Brouillon 3

Léo Burgund

May 19, 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

We make the hypothesis that $d_k < d_e < d_s$.

1.2 Matrix operations in a self-attention block

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 $Y := HW_O + X \in \mathbb{R}^{d_s \times d_e}$

For now, we study the case $h = 1$.

🔥 We omit the $\frac{1}{\sqrt{d_k}}$ scaling for the S matrix, it can cause problem with growing, so for growing we will make it a learnable parameter (and initialize it at $\frac{1}{\sqrt{d_{k_{\text{initial}}}}}$?).

2 Problem

Goal:

$$\min_f \mathcal{L}(f).$$

We will study the variations of the loss made by the variations of S, with other parameters fixed. Hence we will study

$$\arg\min_S \mathcal{L}(S)$$

Let $G := \nabla_S \mathcal{L}(S)$.

We have

$$\mathrm{rk}(G) \leq d_s, \quad \mathrm{rk}(S) \leq d_k$$

We have the first order approximation:

$$\mathcal{L}(S + \gamma\,\mathrm{d}S) = \mathcal{L}(S) + \gamma\langle G, \mathrm{d}S\rangle + o(\|(\mathrm{d}S)\|)$$

We introduce γ, similar to a step size, and we consider the problem

$$\arg\min_{\mathrm{d}S} \langle G, \mathrm{d}S\rangle_F \;\text{ s.t. } \|\mathrm{d}S\| \leq \gamma \wedge \mathrm{rk}(\mathrm{d}S) \leq d_k$$

Let $Z = W_Q W_K^\top$, with $\mathrm{rk}(Z) \leq d_k$, we then have

$$\begin{aligned} S &= X W_Q W_K^\top X^\top \\ &= X Z X^\top \end{aligned}$$

and

🔥 To verify

$$\begin{aligned} \mathrm{d}S &= X\left(W_{Q_{+1}} W_{K_{+1}}\right)X^\top - X W_Q W_K^\top X^\top \\ &= X(Z + \mathrm{d}Z)X^\top - XZX^\top \\ &= X\,\mathrm{d}Z X^\top \end{aligned}$$

Hence

$$\begin{aligned} \langle G, \mathrm{d}S\rangle_F &= \left\langle G, X\,\mathrm{d}ZX^\top\right\rangle_F \\ &= \mathrm{tr}\left(GX\,\mathrm{d}Z^\top X^\top\right) \\ &= \left\langle X^\top G X, \mathrm{d}Z\right\rangle_F \end{aligned}$$

The problem becomes

$$\arg\min_{\mathrm{d}Z}\left\langle X^\top G X, \mathrm{d}Z\right\rangle \;\text{ s.t. } \left\|X\,\mathrm{d}ZX^\top\right\| \leq \gamma \wedge \mathrm{rk}(\mathrm{d}Z) \leq d_k$$

🔥 Problem with the norm constraint:

We can either

- Solve the problem $\min_{X\,\mathrm{d}ZX^\top}\left\langle G, X\,\mathrm{d}ZX^\top\right\rangle$ s.t. $\left\|X\,\mathrm{d}ZX^\top\right\| \leq \gamma \wedge \mathrm{rk}(\mathrm{d}Z) \leq d_k$, expensive but ok.
- Try to relax the norm constraint, but that could cause some space warping? and then $\mathrm{d}Z = -X^\top G X$ could not be the best direction? (Then search gamma with a line search)

-> Test both to see if the second works?

3 Problem with relaxed norm constraint

Let $\mathrm{d}Z^0$ be the best direction for $\mathrm{d}Z$.

We consider we can get $\mathrm{d}Z^0 = -X^\top G X$ from the problem, up to a rank constraint. We will scale with gamma later.

In practice, we could accumulate the $\mathrm{d}Z^0$:

$$\mathrm{d}Z^0 = \mathbb{E}_X\left[-X^\top G X\right]$$

Then do a line search, either

$$\lambda^\star_{\mathrm{FR}} = \mathcal{L}(Z + \lambda\,\mathrm{d}Z^0)$$

$$\lambda^\star_{\mathrm{LR}} = \mathcal{L}\left(\left(Z + \lambda\,\mathrm{d}Z^0\right)_{\mathrm{LR}}\right)$$

Then get the new weight matrices

$$W_{Q_{+1}}, W_{K_{+1}} = \mathrm{SVD}_{\mathrm{LR}}(Z + \lambda^{\star}\, \mathrm{d}Z^{0})$$

4 Full problem