ATTENTION IS ALL YOU NEED

Pavan Sanjana Cirruguri

XK48735

PAPER SUMMARY

- Introduces the Transformer model
- Solely based on attention mechanisms
- Eliminates RNN and CNN

PAPER SUMMARY - THE TRANSFORMER MODEL

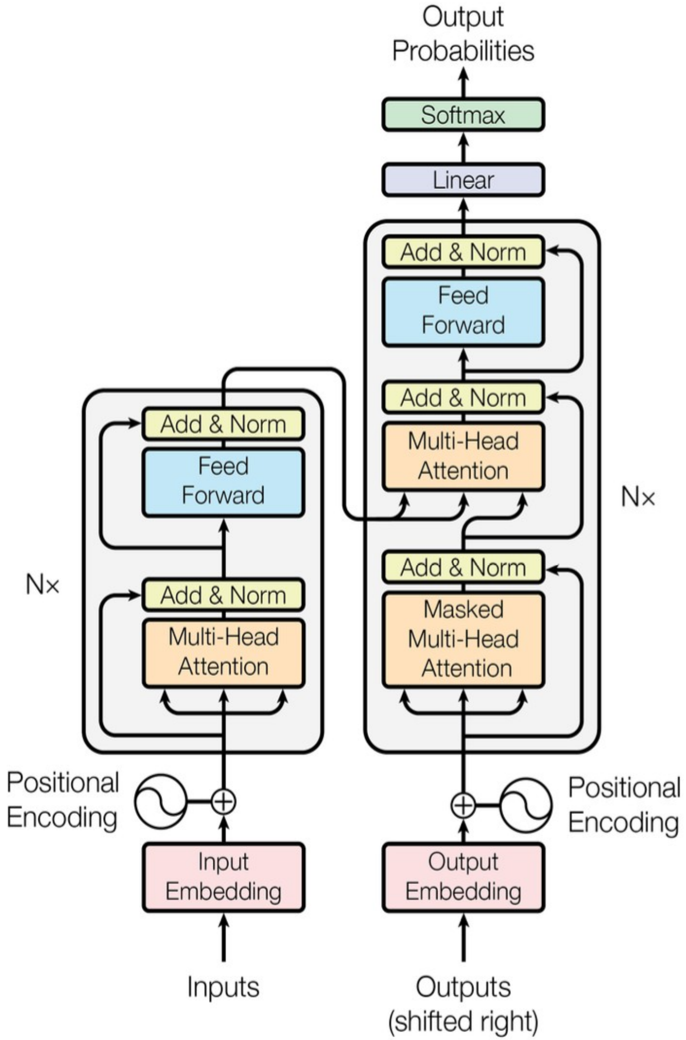
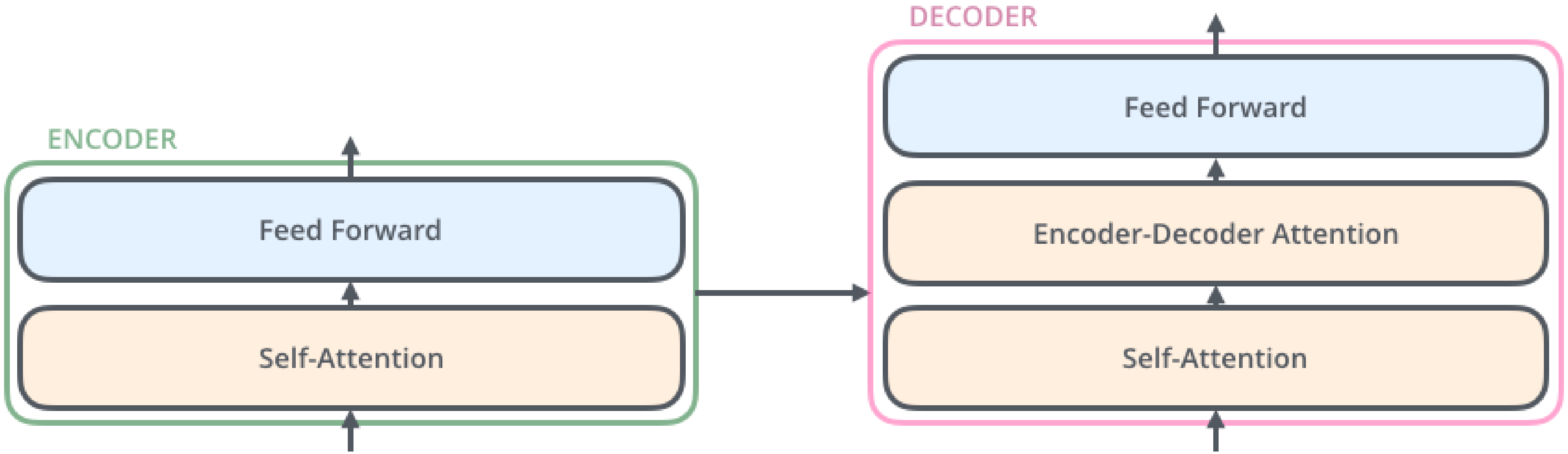


Figure 1: The Transformer - model architecture.

