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

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:

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

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

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

Which is a low rank regression limited by d_k (if $d_k < d_e$). 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.

If we want to grow the inner dimension of the attention matrix by p neurons, we can instead approximate by taking the first $d_{k'} \coloneqq d_k + p$ singular values.

Hence, instead of approximating a matrix $\underbrace{\big(W_Q + \partial W_Q\big)}_{d_e \times d_k}\underbrace{(W_K + \partial W_K)^\top}_{d_k \times d_e}$, we approximate

$$\underbrace{Z}_{d_e \times d_e} = \underbrace{\mathring{W}_Q}_{d_e \times (d_{k'})} \underbrace{\mathring{W}_K^\top}_{(d_{k'}) \times d_e} = \left[W_Q + \partial W_Q \mid \underbrace{\widetilde{W}_Q}_{d_e \times p} \right] \left[W_K + \partial W_K \mid \underbrace{\widetilde{W}_K}_{d_e \times p} \right]^\top$$

with $\operatorname{rank}(Z) \leq d_{k'}$ (we make the hypothesis that $d_{k'} < d_e$).

We then have the optimization problem

$$\arg\min_Z \left\| B - XZX^\top \right\|_F^2 \text{ subject to } \operatorname{rank}(Z) \leq d_{k'}.$$

Which is a low rank regression problem, limited by $d_{k'}$.

Let f such that

$$f(Z) = \left\| B - XZX^\top \right\|_F^2,$$

f is convex.

We have

$$\nabla_Z f = -2X^\top (B - XZX^\top) X,$$

so

$$\nabla_Z f = 0 \Leftrightarrow X^\top X Z X^\top X = X^\top B X.$$

In the case where $d_e \leq d_s$ and $\operatorname{rank}(X) = d_e$, then $X^\top X$ is non-singular, and we have the solution

$$Z^\star = (X^\top X)^{-1} X^\top B X (X^\top X)^{-1}.$$

In the general case,

$$Z^\star = X^+ B (X^+)^\top,$$

with $X^+ = \left(X^\top X\right)^{-1} X^\top$ the Moore-Penrose inverse.

If we had $d_{k'} \geq d_e$, we could use the trivial factorization $\mathring{W}_Q = Z^\star, \mathring{W}_K = I_{d_e}$.

As most of the time $d_{k'} < d_e$, we have to approximate the factorization.

According to the Eckart–Young–Mirsky theorem, the best approximation $Z_k^\star$ of $X^+ B (X^+)^\top$ with $\operatorname{rank}(Z_k^\star) = d_{k'}$ is obtained with a truncated SVD.

Indeed, we have

$$Z^\star = U \Sigma V^\top,\ \Sigma = \operatorname{diag}\left(\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{d_e}\right).$$

We keep the $d_{k'}$ largest singular values

$$U_k = \left[u_1, \ldots, u_{d_{k'}}\right],\ V_k = \left[v_1, \ldots, v_{d_{k'}}\right],\ \Sigma_k = \operatorname{diag}\left(\sigma_1, \ldots, \sigma_{d_{k'}}\right).$$

We get

$$Z_k^\star = U_k \Sigma_k V_k^\top, \quad \operatorname{rank}(Z_k^\star)$$

$$\mathring{W}_Q^\star = U_k \Sigma_k^{\frac{1}{2}},\ \mathring{W}_K^\star = V_k \Sigma^{\frac{1}{2}}.$$

Remark 2.1:

$$\min_{\mathring{W}_Q, \mathring{W}_K} \left\| B - X \mathring{W}_Q \mathring{W}_K{}^\top X^\top \right\|_F^2 = \sum_{i > d_{k'}} \sigma_i^2, \quad \text{subject to } \operatorname{rank}\left(\mathring{W}_Q \mathring{W}_K{}^\top\right) \leq d_{k'}$$

References

Email address: leo.burgund@gmail.com