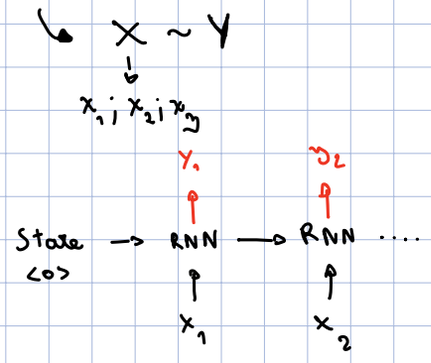
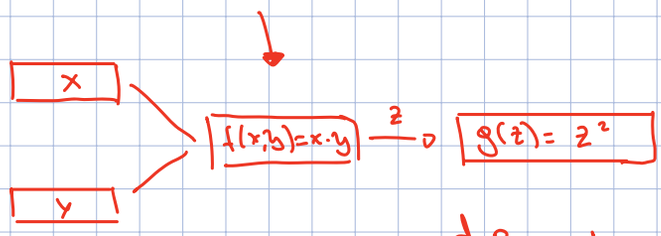
transformer from scratch

RNN



Problem:

1. Slow computation
2. Vanishing or exploding gradients
3. Difficulty in accessing info from long time ago

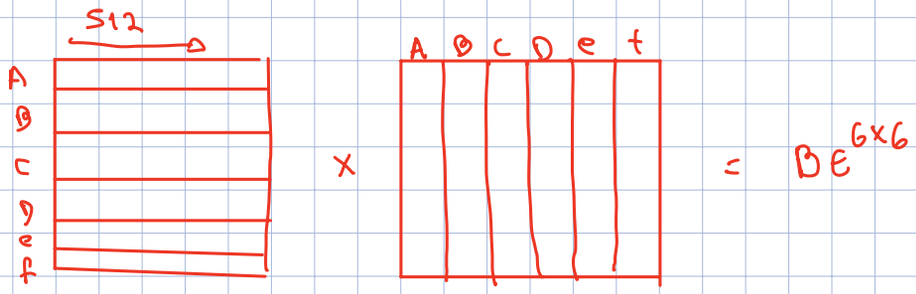


$$\frac{dg}{dx} = \underbrace{\frac{dg}{df}}_{1/2} \cdot \underbrace{\frac{df}{dx}}_{1/2}$$

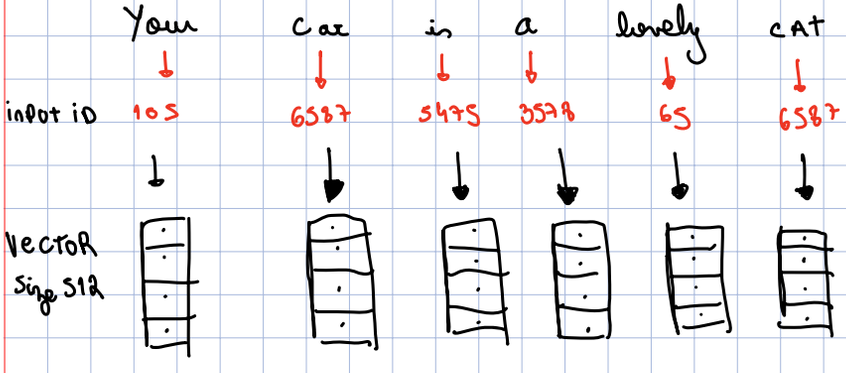
transformer

Notations

$A \in \mathbb{R}^{6 \times 512}$ $\quad (A)^T = A \in \mathbb{R}^{512 \times 6}$



Encoder



positional encoding → represent a position that can learn

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

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

Positional embedding (512)

=

Encoder input

- We only need to compute the positional encodings once and then reuse them for every sentence, no matter if it is training or inference

Multi-head attention

↳ Self attention

↳ $\text{attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

seq: 6 ∧ d_{model}: $d_k = 512$

(6,6)

softmax $\left(\frac{Q \; (6,512) \times K^T \; (512,6)}{\sqrt{512}} \right)$

diagonal

	YOUR	CAT	IS	A	LOVELY	CAT
YOUR	0.268	0.119	0.134	0.148	0.179	0.152
CAT	0.124	0.278	0.201	0.128	0.154	0.115
IS	0.147	0.132	0.262	0.097	0.218	0.145
A	0.210	0.128	0.206	0.212	0.119	0.125
LOVELY	0.146	0.158	0.152	0.143	0.227	0.174
CAT	0.195	0.114	0.203	0.103	0.157	0.229

(6, 6)

X **V** (6, 512) = **Attention** (6, 512)

Each row in this matrix captures not only the meaning (given by the embedding) or the position in the sentence (represented by the positional encodings) but also each word's interaction with other words.

