Transformers

Monday, 15 April 2024 09:47

Attention based models

"Attention is all you need"

Sequence-to-Sequence mapping tasks
- Machine Translation

Sequence: $X = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_T)$

T: length of sequence

$\bar{x}_t$: Representation for the tth element

Regression Techniques:

Linear model for regression

$$\mathcal{D} = \{\bar{x}_n, t_n\}_{n=1}^{N}$$

Polynomial regression
$$y_n = \sum_{i=1}^{D} w_i \phi_i(\bar{x}_n) + w_0$$

d: Dimensions of $\bar{x}_n$

D: No. of basis functions

$\phi_i(\bar{x}_n)$: monomial of a degree

m: Regression degree.

Interpolation method

$\bar{x}$ y

$$y = \sum_{i=1}^{D} w_i \phi_i(\bar{x}) + w_0$$

$$\phi_i(\bar{x}) = e^{-\frac{\|\bar{x} - \bar{\mu}_i\|}{\sigma}}$$

$$y = \sum_{n=1}^{N} \phi(\bar{x}, \bar{x}_n) \cdot t_n$$

$$\phi(\bar{x}, \bar{x}_n) = e^{-\frac{\|\bar{x} - \bar{x}_n\|}{\sigma}}$$

$\bar{x}$: Query

$\bar{x}_n$: Key

t_n: Value

Attention-based processing of a sequence

$$X = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_T)$$

Query vectors: $\bar{q}_1, \bar{q}_2, \dots, \bar{q}_T$

Key vectors: $\bar{k}_1, \bar{k}_2, \dots, \bar{k}_T$

Value vectors: $\bar{v}_1, \bar{v}_2, \dots, \bar{v}_T$

Scaled dot-product attention (SDPA)

$\bar{q}_t, \bar{k}_t$

$$\alpha_{st} = \frac{\langle \bar{q}_s, \bar{k}_t \rangle}{\sqrt{d}}$$

Attention scores between $\bar{q}_s, \bar{k}_t$
$$a_{st} = \frac{e^{\alpha_{st}}}{\sum_{t'=1}^{T} e^{\alpha_{st'}}}$$

Context vectors
$$\bar{c}_t = \sum_{s=1}^{T} a_{st} \cdot \bar{v}_s$$

Simple-Self Attention (SSA)

$$\bar{z} = W \bar{x}$$
$d \times 1$ $d \times d$ $d \times d$

Self-Attention (SA)

Linear transformation on $\bar{x}_t$ to a lower dimensional space (Subspace projection)

$$\bar{q}_t = W^Q \bar{x}_t$$
$d \times 1$ $d \times d$ $d \times 1$

$$\bar{k}_t = W^K \bar{x}_t$$

$$\bar{v}_t = W^V \bar{x}_t$$

Hence: $\{W^Q, W^K, W^V\}$

$$\bar{c}_t = SDPA(\bar{q}_t, \mathcal{K}, V)$$

Key matrix $\mathcal{K} = [\bar{k}_t]_{t=1}^{T}$

Value matrix $V = [\bar{v}_t]_{t=1}^{T}$

$d \times T$

$\bar{c}_t$: A new representation of $\bar{x}_t$
$d \times 1$ $d \times 1$

Multi-Head Attention (MHA)
- Self-Attention (SA)

j^{th} head: $\{W_j^Q, W_j^K, W_j^V\}$

$$z_{jt} = SDPA(\bar{q}_{jt}, \mathcal{K}_j, V_j)$$

$$\bar{q}_{jt} = W_j^Q \bar{x}_t$$

$$\bar{k}_{jt} = W_j^K \bar{x}_t \qquad \mathcal{K}_j = [\bar{k}_{jt}]_{t=1}^{T}$$

$$\bar{v}_{jt} = W_j^V \bar{x}_t \qquad V_j = [\bar{v}_{jt}]_{t=1}^{T}$$
$d \times 1$ $d \times d$

No. of heads: h

$$d = \frac{d}{h}$$

$$\underset{d \times 1}{\bar{x}_t} \longrightarrow \underset{d \times 1}{\bar{z}_t} = \text{concatenate}(\bar{z}_{1t}, \bar{z}_{2t}, \dots, \bar{z}_{ht})$$

$\bar{z}_{jt}$, $j = 1, 2, \dots, h$
$d \times 1$

h d-dimensional vectors

$$\underset{d \times 1}{\bar{z}_t} = \underset{d \times d}{W^O} \underset{d \times 1}{\bar{z}_t}$$

$$\underset{d \times 1}{\bar{x}_t} \xrightarrow{\text{MHA}} \underset{d \times 1}{\bar{z}_t}$$

Stack

$$\underset{d \times T}{X} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_T)$$

MHA

$$Z = (\bar{z}_1, \bar{z}_2, \dots, \bar{z}_T)$$

Position-wise FFNN

PWFFNN:

$$\underset{d \times 1}{\bar{z}_t} \boxed{W_{FF}} \boxed{\bar{w}_{FF}} \underset{d \times 1}{\bar{z}_t}$$

Linear weight Linear

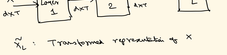
$$Z = (\bar{z}_1, \bar{z}_2, \dots, \bar{z}_T)$$

$$\text{PWFFNN} \downarrow$$

$$\underset{d \times T}{\tilde{X}} = (\underset{d \times 1}{\tilde{z}_1}, \underset{d \times 1}{\tilde{z}_2}, \dots, \underset{d \times 1}{\tilde{z}_T})$$

Transformer Encoder:

L layers



$\tilde{X}_L$: Transformed representation of X

Hyperparameters:

No. of layers: L

No. of heads: h

No. of units in HL of PWFFNN: J

one Layer
$$\begin{cases} \{W_j^Q, W_j^K, W_j^V\}, & j = 1, 2, \dots, h \\ W^O \\ W_{FF}^1, W_{FF}^2 \end{cases}$$

MHA: $d \times d \times h \times 3$

W^O $d \times d$

W_{FF}^1 $d \times J$

W_{FF}^2 $J \times d$

Position Embedding:

$$X = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_T)$$

$$P = (\bar{p}_1, \bar{p}_2, \dots, \bar{p}_T)$$

$$\hat{X} = X + P$$

$$\hat{X} = (\bar{x}_1 + \bar{p}_1, \bar{x}_2 + \bar{p}_2, \dots, \bar{x}_T + \bar{p}_T)$$

$$\hat{X} = (\hat{\bar{x}}_1, \hat{\bar{x}}_2, \dots, \hat{\bar{x}}_T)$$

$j = 0$ $p_{t,0}$ $\longrightarrow p_{t,2j} = \sin\left(\frac{t}{n^{2j/d}}\right)$, for even j

n is a hyperparameter

$n = 10000$

$p_{t,1}$ $p_{t,2j+1} = \cos\left(\frac{t}{n^{2j/d}}\right)$, for odd j

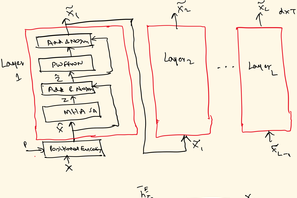
$j = 1$ $p_{t,2}$

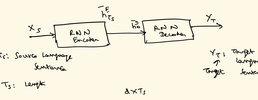
$p_{t,3}$

$\vdots$

$j = 0, 1, 2, \dots, d/2 - 1$

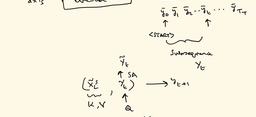
$\underbrace{p_{t,0}, p_{t,1}}_{\bar{p}_t}$





X_S: Source language sentence

T_S: Length



Cross-Attention