Attention is all you need

RNN is unnecessary

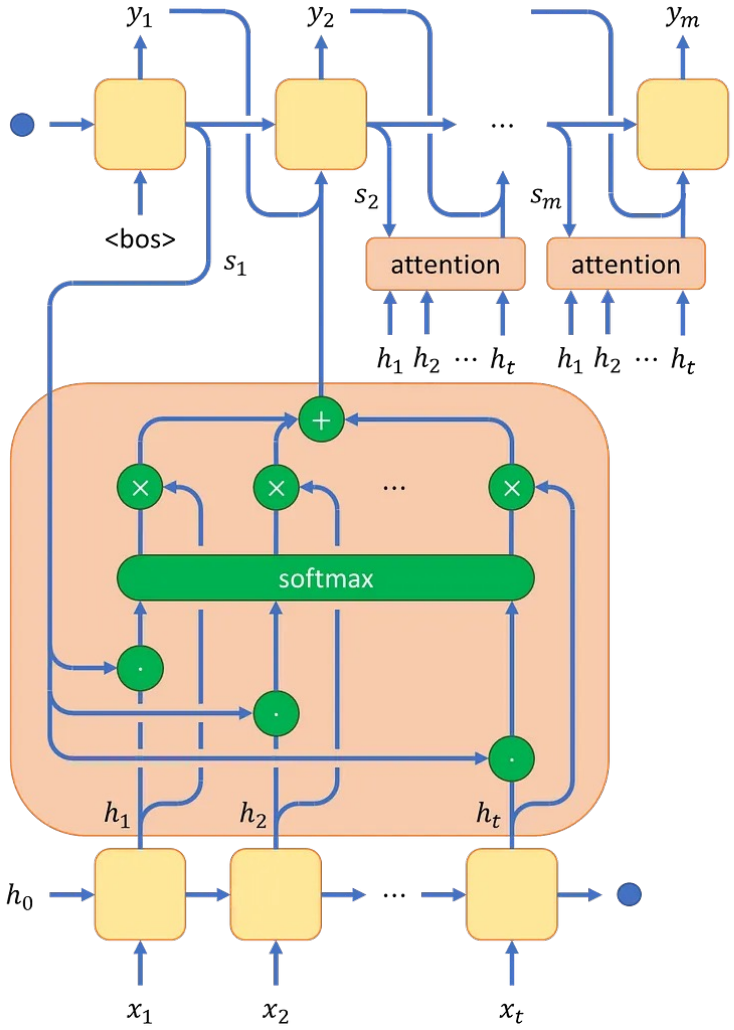
Liang Hong

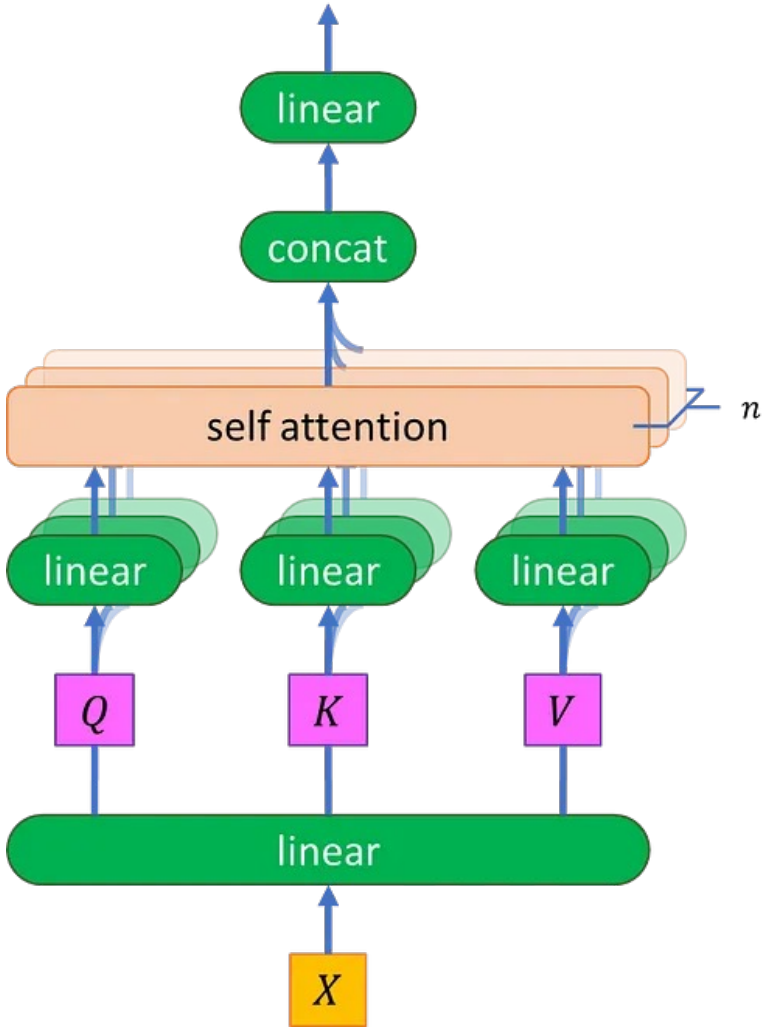
17 Mar, 2023

Outline

- How inference change from RNN style to Transformer style
- How attention deal with token order
