Internship report, Attention growing networks

Léo Burgund

April 25, 2025

1 Nomenclature

1.1 Dimensions

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

1.2 Matrix operations in an attention block

We will first place ourselves in the case where $b = 1$, we study only one instance.

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* 1.1.** 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.** 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:

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

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

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

We have the following optimization problem:

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

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

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

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

We can approximate $X\underbrace{\left(W_Q + \partial W_Q\right)}_{d_e \times d_k}\underbrace{\left(W_K + \partial W_K\right)^\top}_{d_k \times d_e} X^\top$ 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 $\underbrace{\left(W_Q + \partial W_Q\right)}_{d_e \times d_k}\underbrace{\left(W_K + \partial W_K\right)^\top}_{d_k \times d_e}$, 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

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

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

$$\nabla_Z f = 0 \Longleftrightarrow X^\top X Z^\star X^\top X = X^\top BX. \tag{2.1}$$

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

$$Z^\star = \left(X^\top X\right)^{-1} X^\top BX\left(X^\top X\right)^{-1}.$$

In the general case,

$$\boxed{Z^\star = X^+ B\left(X^+\right)^\top,}$$

with X^+ the pseudoinverse (Moore-Penrose).

Proof. Suppose $Z^\star = X^+ B\left(X^+\right)^\top$. Then,

$$\begin{aligned} X^\top X Z^\star X^\top X &= X^\top X X^+ B\left(X^+\right)^\top X^\top X \\ &= X^\top X X^+ B\left(X^\top X X^+\right)^\top. \end{aligned}$$

We have

$$\begin{aligned} X^\top X X^+ &= X^\top \left(X X^+\right)^\top && \text{by definition of the pseudoinverse} \\ &= X^\top \left(X^+\right)^\top X^\top \\ &= \left(X X^+ X\right)^\top \\ &= X^\top && \text{by definition .} \end{aligned}$$

Then

$$X^\top X Z^\star X^\top X = X^\top BX$$

we have verified equation (2.1). □

2.1 Factorization

We now have $Z^\star$, which is equal to $\mathring{W}_Q \mathring{W}_K^\top$, and want to factorize it to find $\mathring{W}_Q$ and $\mathring{W}_K$.

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

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

According to the Eckart–Young–Mirsky theorem, the best approximations $\breve{W}_Q$ and $\breve{W}_K$ to get $\breve{W}_Q \breve{W}_K^\top \approx Z^\star$ with $\operatorname{rank}\left(\breve{W}_Q \breve{W}_K^\top\right) = d_{k'}$ 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

$$Z_{k'}^\star = U_{k'} \Sigma_{k'} V_{k'}^\top, \quad \operatorname{rank}\left(Z_{k'}^\star\right) = d_{k'}$$

$$\breve{W}_Q = U_{k'} \Sigma_{k'}^{\frac{1}{2}},\ \breve{W}_K = V_{k'} \Sigma_{k'}^{\frac{1}{2}}.$$

***Remark* 2.1.**

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

***Remark* 2.2.** For implementation:

Keep the matrices apart, for example for the weight matrix of Q :

$$\breve{W}_Q = W'_Q + \partial W'_Q + W_Q^{\text{new}}$$

with (remind that $d_{k'} = d_k + p$)

$$\underset{d_e \times (d_k+p)}{W'_Q} = \left[w_1 \ \dots \ w_k \,\middle|\, \mathbf{0}_1 \ \dots \ \mathbf{0}_p\right]$$

$$\underset{d_e \times (d_k+p)}{\partial W'_Q} = \left[\partial w_1 \ \dots \ \partial w_k \,\middle|\, \mathbf{0}_1 \ \dots \ \mathbf{0}_p\right]$$

$$\underset{d_e \times (d_k+p)}{W_Q^{\text{new}}} = \left[\mathbf{0}_1 \ \dots \ \mathbf{0}_k \,\middle|\, w_1^{\text{new}} \ \dots \ w_p^{\text{new}}\right]$$

with any vector $w \in \mathbb{R}^{d_e}$, and $\mathbf{0} \in \mathbb{R}^{d_e}$ the 0 vector.

If we wanted to account for the bias, it's the same but include a new last row for each matrix, each vector has one more element.

2.2 Summary

$$\begin{aligned} Z &= X^+\left(\nabla_S \mathcal{L}(S) + XW_QW_K^\top X^\top\right)(X^+)^\top \\ &= X^+\nabla_S \mathcal{L}(S)(X^+)^\top + X^+XW_QW_K^\top X^+X \end{aligned}$$

and

$$U_{k'}\Sigma_{k'}V_{k'}^\top = \text{SVD}_{\text{trunc } k'}(Z)$$

$$\breve{W}_Q = U_{k'}\Sigma_{k'}^{\frac{1}{2}}, \ \breve{W}_K = V_{k'}\Sigma_{k'}^{\frac{1}{2}}.$$

2.3 Notes on Computing

2.3.1 Mini-batch

Note: The "Mini-batch size" can refer either to the machine batch size taken in by the GPU which can optimize computations, or the statistical batch size used to estimate a statistic (this is important as a machine batch may not be of size large enough to get a good estimation of a statistic). In this section, the mini-batch will refer to the statistical batch.

Let b be the mini-batch size, and $i \in \{1, \dots, b\}$.

As Z depends on $\nabla_S \mathcal{L}(S)$, the "quality" of the new weight matrices is dependant on b.

To account for the batch, we identified two possibilities:

(i) For each instance, calculate Z_i, get the empirical mean $\bar{Z}_b$ then do $\text{SVD}\left(\bar{Z}_b\right)$ to find $\breve{W}_Q, \breve{W}_K$.

$$\bar{Z}_b = \mathbb{E}_X[Z_i]$$

$$U_{k'}\Sigma_{k'}V_{k'}^\top = \text{SVD}_{\text{trunc } k'}\left(\bar{Z}_b\right)$$

$$\breve{W}_Q = U_{k'}\Sigma_{k'}^{\frac{1}{2}}, \ \breve{W}_K = V_{k'}\Sigma_{k'}^{\frac{1}{2}}.$$

We do one SVD per mini-batch.

(ii) For each instance, calculate Z_i, do $\mathrm{SVD}(Z_i)$ to get $\breve{W}_{Q,i}, \breve{W}_{K,i}$, then get the empirical means $\overline{\breve{W}}_Q, \overline{\breve{W}}_K$.

$$U_{k',i}\Sigma_{k',i}V_{k',i}^{\top} = \mathrm{SVD}_{\text{trunc } k'}(Z_i)$$

$$\breve{W}_{Q,i} = U_{k',i}\Sigma_{k',i}^{\frac{1}{2}},\ \breve{W}_{K,i} = V_{k',i}\Sigma_{k',i}^{\frac{1}{2}}$$

$$\breve{W}_Q = \overline{\breve{W}}_Q = \mathbb{E}_X\left[\breve{W}_{Q,i}\right],\ \breve{W}_K = \overline{\breve{W}}_K = \mathbb{E}_X\left[\breve{W}_{K,i}\right].$$

Here, we do one SVD for each instance.

Note: This is not counting the SVD we will have to do to find X^+.

2.3.2 Computing Z