source - https://jalammar.github.io/illustrated-transformer/

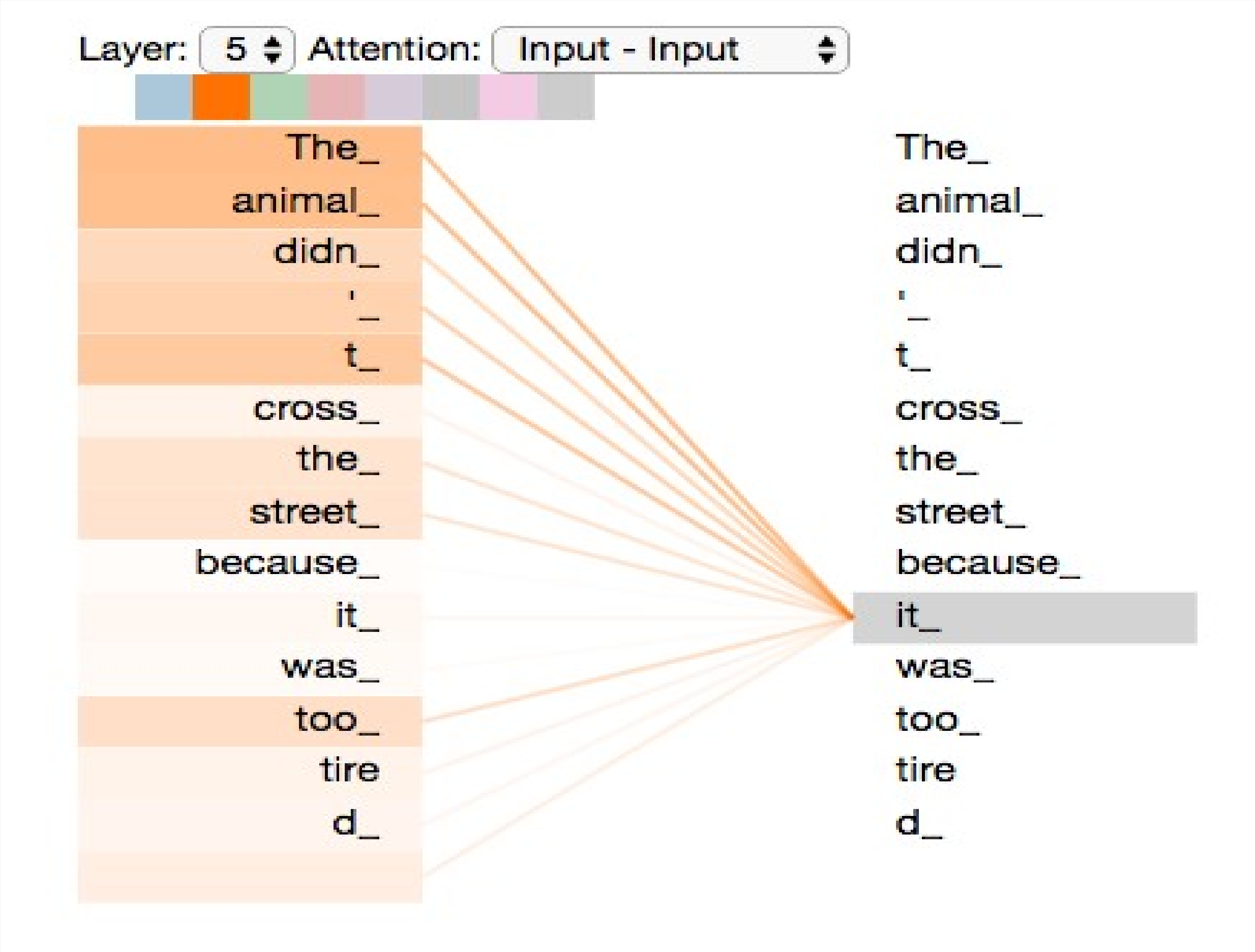
PAPER SUMMARY - TRANSFORMER SIMPLIFIED



KEY INNOVATIONS - ATTENTION MECHANISM

The animal didn't cross the street because it was too tired