$\sqrt{d_k}$

d_{model} / h

$s \times d_k$

Multi-head Attention

input

- $Q \times W^Q = Q'$ → will see $t \in \mathbb{R}$
- $K \times W^K = K'$
- $V \times W^V = V'$

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

↳ H $\times$ W^O $= MH\text{-}A$

(req, Hxd_k) (hxd_k, d_{model})

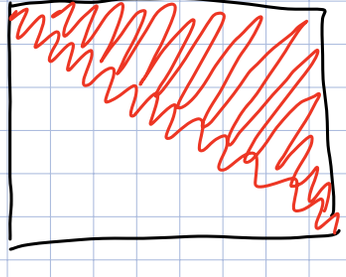
Why Q, K, V → amplify some QK

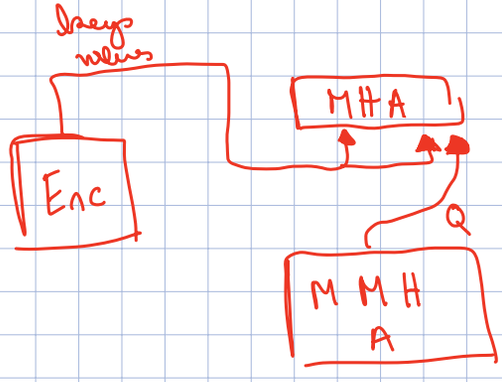
Layer normalization → $\hat{x}_j = \frac{x_j - \mu_j}{\sqrt{\sigma_j^2 + \varepsilon}}$