- What’s next

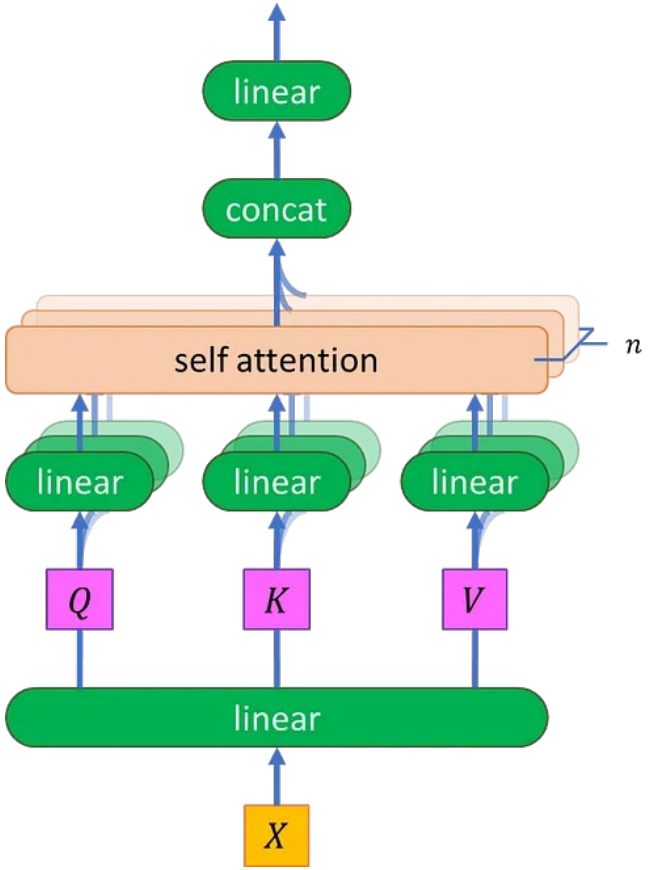
Inference style change





Token order

- Without sequential input, how token order is perceived?



Attention is performed with dot product, and the previous linear layers are shared.
-> permutation irrelevant