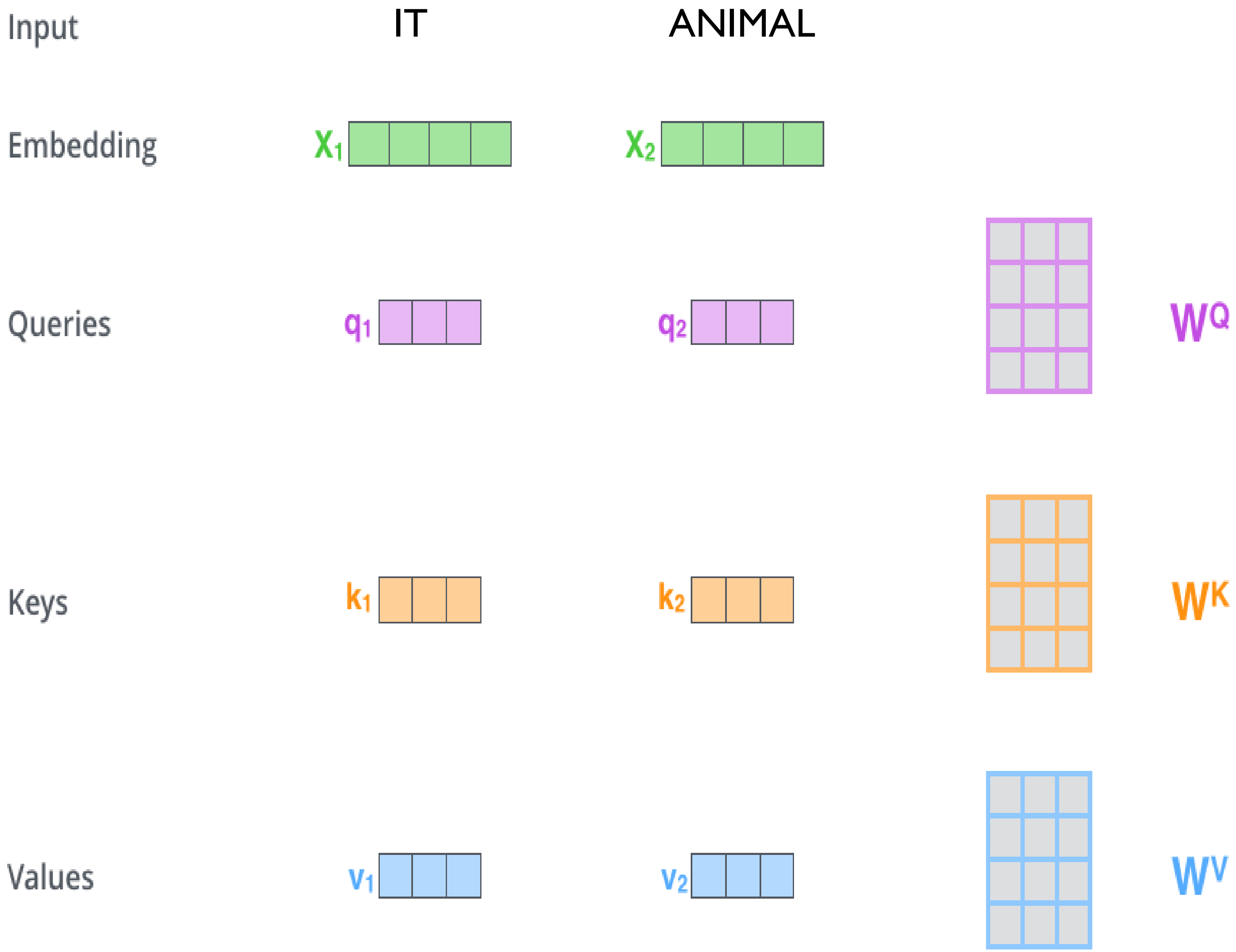
KEY INNOVATIONS - ATTENTION MECHANISM



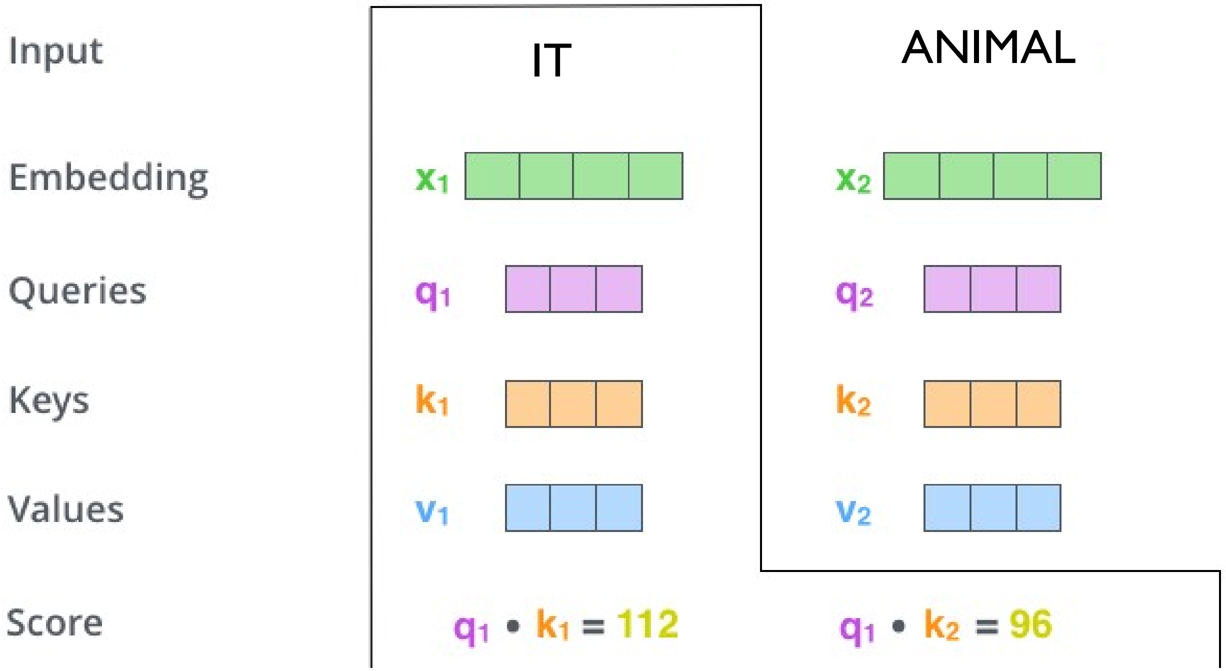
Layer: 5

Attention: Input - Input

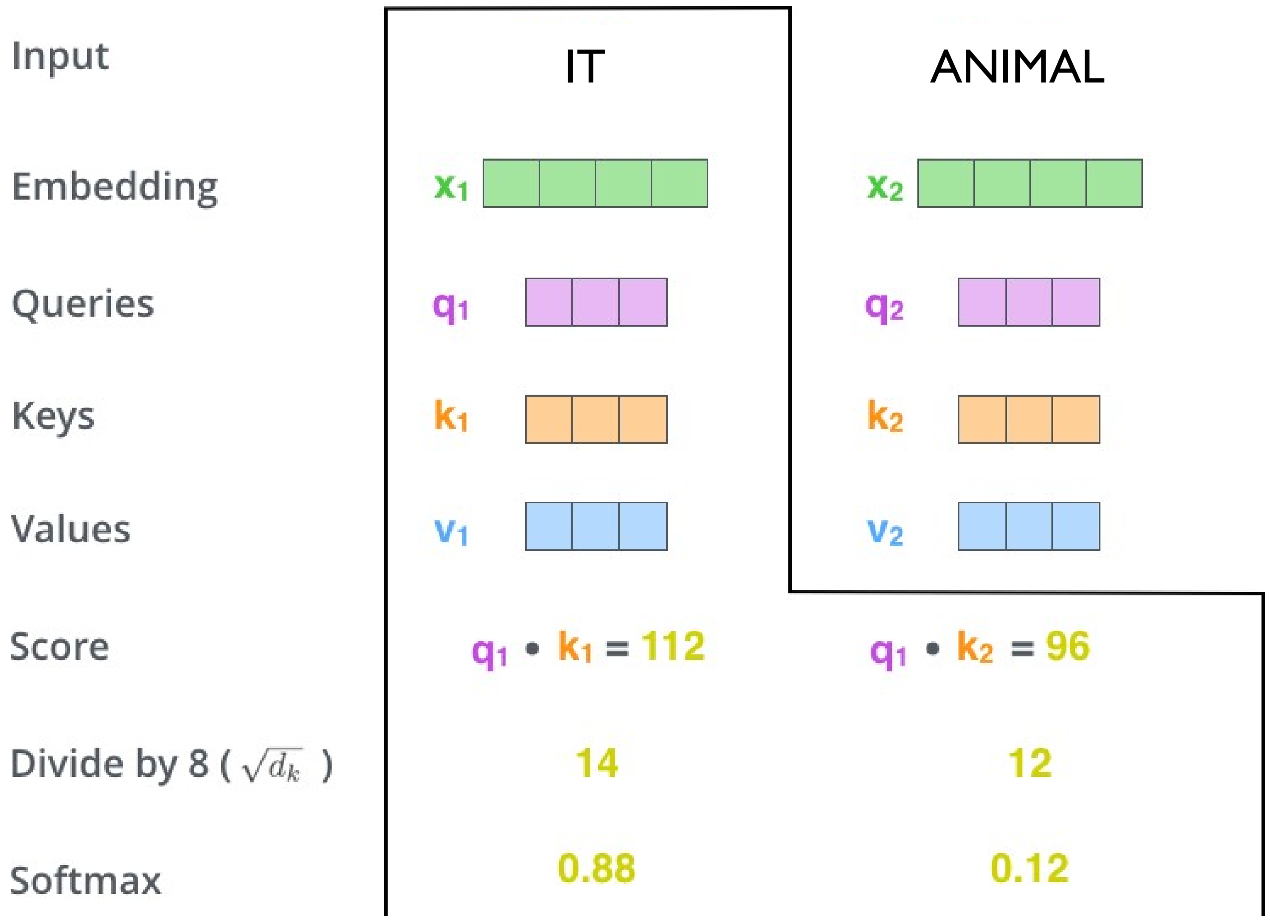
