INTERNSHIP REPORT, ATTENTION GROWING NETWORKS

LÉO BURGUND

1. Nomenclature

1.1. Dimensions.

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

1.2. Matrix.

In the case of multi-head attention, for each head $i = 1, ..., h$, we have:

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

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 $h \in \mathbb{N}_+^*$.

Remark 1.2.2: We can easily consider the bias by augmenting the matrices:

$$X' = [X \mid \mathbf{1}] \in \mathbb{R}^{d_s \times (d_e + 1)}$$

$$H' = [H \mid \mathbf{1}] \in \mathbb{R}^{d_s \times (d_v + 1)}$$

And adding a row of parameters to $W_{Q_i}, W_{K_i}, W_{V_i}, W_O$. For example:

$$W'_{Q_i} = \begin{pmatrix} W_{Q_i} \\ (b^Q)^\top \end{pmatrix} \in \mathbb{R}^{(d_e + 1) \times \frac{d_k}{h}}.$$

2. Problem

We study the case where $h = 1$.

We are interested in growing the d_k dimension. We consider the first order approximation, using the functional gradient,

$$\mathcal{L}(f + \partial f(d\theta, d\mathcal{A})) = \mathcal{L}(f) + \left\langle \nabla_f \mathcal{L}(f), \partial f(\partial\theta, \partial\mathcal{A}) \right\rangle + o(\|\partial f(\partial\theta, \partial\mathcal{A})\|).$$

To avoid the softmax's non linearity, we will consider the gradient with respect to the matrix S, just before the softmax.

We then have

$$\mathcal{L}(S + \partial S) = \mathcal{L}(S) + \langle \nabla_S \mathcal{L}(S), \partial S \rangle + o(\|\partial S\|)$$

with

$$\partial S = X\big(W_Q + \partial W_Q\big)(W_K + \partial W_K)^\top X^\top - X W_Q W_K^\top X^\top.$$

We have the following optimization problem:

$$\arg\min_{\partial S} \langle \nabla_S \mathcal{L}(S), \partial S \rangle, \text{such that } \|\partial S\| \leq \gamma$$

$$\arg\min_{\partial W_Q, \partial W_K} \left\| B - X\big(W_Q + \partial W_Q\big)(W_K + \partial W_K)^\top X^\top \right\|^2$$

$$\text{with } B \coloneqq \nabla_S \mathcal{L}(S) + X W_Q W_K^\top X^\top$$

Which is a low rank regression (limited by d_k). B is known.

We can approximate $X\underbrace{\big(W_Q + \partial W_Q\big)}_{d_e \times d_k}\underbrace{(W_K + \partial W_K)^\top}_{d_k \times d_e} X^\top$ with a truncated SVD, taking the first d_k singular values.

To grow S for the next training iteration, we can instead approximate by

References

Email address: leo.burgund@gmail.com