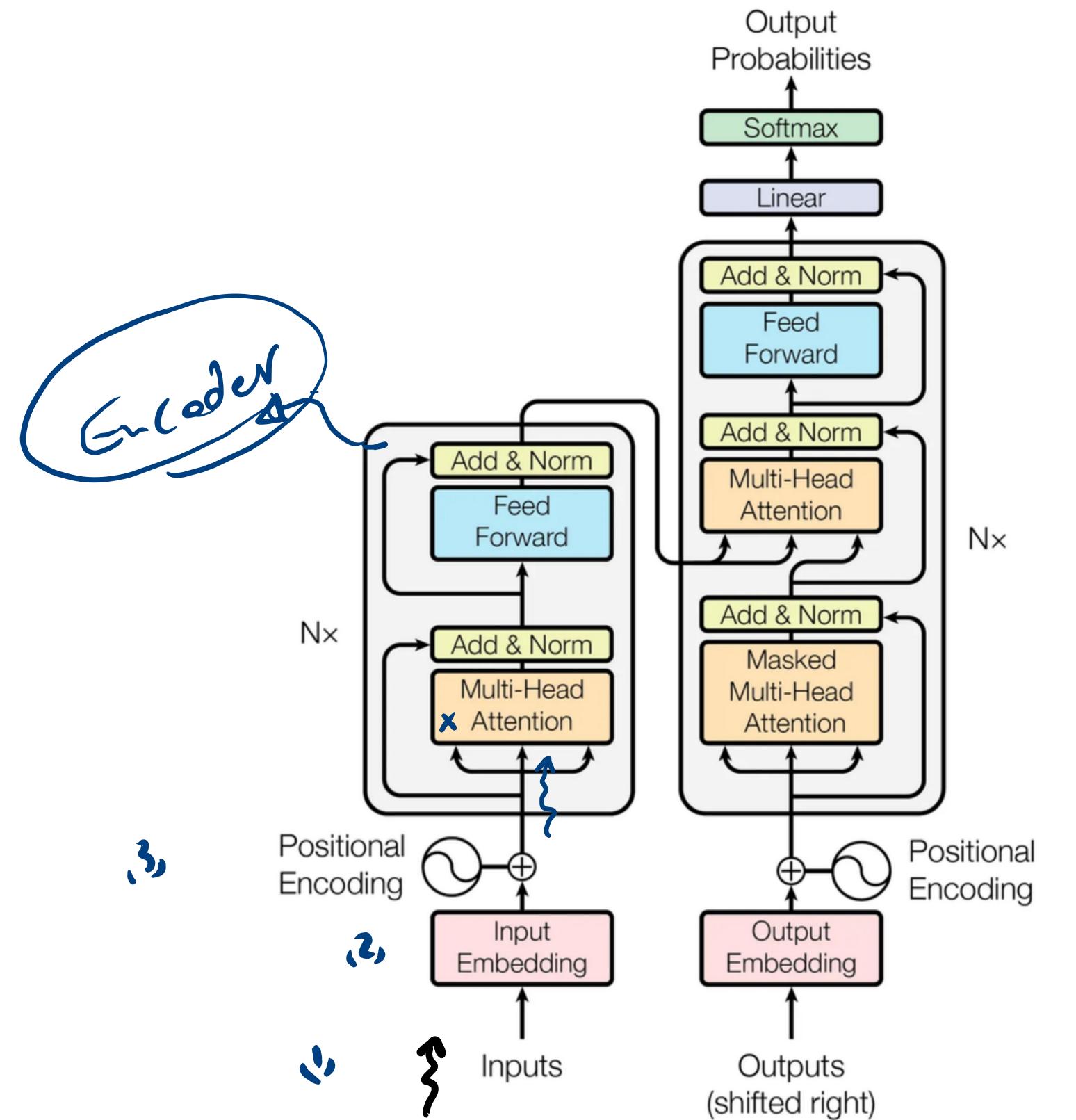
Translation:  
sequence-to-sequence task



## شبکه ترانسفومر، کمی جزیی تر



## شبکه ترانسفورمر، دقیق تر



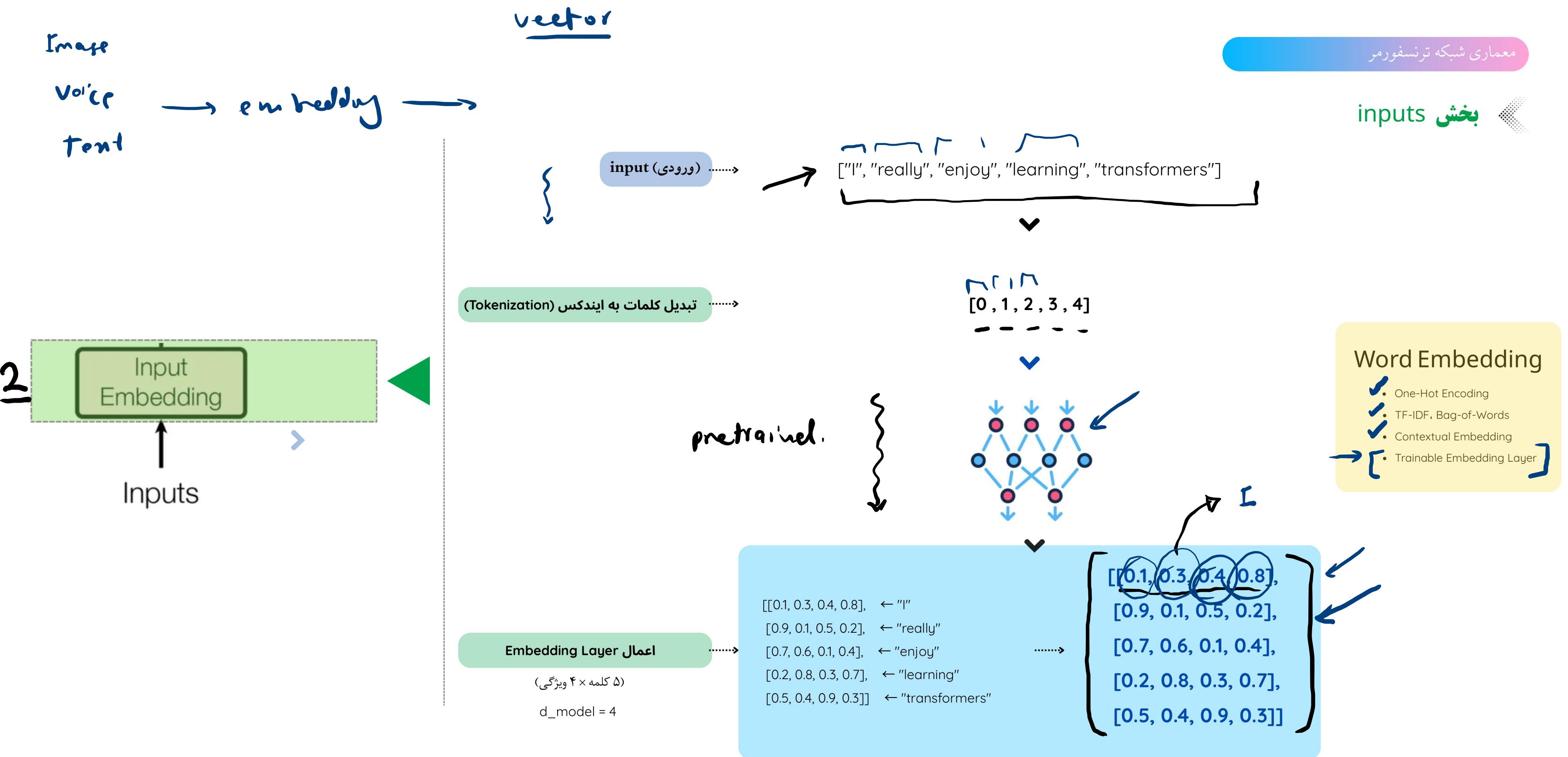
## بخش inputs



~~~~> I  
Inputs >

I really enjoy learning transformers

→ ["I", "really", "enjoy", "learning", "transformers"]

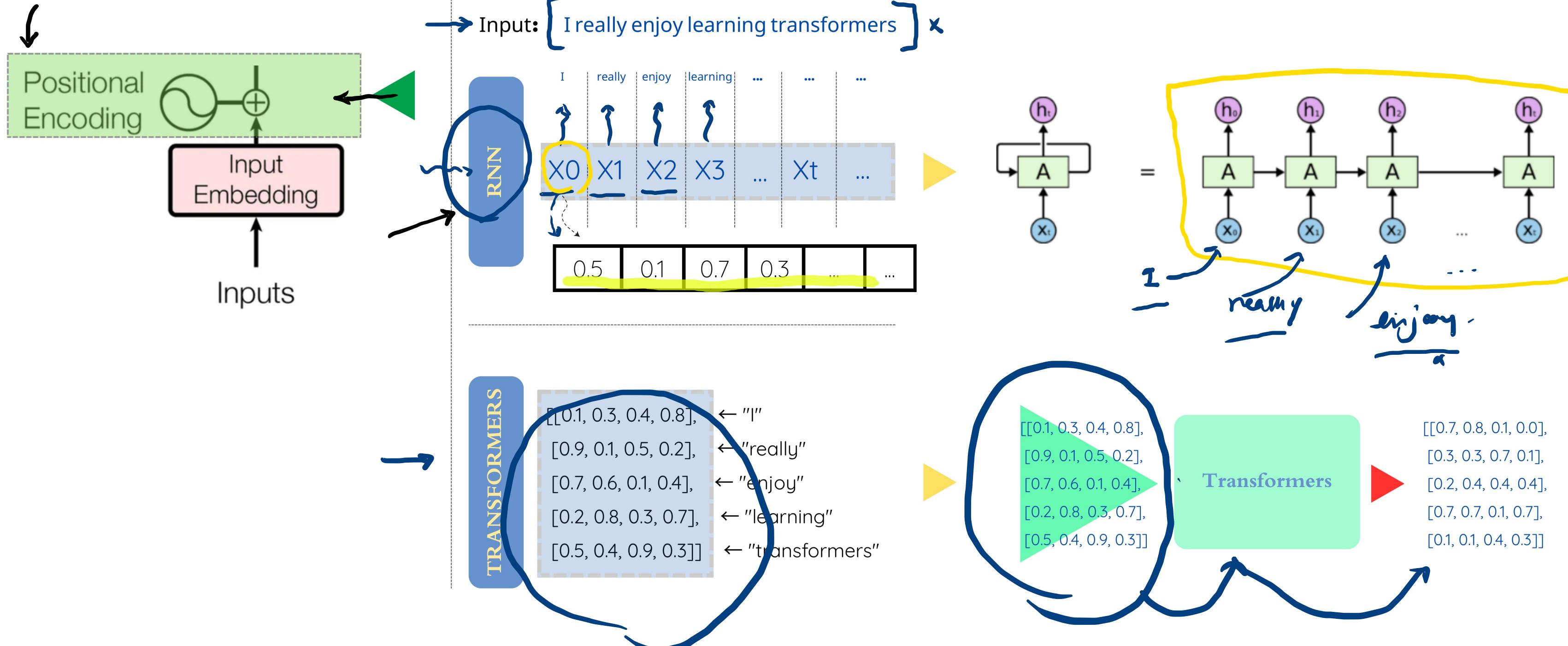


## بخش Positional Encoding (افزودن موقعیت)

### چرا position encoding

برخلاف RNN‌ها که توکن‌ها را به ترتیب پردازش می‌کنند، Transformer کل جمله را یکجا (موازی) پردازش می‌کند.