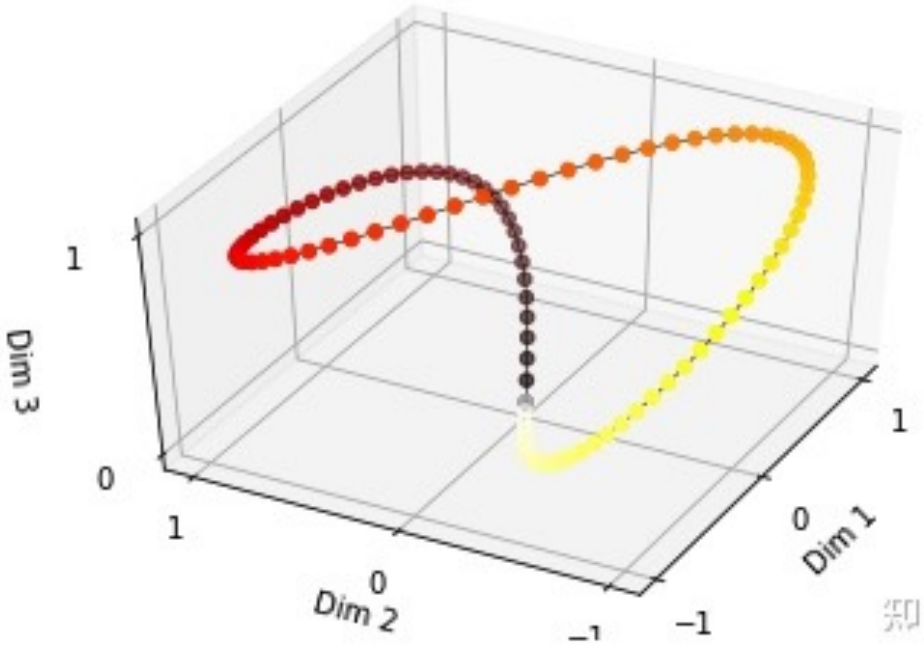
Positional encoding

$$\begin{cases} PE(pos, 2i) = \sin\left(pos/10000^{2i/d_{model}}\right) \\ PE(\text{pos}, 2i+1) = \cos\left(pos/10000^{2i/d_{model}}\right) \end{cases}$$

Why this frequency

❌ int type
❌ 0~1 representation

Credit to @lemonround



Positional encoding

$$\begin{cases} PE(pos, 2i) = \sin\left(pos/10000^{2i/d_{model}}\right) \\ PE(\text{pos}, 2i+1) = \cos\left(pos/10000^{2i/d_{model}}\right) \end{cases}$$

(now we have unique and continuous position rep vectors)

How can we get relative position from position encoding?

$$\begin{aligned} PE_t^T * PE_{t+\triangle t} &= \sum_{i=0}^{\frac{d_{model}}{2}-1} [sin(w_i t) sin(w_i(t+\triangle t) + cos(w_i t) cos(w_i(t+\triangle t)] \\ &= \sum_{i=0}^{\frac{d_{model}}{2}-1} cos(w_i(t-(t+\triangle t))) \\ &= \sum_{i=0}^{\frac{d_{model}}{2}-1} cos(w_i \triangle t) \end{aligned}$$

$$PE_t^{(i)} = \begin{cases} \sin(w_i t), & if\ k = 2i \\ \cos(w_i t), & if\ k = 2i+1 \end{cases}$$

Positional encoding

$$\begin{cases} PE(pos, 2i) = \sin\left(pos/10000^{2i/d_{model}}\right) \\ PE(\text{pos}, 2i+1) = \cos\left(pos/10000^{2i/d_{model}}\right) \end{cases}$$

(now we can get relative positions from positional encoding)