KEY INNOVATIONS - ATTENTION MECHANISM



KEY INNOVATIONS - ATTENTION MECHANISM

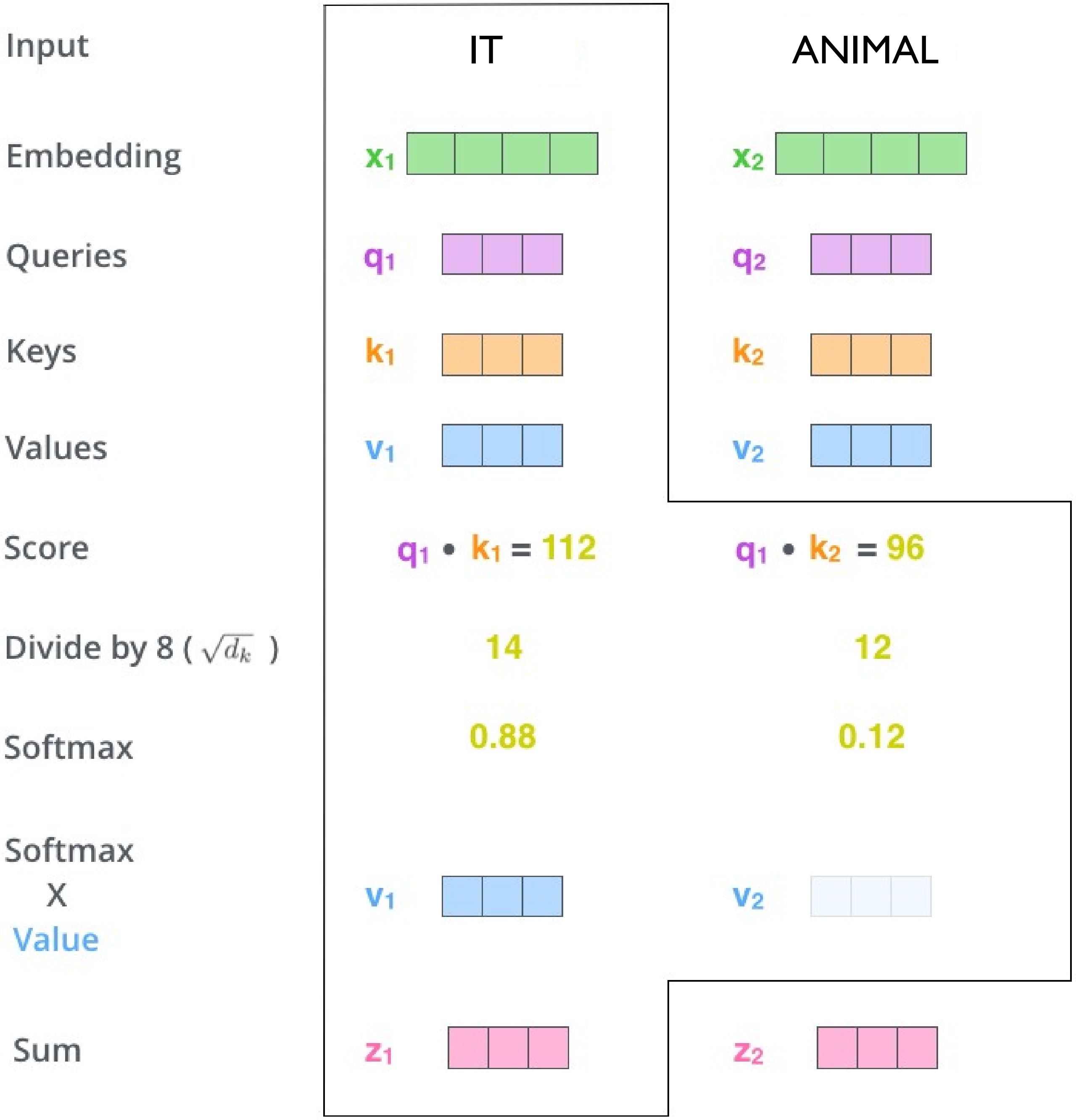


KEY INNOVATIONS - ATTENTION MECHANISM



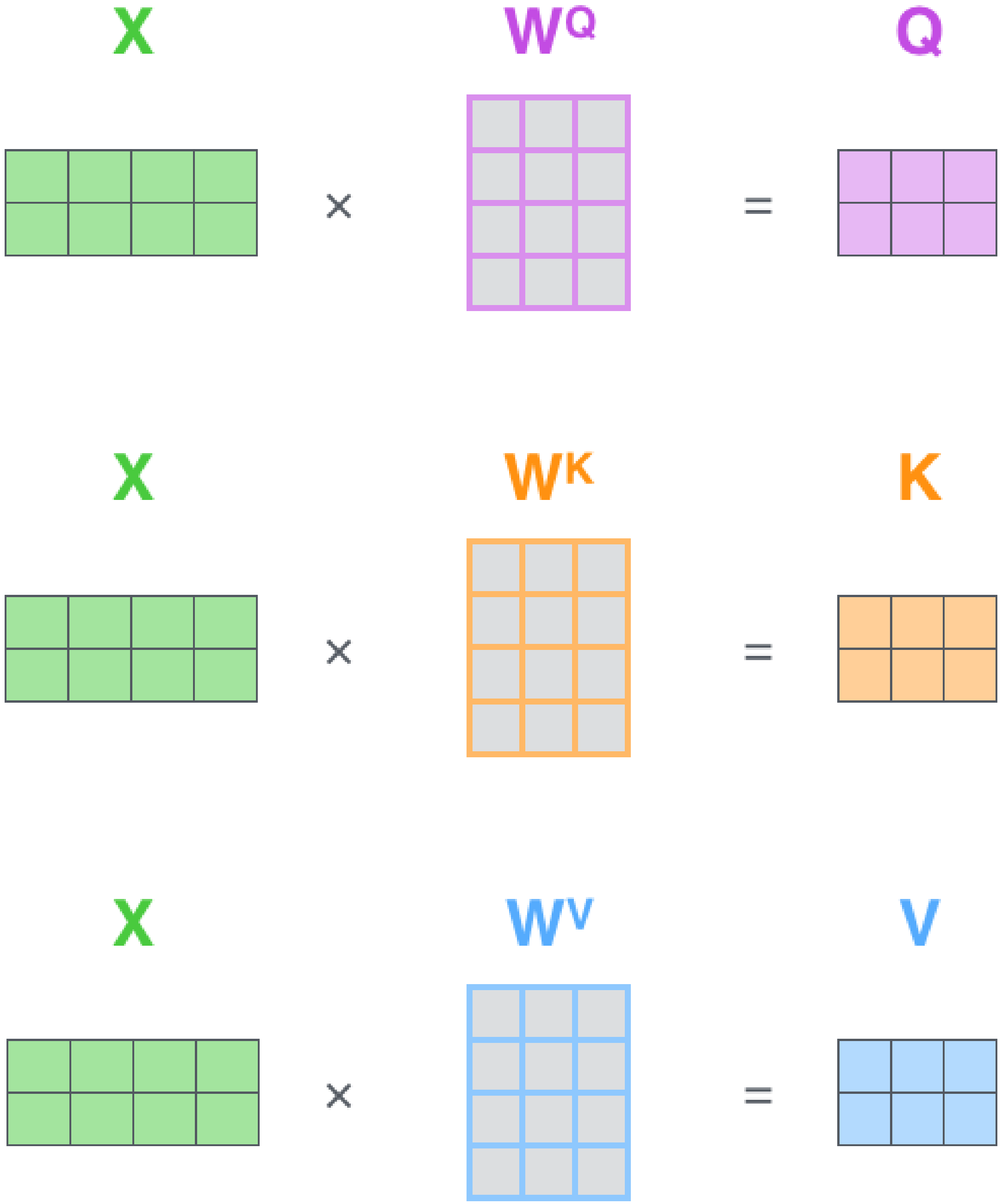
Input	IT	ANIMAL
Embedding	x_1 🟩🟩🟩🟩	x_2 🟩🟩🟩🟩
Queries	q_1 🟪🟪🟪	q_2 🟪🟪🟪
Keys	k_1 🟧🟧🟧	k_2 🟧🟧🟧
Values	v_1 🟦🟦🟦	v_2 🟦🟦🟦
Score	$q_1 \bullet k_1 = 112$	$q_1 \bullet k_2 = 96$
Divide by 8 ($\sqrt{d_k}$)	14	12
Softmax	0.88	0.12