به صورت ذاتی نمی‌داند ترتیب کلمات چیست!



# → (سینوسی و کوسمینوسی)

## چگونه (sinusoidal positional encoding) & positional encoding

برای اینکه مدل بتواند ترتیب توکن‌ها را تشخیص دهد، به هر توکن یک بردار موقعیتی (Positional Encoding) اضافه می‌شود.

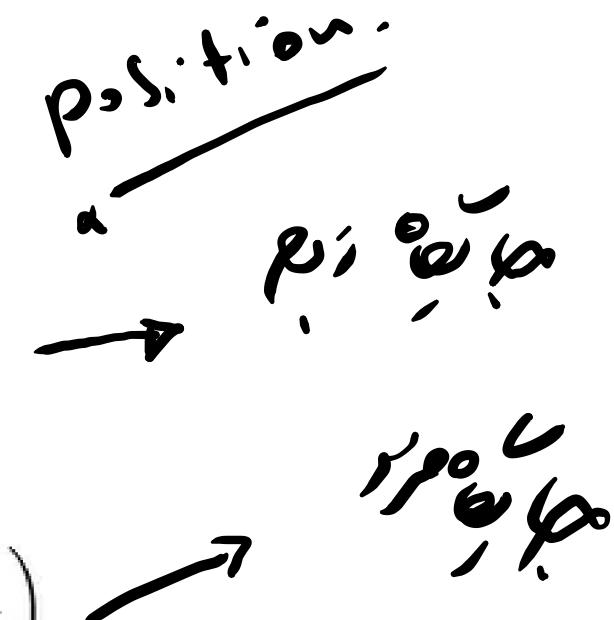
دادن اطلاعات «موقعیت» به هر توکن به طوری که مدل بفهمد "I" اول جمله است و "آخر" آخر "transformers" است.

:Positional Encoding

I ... Branche .

