INTERSHIP REPORT, ATTENTION GROWING NETWORKS

LÉO BURGUND

1. Nomenclature

1.1. Dimensions.

- b Batch size
- d_e Embeding dimension
- d_s Sequence length
- d_k Query/Keys dimension
- d_v Value dimension
- h Number of heads

1.2. Matrixes.

- Input $X \in \mathbb{R}^{d_s \times d_e}$
- Multi head, for head $i = 1, ..., h$
  - $W_{Q_i} \in \mathbb{R}^{d_e \times \frac{d_k}{h}}, Q_i := X W_{Q_i} \in \mathbb{R}^{d_s \times \frac{d_k}{h}}$
  - $W_{K_i} \in \mathbb{R}^{d_e \times \frac{d_k}{h}}, K_i := X W_{K_i} \in \mathbb{R}^{d_s \times \frac{d_k}{h}}$
  - $S_i := \frac{Q_i K_i^\top}{\sqrt{\frac{d_k}{h}}} \in \mathbb{R}^{d_s \times d_s}$
  - $A_i := \operatorname{softmax}_{\text{row}}(S)$
  - $W_{V_i} \in \mathbb{R}^{d_e \times \frac{d_v}{h}}, V_i := X W_{V_i} \in \mathbb{R}^{d_s \times \frac{d_v}{h}}$
  - $H_i := A_i V_i \in \mathbb{R}^{d_s \times \frac{d_v}{h}}$, $H = [H_1, ..., H_h] \in \mathbb{R}^{d_s \times d_v}$
  - $W_O \in \mathbb{R}^{d_v \times d_e}$
  - $Y := H W_O + X \in \mathbb{R}^{d_s \times d_e}$

Remark 1.2.1: The number of parameters to learn

$$\left(\underbrace{2\left(d_e \frac{d_k}{h}\right)}_{W_{Q_i}, W_{K_i}} + \underbrace{d_e \frac{d_v}{h}}_{W_{V_i}}\right) h + \underbrace{d_v d_e}_{W_O}$$

is the same for any $h \in \mathbb{N}_+^*$.

2. Problem

We study the case where $h = 1$.

References

Email address: leo.burgund@gmail.com