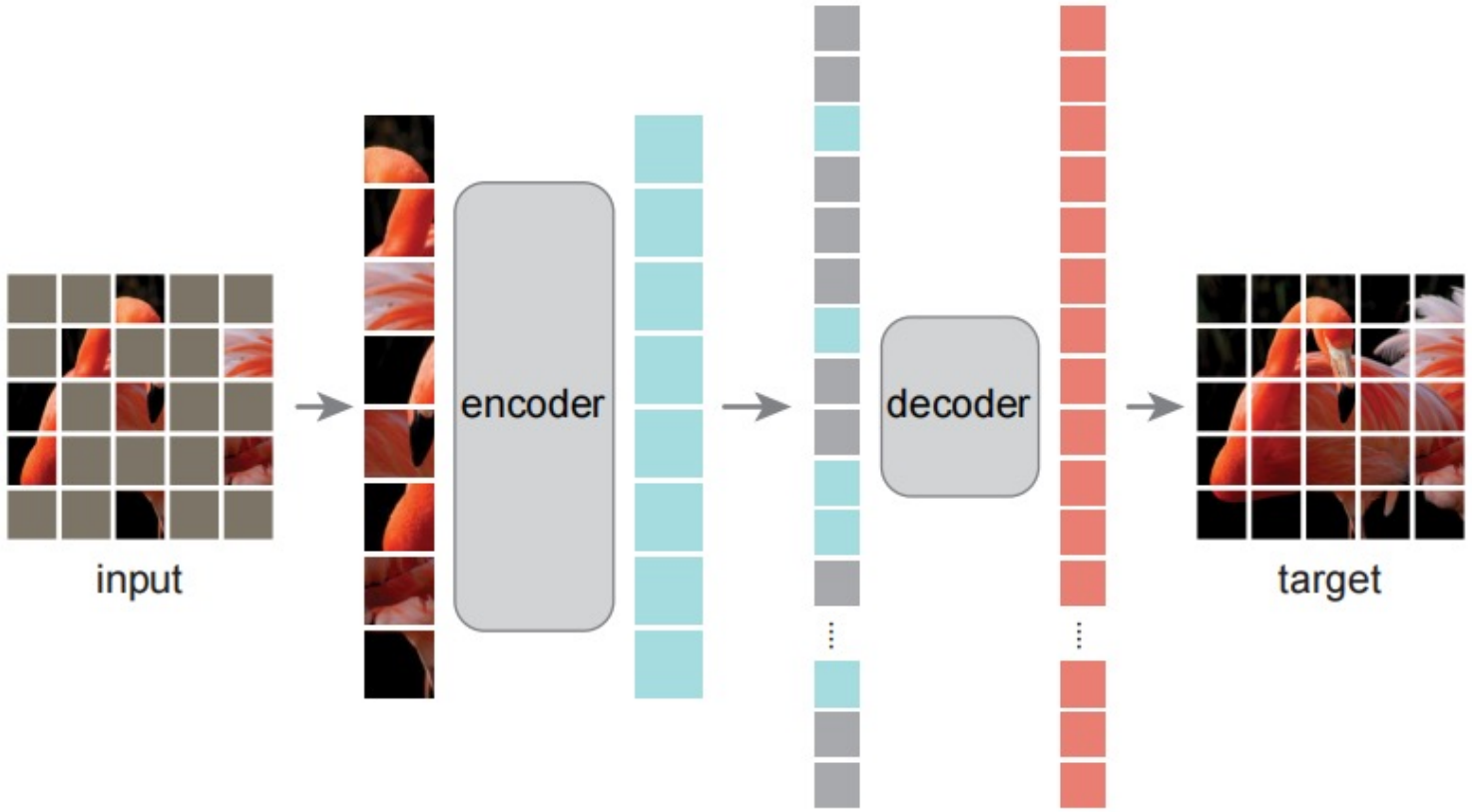
$$\begin{aligned} PE_t^T * PE_{t+\triangle t} &= \sum_{i=0}^{\frac{d_{model}}{2}-1} [sin(w_i t) sin(w_i(t+\triangle t) + cos(w_i t) cos(w_i(t+\triangle t)] \\ &= \sum_{i=0}^{\frac{d_{model}}{2}-1} cos(w_i(t-(t+\triangle t))) \\ &= \sum_{i=0}^{\frac{d_{model}}{2}-1} cos(w_i \triangle t) \end{aligned}$$