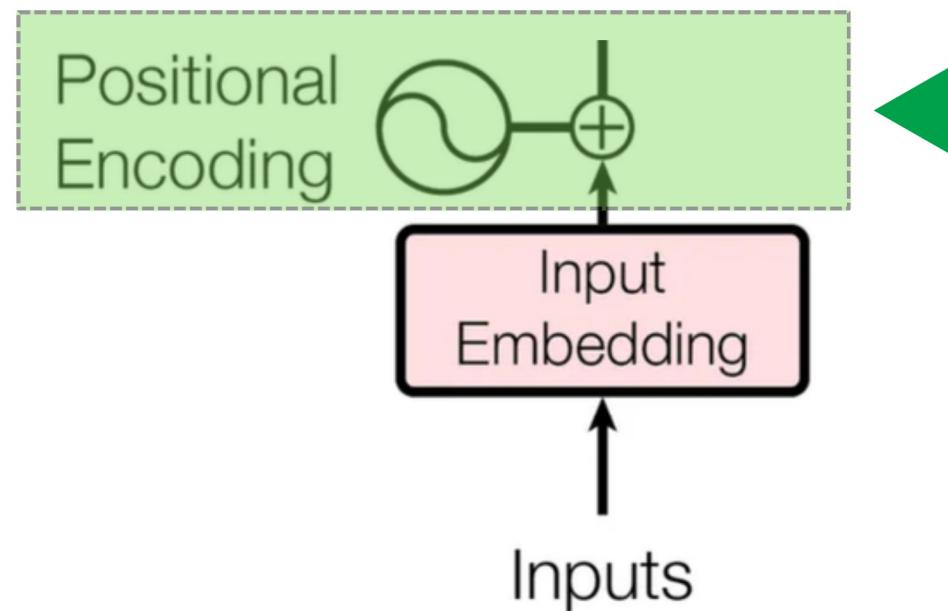
For each position  $pos$  and dimension  $i$ :

$$\left[ \begin{array}{l} PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \\ PE(pos, 2i + 1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \end{array} \right]$$



Here,  $d_{model}$  is the dimension of the model.

$d_{model}$  = embedding dimension



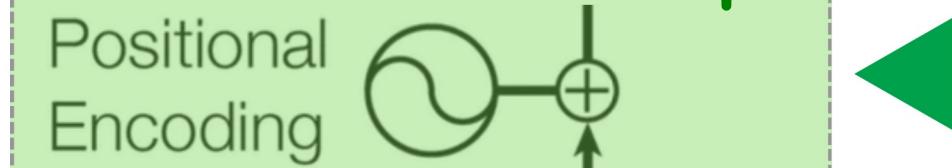
دستورات مخصوصی را در متن معرفی کنید

$I \rightsquigarrow pos = 0 \rightsquigarrow d = 4 \rightarrow [i=0, i=1, i=2, i=3]$

$P_E = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$  (افزودن موضعی تکراری) Positional Encoding بخش

$$PE(I, i=0) = \sin\left(\frac{0}{10000^{\frac{2 \times 0}{4}}}\right)$$

$$PE(I, i=0) = \sin(0^\circ) = 0$$



$$i=0 \\ 2i+1=1 \\ pos = 4 \\ i=1 \rightarrow \cos\left(\frac{4}{10000^{\frac{2 \times 0}{4}}}\right) \rightarrow \cos(4) = -0.65364$$

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

$$PE(pos, 2i+1) = \cos(0^\circ) = 1$$

Sequence

| Index of token (k) | pos = 0 | pos = 1 | pos = 2 | pos = 3 | pos = 4 |
|--------------------|---------|---------|---------|---------|---------|
| x                  | i=0     | i=1     | i=2     | i=3     | i=4     |
| really             | pos = 0 | pos = 1 | pos = 2 | pos = 3 | pos = 4 |
| enjoy              | pos = 0 | pos = 1 | pos = 2 | pos = 3 | pos = 4 |
| learning           | pos = 0 | pos = 1 | pos = 2 | pos = 3 | pos = 4 |
| transformers       | pos = 0 | pos = 1 | pos = 2 | pos = 3 | pos = 4 |

position encoding

$d_{model} = \text{embedding dimension} = 4$

$i=0 > pos/(10000^{(0/4)}) = pos/1$

$i=1 > pos/(10000^{(2/4)}) = pos/10000^{0.5} = pos/100$

$\sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$

$d=4$

| i=0                                         | i=1                                         | i=2                          | i=3                          |
|---------------------------------------------|---------------------------------------------|------------------------------|------------------------------|
| $i=0, pos=0$<br>$\sin(0/10000^{(0/4)})=0.0$ | $i=1, pos=0$<br>$\cos(0/10000^{(0/4)})=1.0$ | $\dots$                      | $\dots$                      |