KEY INNOVATIONS - ATTENTION MECHANISM

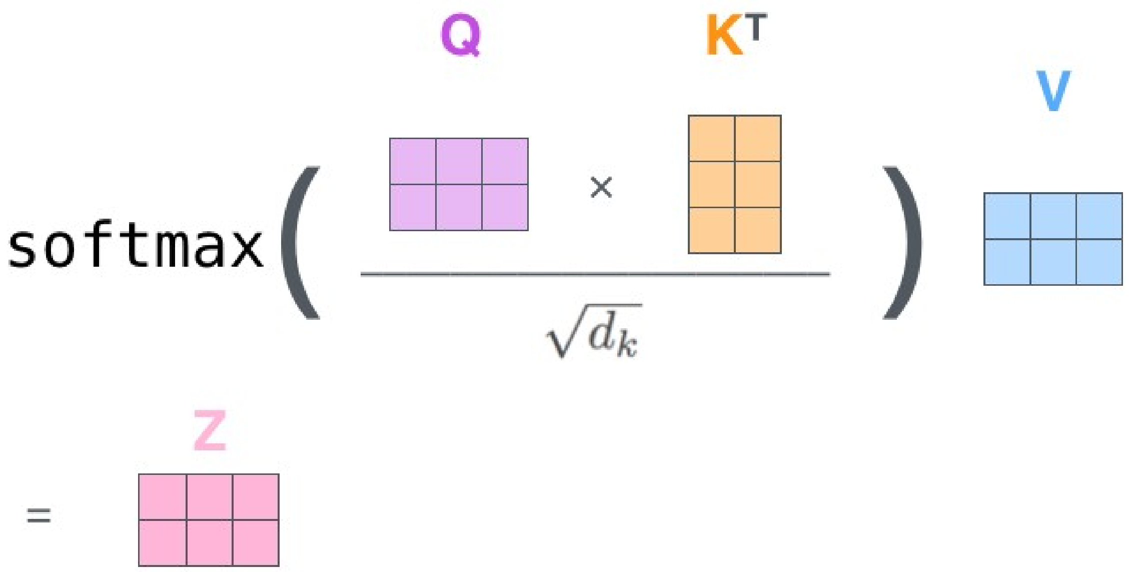


Input	IT	ANIMAL
Embedding	x_1	x_2
Queries	q_1	q_2
Keys	k_1	k_2
Values	v_1	v_2
Score	$q_1 \bullet k_1 = 112$	$q_1 \bullet k_2 = 96$
Divide by 8 ($\sqrt{d_k}$)	14	12
Softmax	0.88	0.12
Softmax X Value	v_1	v_2
Sum	z_1	z_2