✅ distance aware

❌ direction aware

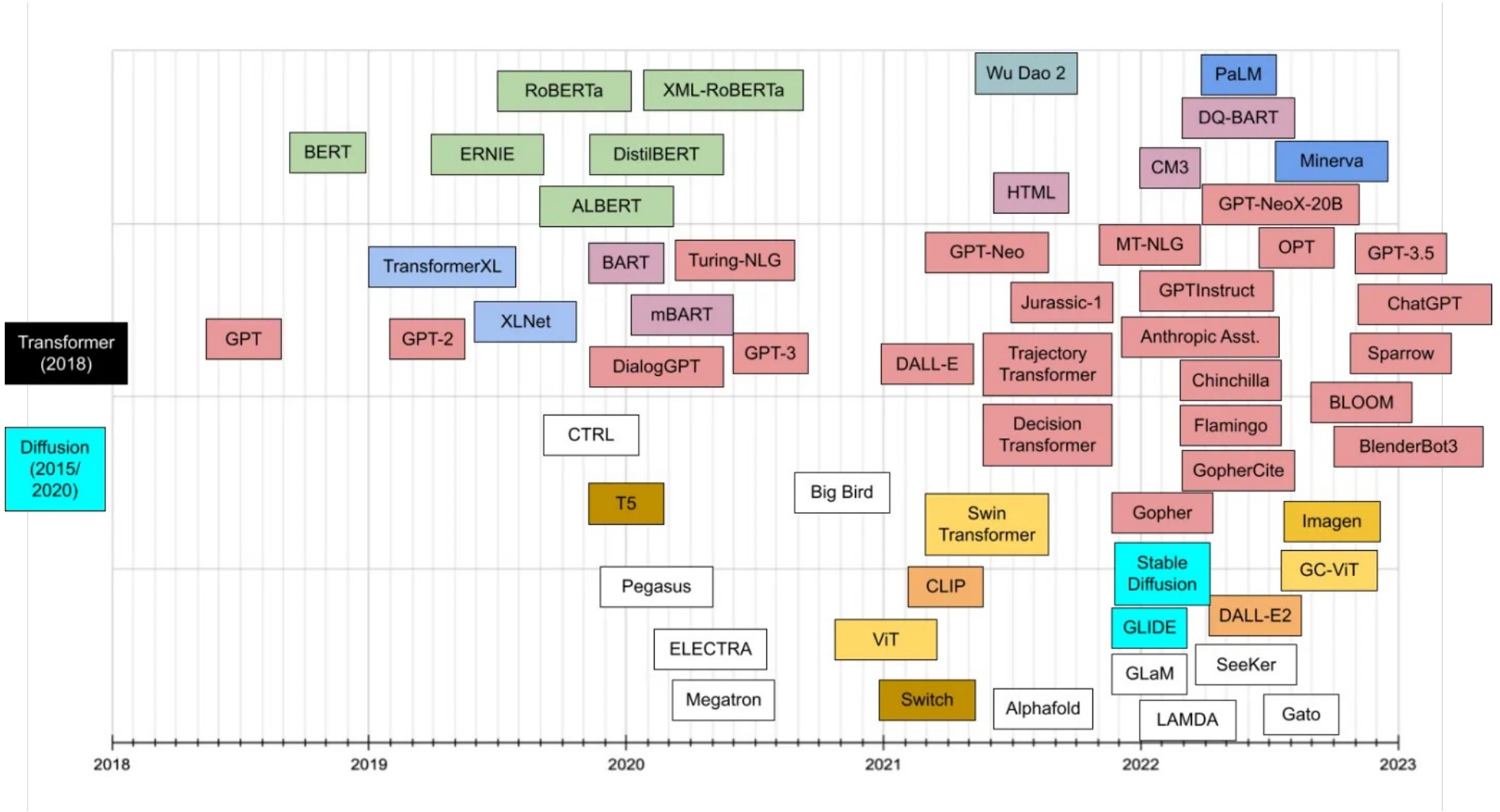
$$PE_t^{(i)} = \begin{cases} \sin(w_i t), & if\ k = 2i \\ \cos(w_i t), & if\ k = 2i+1 \end{cases}$$

What’s next



- MAE: Masked Autoencoders Are Scalable Vision Learners

What’s next


Transformer (2018)
Diffusion (2015/ 2020)
RoBERTa
XML-RoBERTa
BERT
ERNIE
DistilBERT
ALBERT
TransformerXL
BART
Turing-NLG
GPT
GPT-2
XLNet
mBART
DialogGPT
GPT-3
CTRL
T5
Pegasus
ELECTRA
Megatron
Wu Dao 2
PaLM
DQ-BART
Minerva
CM3
HTML
GPT-NeoX-20B
GPT-Neo
MT-NLG
OPT
GPT-3.5
Jurassic-1
GPTInstruct
ChatGPT
DALL-E
Trajectory Transformer
Anthropic Asst.
Sparrow
Chinchilla
BLOOM
Decision Transformer
Flamingo
BlenderBot3
GopherCite
Big Bird
Swin Transformer
Gopher
Imagen
Stable Diffusion
GC-ViT
CLIP
DALL-E2
GLIDE
ViT
SeeKer
GLaM
Switch
Alphafold
LAMDA
Gato
2018
2019
2020
2021
2022
2023