| $\sin(1/10000^{(0/4)})=0.84$                | $\cos(1/10000^{(0/4)})=0.54$                | $\sin(1/10000^{(2/4)})=0.01$ | $\cos(1/10000^{(2/4)})=0.99$ |
| 0.90930                                     | -0.41615                                    | 0.02000                      | 0.99980                      |
| 0.14112                                     | -0.98999                                    | 0.02999                      | 0.99955                      |
| -0.75680                                    | -0.65364                                    | 0.03998                      | 0.99920                      |

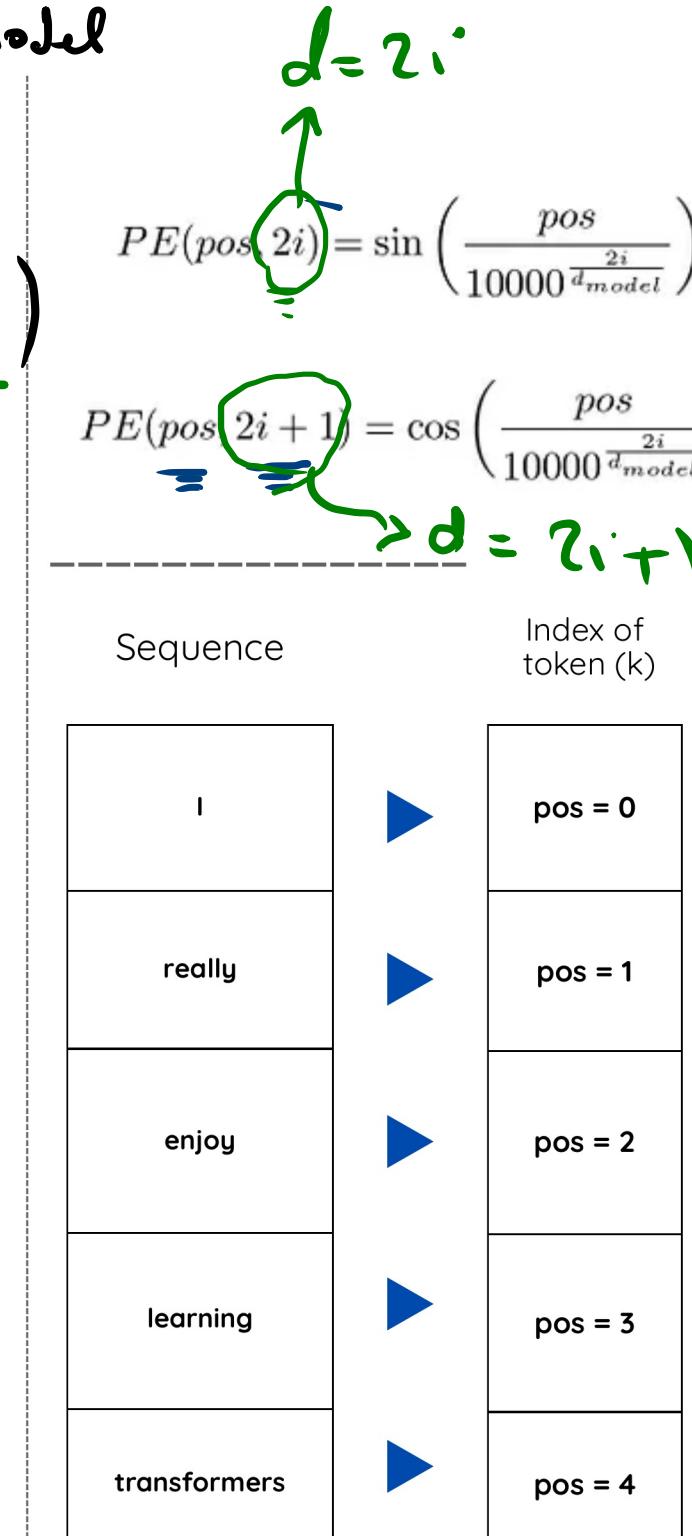
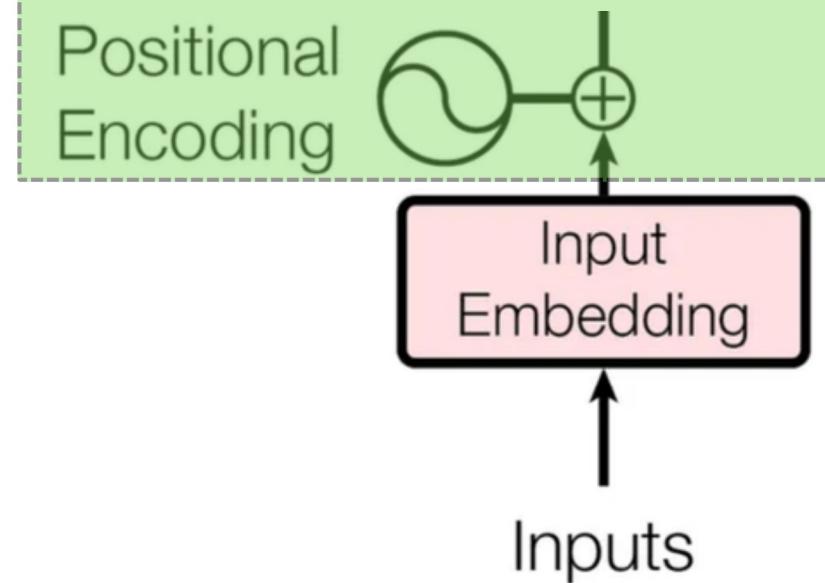
بخش Positional Encoding (افزودن موقعیت توکن ها)

## چگونه position encoding

$$PE(pos, d) = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

$$PE(4, 1) = \cos\left(\frac{4}{10000^{\frac{2 \times 0}{4}}}\right)$$

$$2i+1 = 1 \rightarrow i=0$$

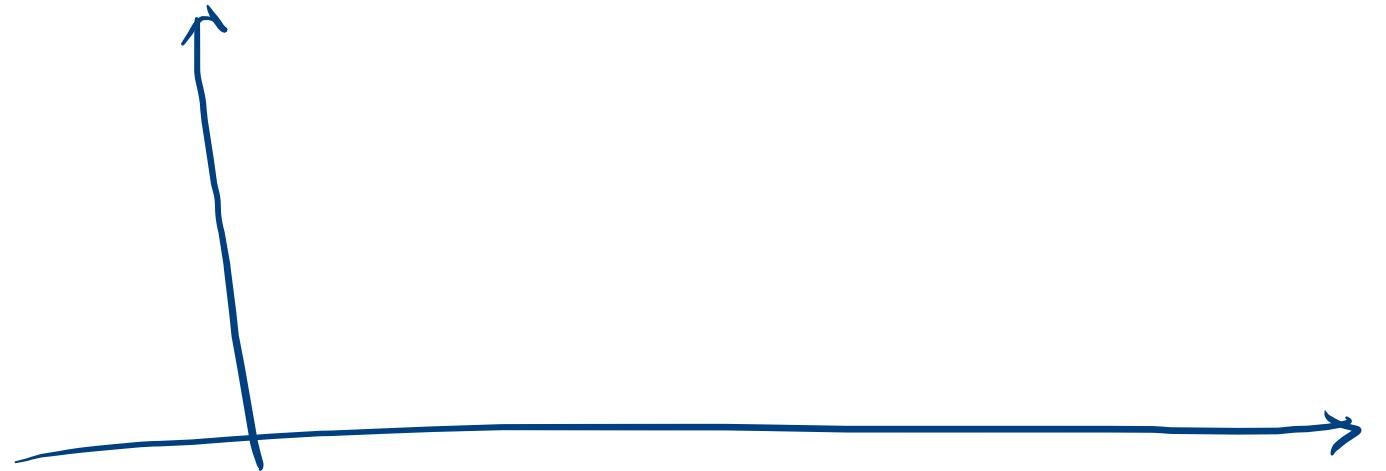


$d_{model} = \text{embedding dimension} = 4$

$i=0 > pos/(10000^{(0/4)}) = pos / 1$

$i=1 > pos/(10000^{(2/4)}) = pos / 10000^{0.5} = pos/100$

| $d=0$                        | $d=1$                        | $d=2$                        | $d=3$                        |
|------------------------------|------------------------------|------------------------------|------------------------------|
| $i=0$                        | $i=1$                        | $i=2$                        | $i=3$                        |
| $\sin(0/10000^{(0/4)})=0.0$  | $\cos(0/10000^{(0/4)})=1.0$  | $\sin(0/10000^{(2/4)})=0.0$  | $\cos(0/10000^{(2/4)})=1.0$  |
| $\sin(1/10000^{(0/4)})=0.84$ | $\cos(1/10000^{(0/4)})=0.54$ | $\sin(1/10000^{(2/4)})=0.01$ | $\cos(1/10000^{(2/4)})=0.99$ |
| 0.90930                      | -0.41615                     | 0.02000                      | 0.99980                      |
| 0.14112                      | -0.98999                     | 0.02999                      | 0.99955                      |
| -0.75680                     | <del>-0.65364</del>          | 0.03998                      | 0.99920                      |



$\sin(\rho^s)$

$\cos(\rho^s)$

I really enjoy -  
pos:0 pos1 pos=2 ...