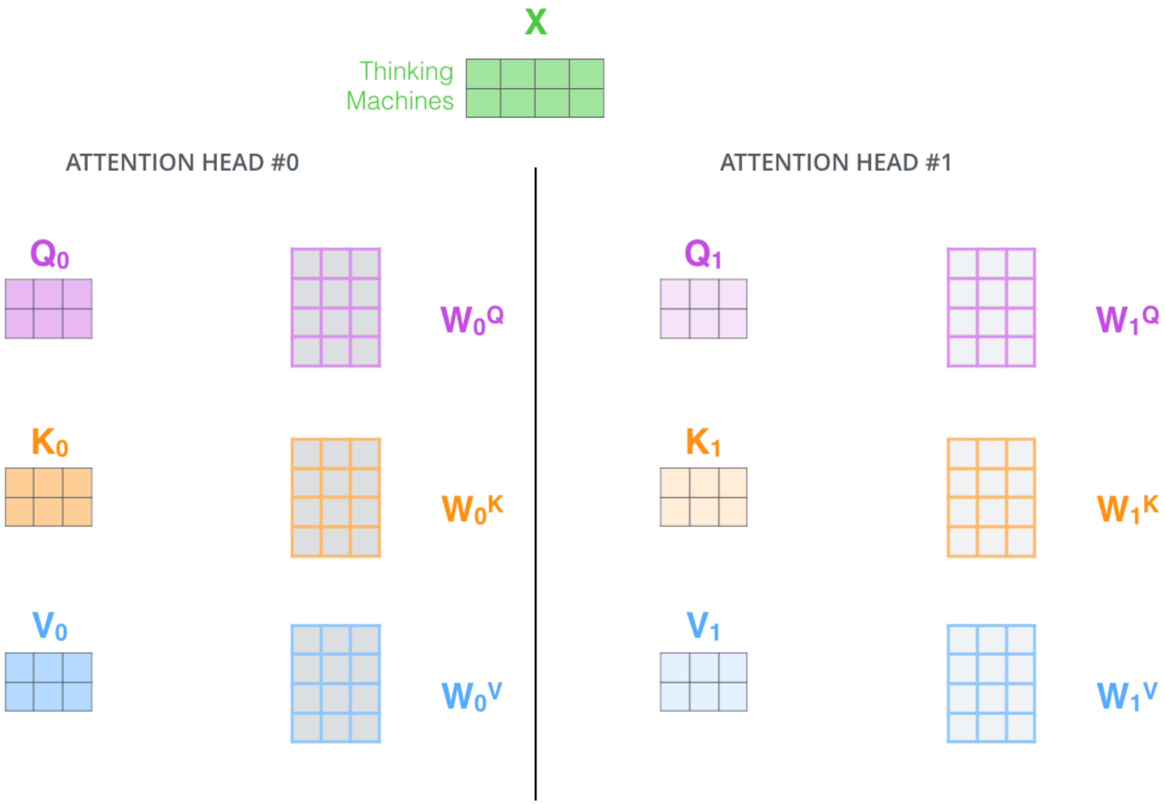
MATRIX CALCULATION OF SELF-ATTENTION



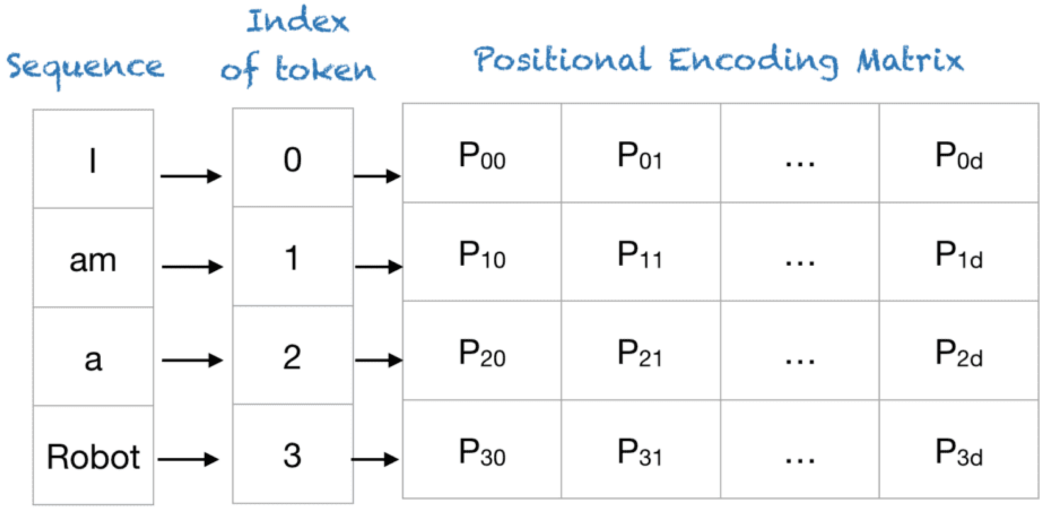
MATRIX CALCULATION OF SELF-ATTENTION



MULTI HEAD ATTENTION



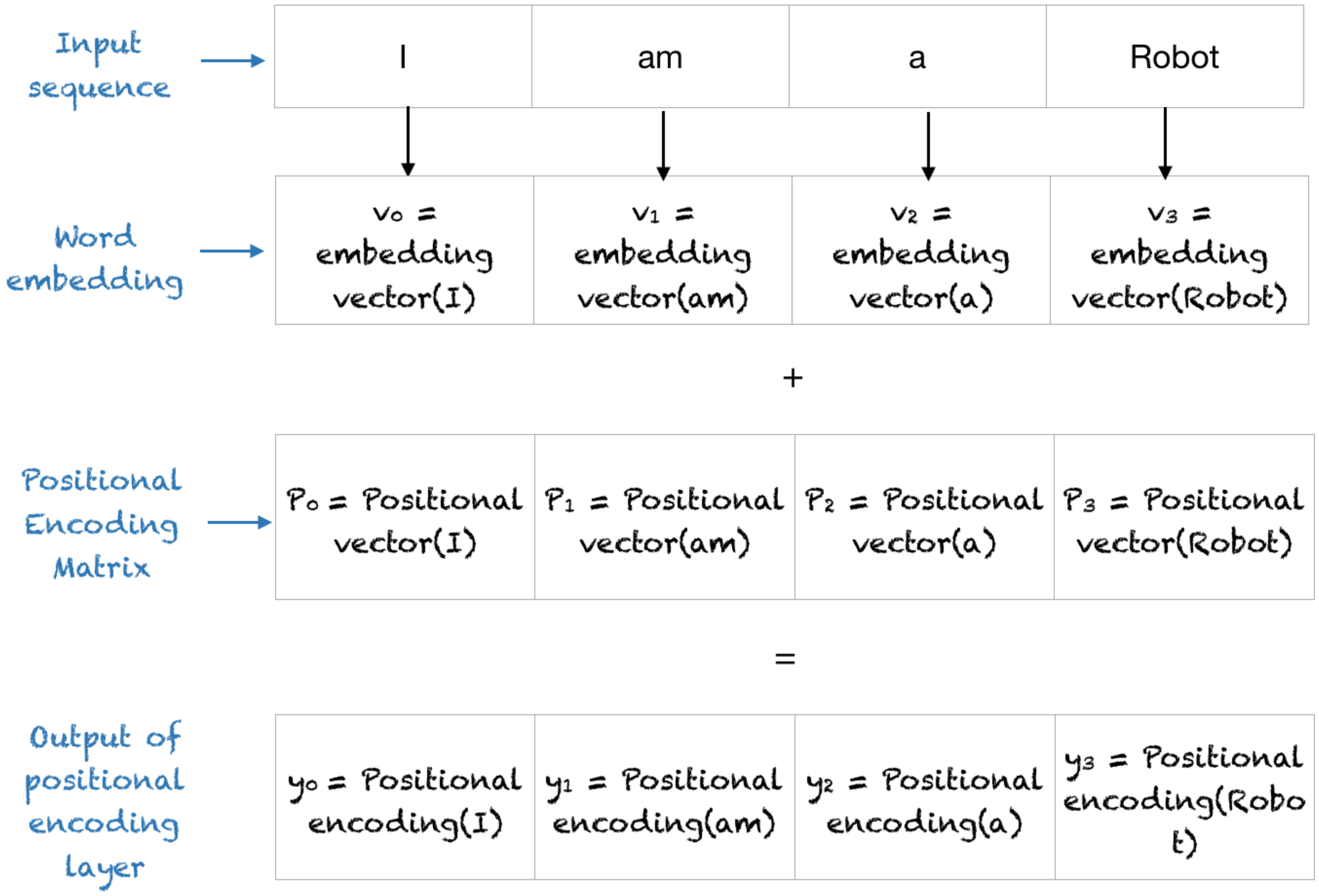
POSITIONAL ENCODING



Sequence	Index of token	Positional Encoding Matrix			
I	0	P_{00}	P_{01}	…	P_{0d}
am	1	P_{10}	P_{11}	…	P_{1d}
a	2	P_{20}	P_{21}	…	P_{2d}
Robot	3	P_{30}	P_{31}	…	P_{3d}

Positional Encoding Matrix for the sequence 'I am a robot'

POSITIONAL ENCODING



Input sequence	I	am	a	Robot
Word embedding	V_0 = embedding vector(I)	V_1 = embedding vector(am)	V_2 = embedding vector(a)	V_3 = embedding vector(Robot)
		+		
Positional Encoding Matrix	P_0 = Positional vector(I)	P_1 = Positional vector(am)	P_2 = Positional vector(a)	P_3 = Positional vector(Robot)
		=		
Output of positional encoding layer	y_0 = Positional encoding(I)	y_1 = Positional encoding(am)	y_2 = Positional encoding(a)	y_3 = Positional encoding(Robot)

STRENGTHS

- Achieves state-of-the-art performance on translation tasks
- Parallelizable, reducing training time
- Captures long-range dependencies effectively
- Interpretability through attention visualizations

WEAKNESSES

- Computationally intensive for very long sequences
- Lack of inherent sequence order modeling (mitigated by positional encoding)
- Potential overfitting on small datasets

RELATION TO STORY GENERATION

- Can generate coherent and contextually relevant text
- Attention mechanism allows for better understanding of context and relationships between words/ideas
- Potential for generating diverse and creative storylines

APPLICATION TO INTERACTIVE FICTION

- Can be used to generate dynamic responses to user inputs
- Attention mechanism helps maintain consistency in long-term narrative
- Potential for creating more engaging and personalized interactive experiences

FUTURE DIRECTIONS

- Exploring Transformer variants for specific story generation tasks
- Investigating methods to control and guide the generation process
- Combining Transformer models with other techniques for enhanced storytelling

CONCLUSION

- Transformer revolutionizes sequence modeling with attention.
- Excels in translation tasks and trains faster than previous models.
- Challenges remain for very long sequences and small datasets.
- Potent for story generation and interactive fiction.
- Transformer, a cornerstone of modern NLP.

REFERENCES

- https://arxiv.org/pdf/1706.03762
- https://jalammar.github.io/illustrated-transformer/