Decoder

↳ Masked Multi-Head Attention

↳ the model must not be able to see future words

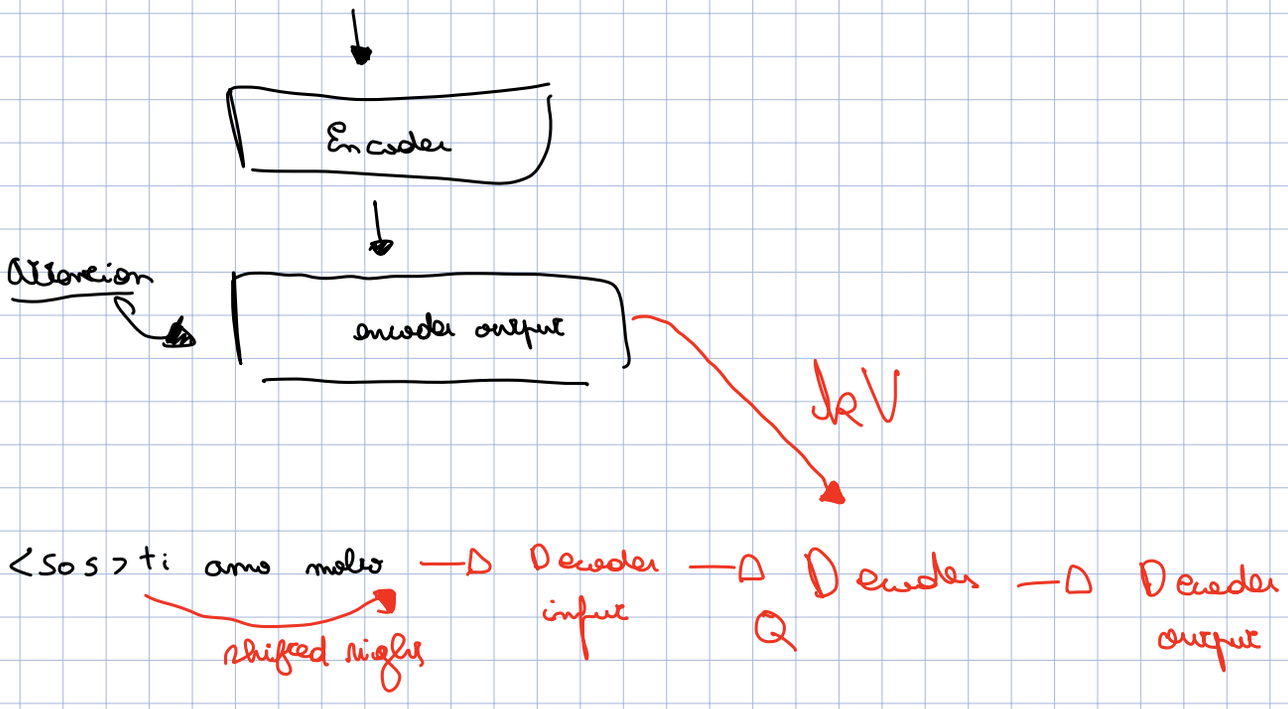




Inference and training

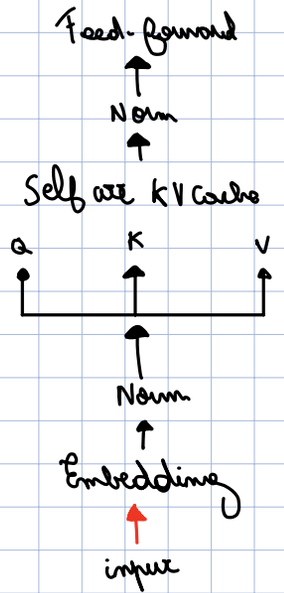
↳ training (translation)

<SOS> I love you very much <EOS>



→ (Seq ; d_{model}) —linear→ (Seq ; vocab_size) → Softmax (seq ; vocab_size)

Llama architecture



GQA → grouped query attention

linear layer

$$G : X \cdot W^{t} + b$$

- The distribution of internal nodes (neurons) of a neural network (weights) changes is referred to as **Internal Covariate Shift**
  - ↳ Normalization $\xrightarrow{\mu,\ \sigma^2}$; $\gamma + \beta$ → learnable parameters

(!!!) AHAAA!!!!

↓ gaus normalization

standard gaussian

$$X \sim N(5;36) \longrightarrow \frac{X-5}{\sqrt{36}} = Z \sim N(0;1)$$

Root Mean Square Normalization

requires < computation

$$\bar{a}_i = \frac{a_i}{RMS(a)} g_i \; ; \quad \text{where } RMS(a) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} a_i^2}$$

Rotary positional encoding

↳ given 2 tokens we create a vector that represents a distance

vanilla transformers they're applied before

absolute positional encoding

$$Att = \left(\frac{QK^T}{\sqrt{d_m}}\right)V$$

relative positional encoding

$$Q\left(K + a_{ij}^{K}\right)^T$$

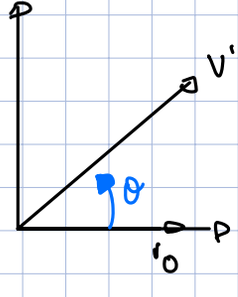
distance

Rotary position embedding

$$\begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix}$$

↓

rotary position encoding



apply a rotation

Long-term decay

↳ the intensity of relationship between two tokens encoded with Rotary Positional embedding will be numerically smaller as the distance between them grows

Multi-query att

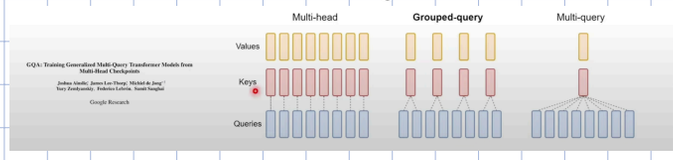
Self-Attention

Attention = $\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

KV-Cache → less computation
↓
only on during inference

Grouped query Attention

- heads for Q
- shared heads for KV



feed-forward

SwiGLU Activation function

FP16 → 2 bytes