$\sin(\underline{\circ})$

$\sin(\underline{1})$

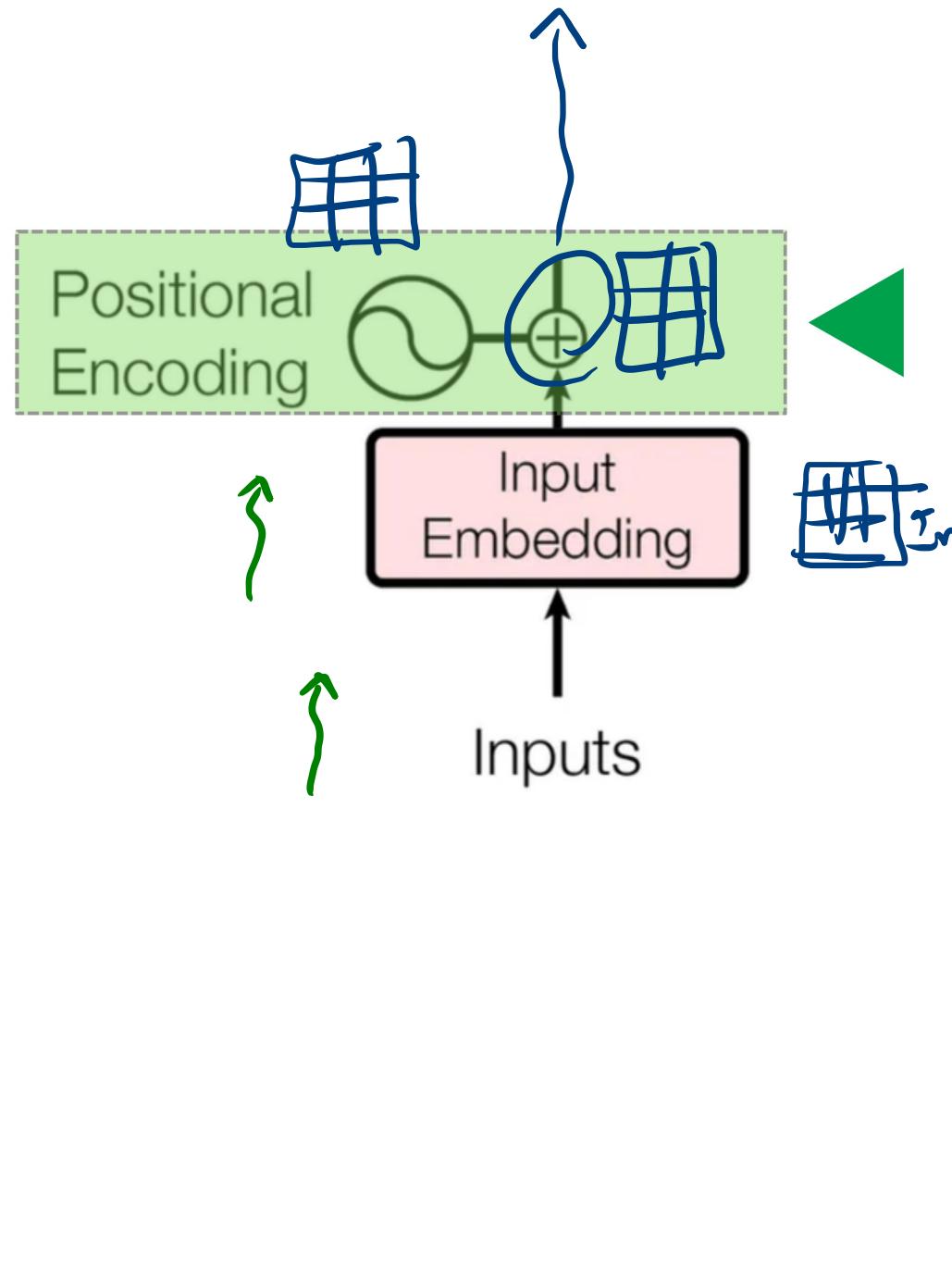
$\sin(\underline{2})$

$\sin(\underline{3})$

نحوه ترکیب اطلاعات ترتیبی (از مرحله ای Input Embedding) و بردار هر توکن (کلمه) (از مرحله ای Positional Encoding)

برای هر توکن (کلمه)، بردار position encoding با بردار embedding جمع می شود.

$$X = E + PE$$



|              | Input |
|--------------|-------|
|              | —     |
| really       |       |
| enjoy        |       |
| learning     |       |
| transformers |       |

|     |     |     |     |
|-----|-----|-----|-----|
| 0.1 | 0.3 | 0.4 | 0.8 |
| 0.9 | 0.1 | 0.5 | 0.2 |
| 0.7 | 0.6 | 0.1 | 0.4 |
| 0.2 | 0.8 | 0.3 | 0.7 |
| 0.5 | 0.4 | 0.9 | 0.3 |

|       |       |       |       |
|-------|-------|-------|-------|
| 0.0   | 1.0   | 0.0   | 1.0   |
| 0.84  | 0.54  | 0.01  | 0.99  |
| 0     | -0.41 | 0.02  | 0.99  |
| 0.14  | -0.98 | 0.029 | 0.99  |
| -0.75 | -0.65 | 0.039 | 0.999 |

|       |       |      |      |
|-------|-------|------|------|
| 0.10  | 1.30  | 0.40 | 1.80 |
| 1.74  | 0.64  | 0.51 | 1.19 |
| 1.60  | 0.18  | 0.12 | 1.39 |
| 0.34  | -0.18 | 0.32 | 1.69 |
| -0.25 | -0.25 | 0.93 | 1.29 |

[concept matching]

[Text 1] , [Text 2]

ایران سرکرد کو ایضاً  
ايران منطق پهلوی را بگوی



## از این ماتریس (خروجی بخش) (Positional Encoding) مدل چگونه ترتیب کلمات را می فهمد؟



Y

X

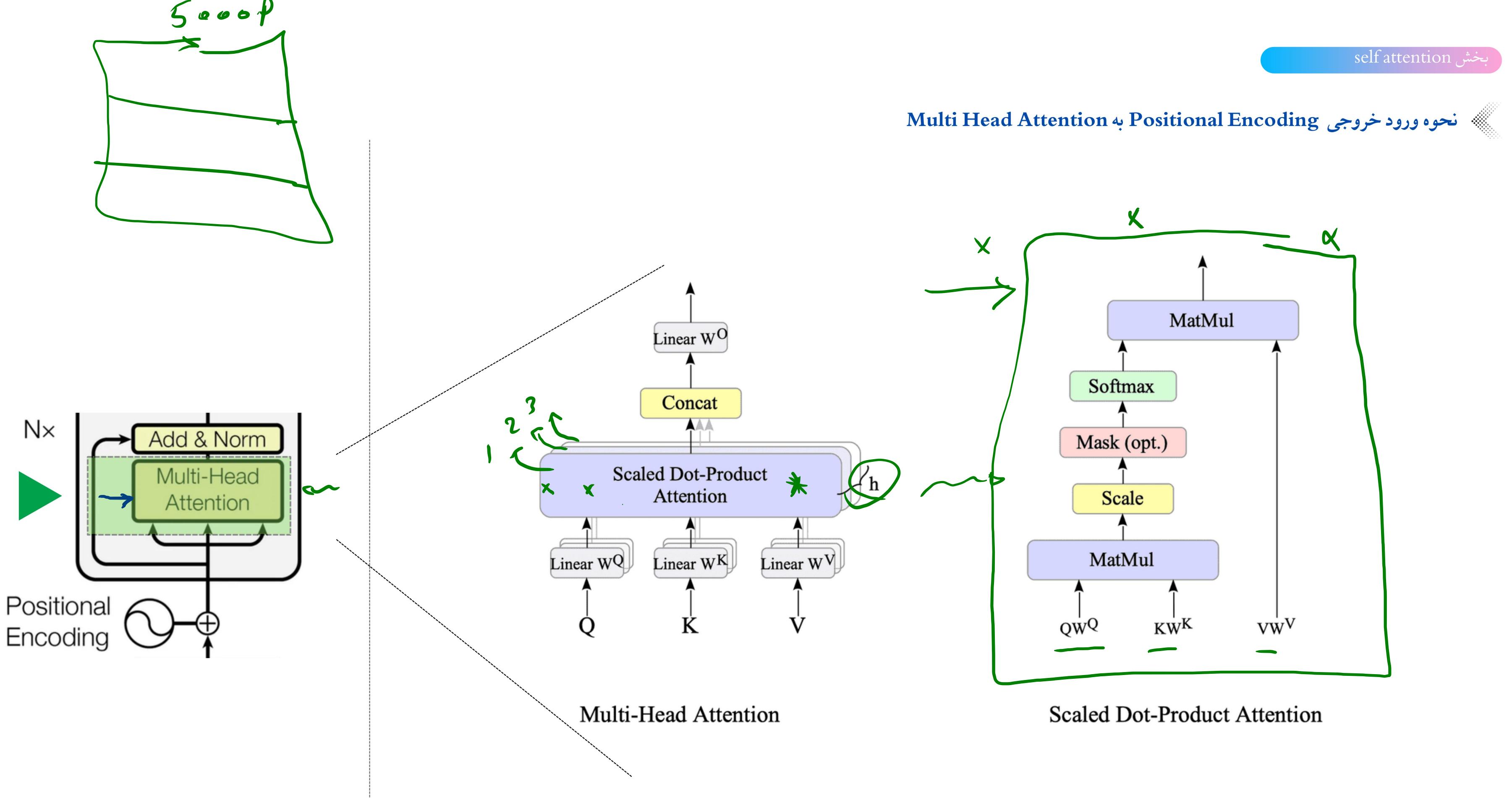
|              |
|--------------|
| I            |
| really       |
| enjoy        |
| learning     |
| transformers |

|       |       |      |      |
|-------|-------|------|------|
| 0.10  | 1.30  | 0.40 | 1.80 |
| 1.74  | 0.64  | 0.51 | 1.19 |
| 1.60  | 0.18  | 0.12 | 1.39 |
| 0.34  | -0.18 | 0.32 | 1.69 |
| -0.25 | -0.25 | 0.93 | 1.29 |

5000P

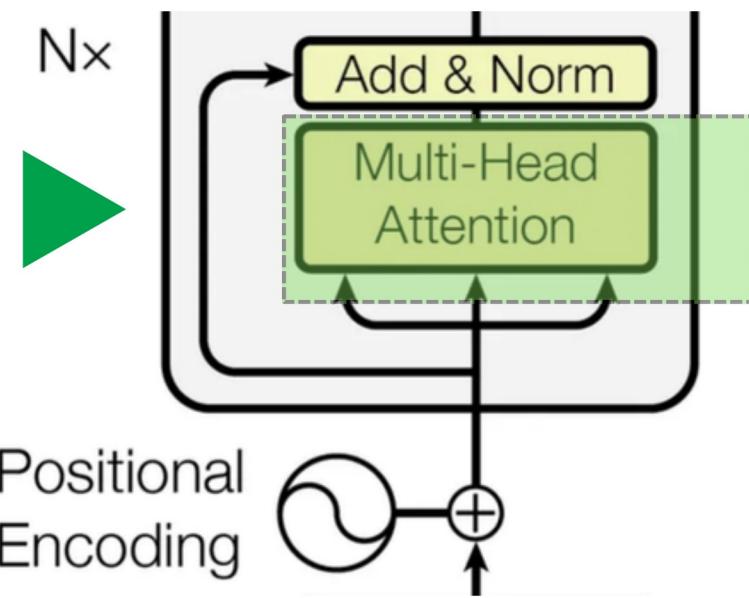
بخش self attention

نحوه ورود خروجی Multi Head Attention & Positional Encoding



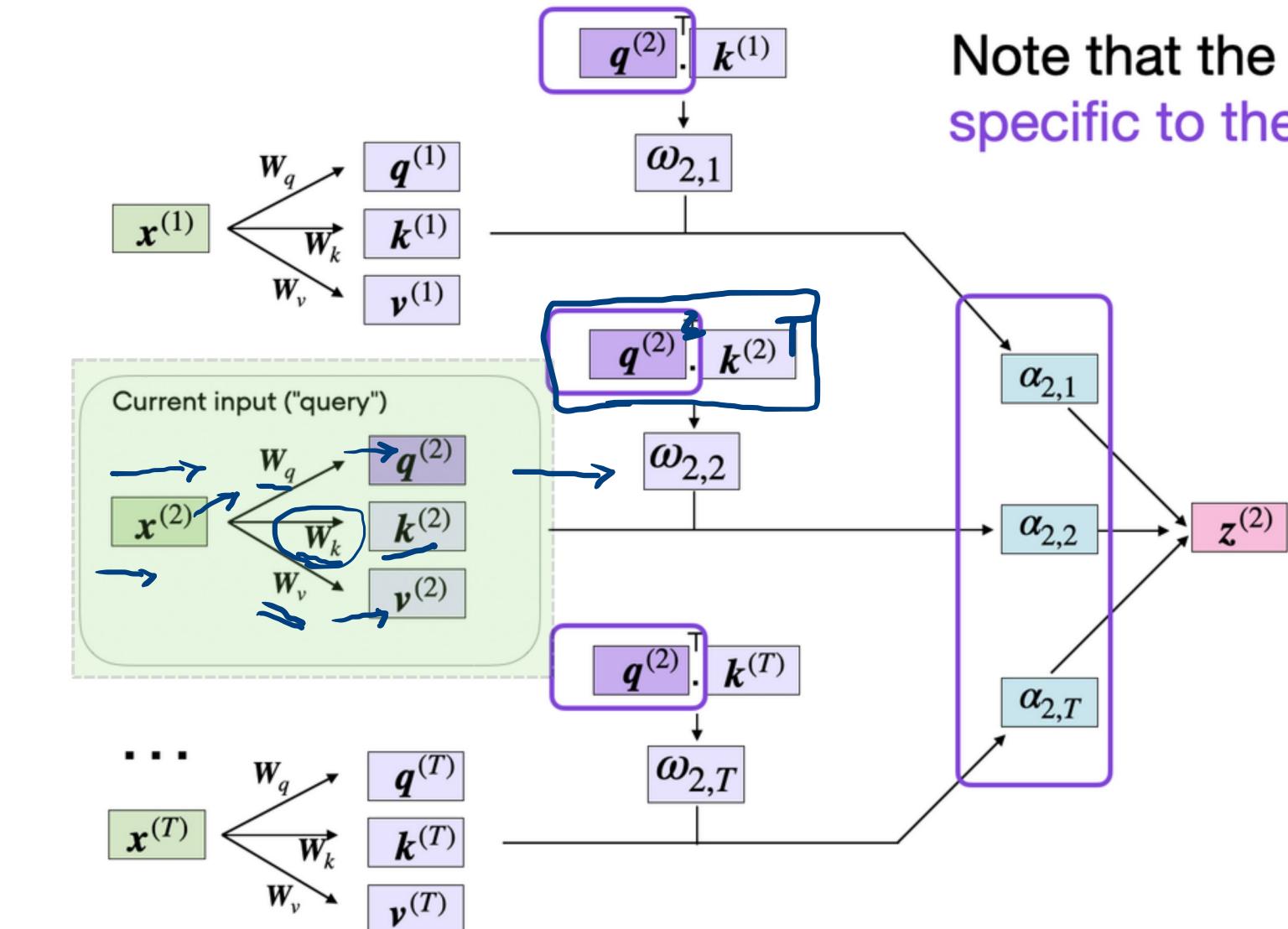
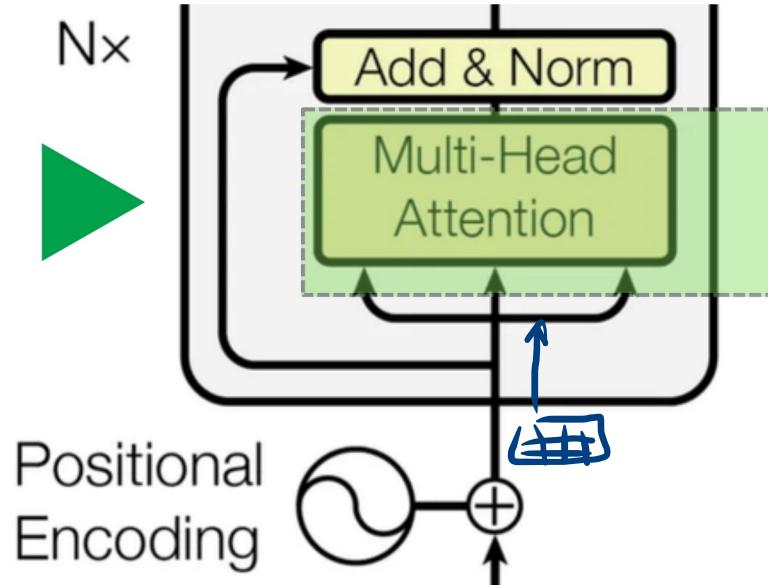
## فرمول مکانیزم

$$\frac{Q K^T}{\sqrt{d_k}} \rightarrow \text{Soft} \left( \frac{Q K^T}{\sqrt{d_k}} \right) \rightarrow$$



$\times \quad \text{Attention}(Q, K, V) = \underbrace{\text{softmax}}_{\rightarrow} \left( \frac{Q K^T}{\sqrt{d_k}} \right) V$

where Q, K and V are Query, Key and Value vectors.



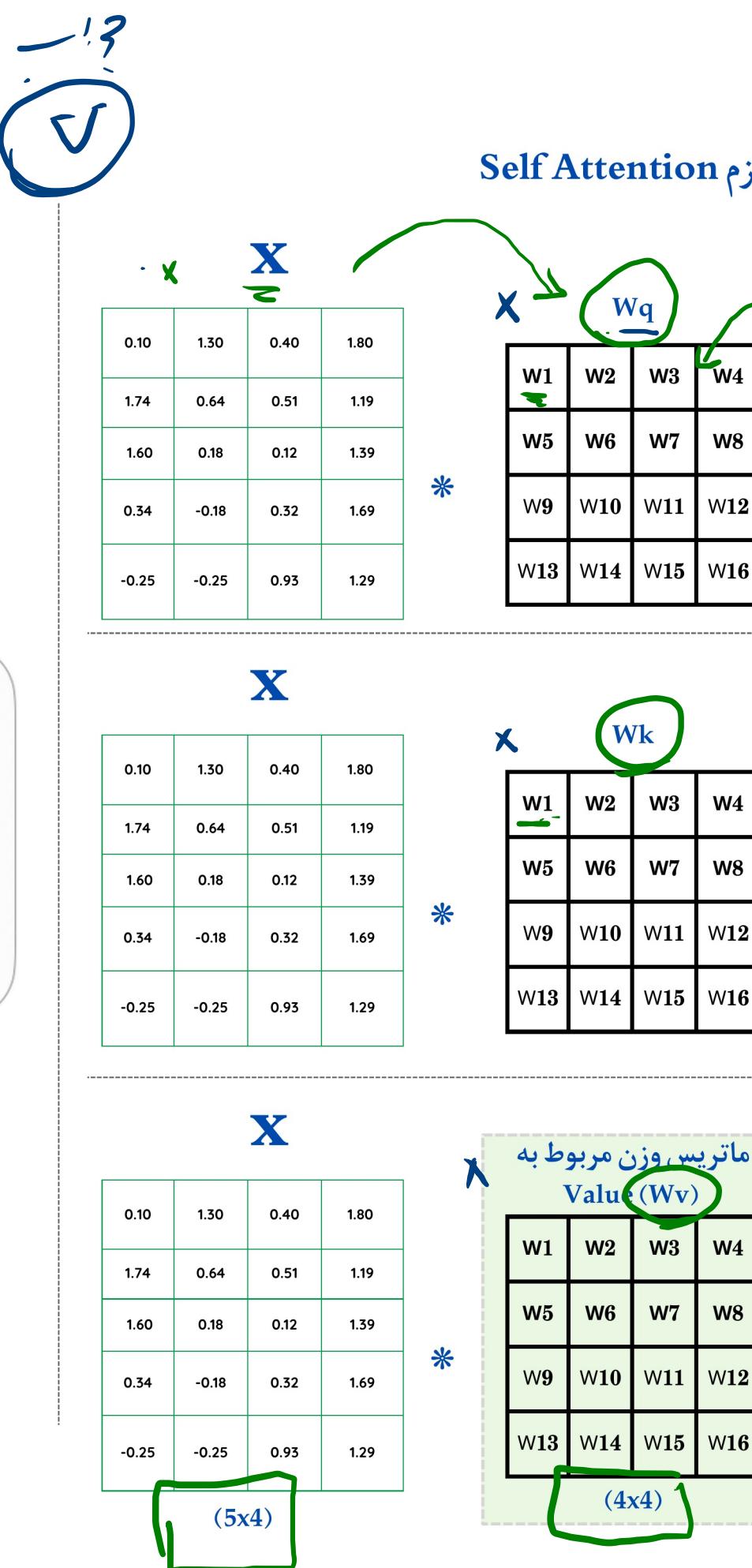
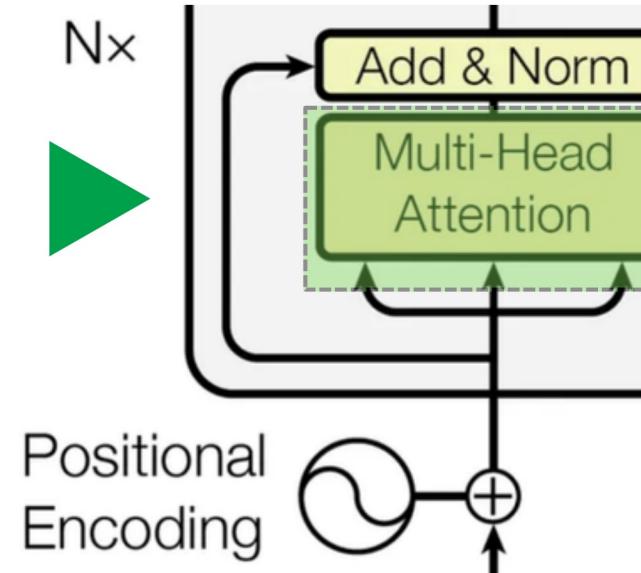
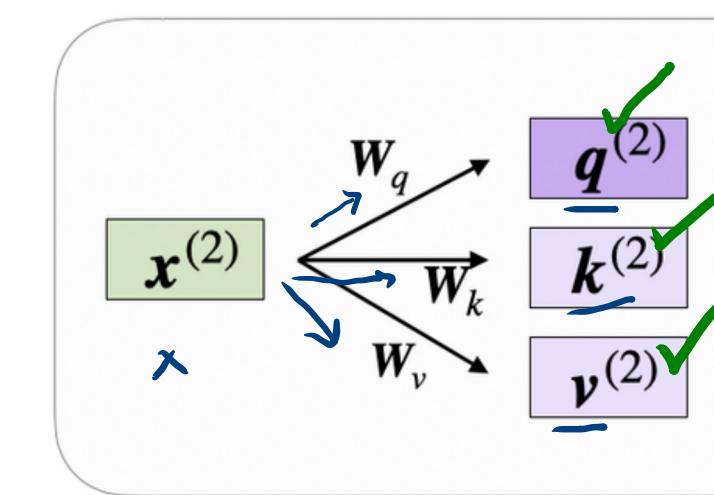
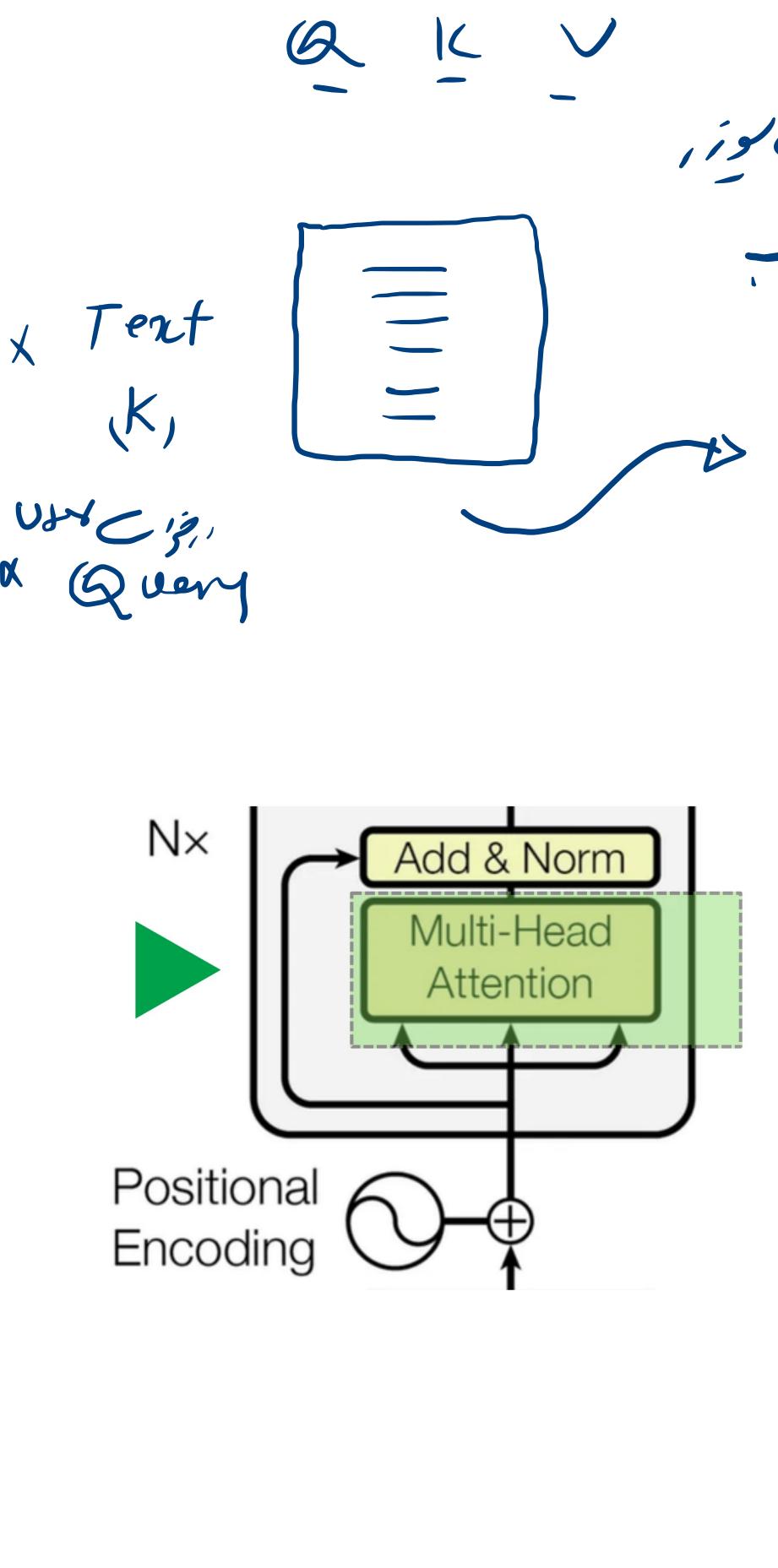
Note that the attention scores are specific to the current input token

$$\text{where } z^{(2)} = \sum_{j=1}^T \alpha_{2,j} v^{(j)}$$

نمایش دیگری از معماری مکانیزم self attention



self attention بخش



One End

Q:

→ I really enjoy

K:

