The scaling factor in self-attention

William Chuang

September 25, 2023

1 Introduction

In the context of deep learning neural networks, attention algorithms are designed for determining connections by weights between two elements in a sequential inputs. Self-attention was proposed by Vaswani et. al.[2] in 2017. Suppose a sequential input $X \in \mathbb{R}^N \times \mathbb{R}^{d_x}$ is given. Let $Q := XW_q \in \mathbb{R}^N \times \mathbb{R}^{d_q}$, $K := XW_k \in \mathbb{R}^N \times \mathbb{R}^{d_k}$, and $V := XW_v \in \mathbb{R}^N \times \mathbb{R}^{d_v}$ where W_k, W_v, and W_q are initiated such that (i) $W_q \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_q}$, $W_k \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_k}$, and $W_v \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_v}$, (ii) the variance of the inner product of each row vector in X with each column vector in either W_q, W_k, or W_v to be unit, (iii) each row vector in Q denoted by q_i, and each row vector in K denoted by k_i to be random variables with zero means, i.e. $\mathbb{E}q_i = 0 = \mathbb{E}k_i, \forall i$, and unit variances, i.e $\mathrm{Var}q_i = 1 = \mathrm{Var}k_i$, where $i \in \{1, ..., d_k\}$, and (iv) q_i and k_j are all independent in the following ways: $\mathrm{Cov}\,(q_i, k_i) = 0, \forall i$, $\mathrm{Cov}\,(q_i, q_j) = 0, \forall i \neq j$, and $\mathrm{Cov}\,(k_i, k_j) = 0, \forall i \neq j$. By convention, let $d_k = d_q$. Then the each scalar product of column vectors in Q and K has variance d_k (see Appendix).

Definition 1 (Self-attention)**.** *The self-attention is the following function*

$$\mathit{Attention}\,(Q, K, V) := \left(\mathit{softmax} \left(\frac{\left(QK^T\right)_i}{\sqrt{d_k}} \right)_{i=1}^{N} \right)_{N \times N} \circ V_{N \times d_v}.$$

Remark: The Softmax function is defined using the Boltzmann distribution function: Given a finite sequence $\{z_i\}, j \in \{1, ..., n\}$, then

$$\text{Softmax}(z_i) := \frac{\exp(z_i)}{\sum\limits_j \exp(z_j)}.$$

Since the sum of the above function adds up to one, i.e. $\sum\limits_{i=1}^{n} \text{Softmax}(z_i) = 1$, it is a probability distribution. In physics, this distribution is called Boltzmann distribution. Furthermore, this distribution can be scaled by using a parameter β, and β is the inverse of the temperature of the physical system described by the distribution:

$$\text{Softmax}(\beta z_i) := \frac{\exp(\beta z_i)}{\sum\limits_j \exp(\beta z_j)}.$$

In self-attention, this β is $\frac{1}{\sqrt{d_k}}$, and T is the standard deviation of the scalar product $\sqrt{d_k}$.

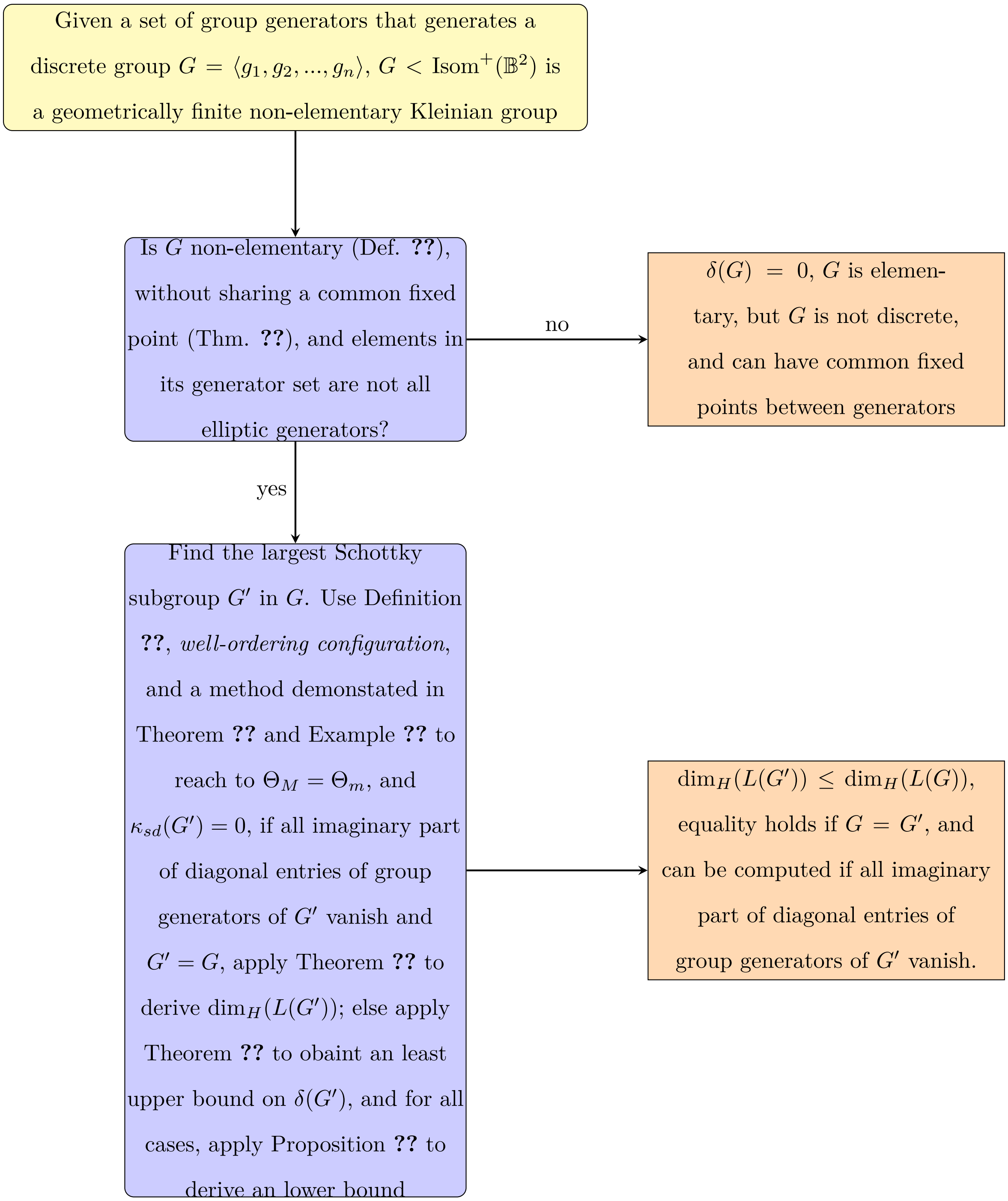
Apply Softmax to column vectors in the matrix $\frac{QK^T}{\sqrt{d_k}}$, meaning each entry of the matrix $\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)_{N \times N}$ is obtained by taking the scalar product on the i-th row of Q to multiply the $j - th$ column of K^T, then divide the scalar product by $\sqrt{d_k}$, and then apply Softmax function on each column of the matrix $\left(\frac{QK^T}{\sqrt{d_k}}\right)_{N \times N}$. Finally, take matrix multiplication between $\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)_{N \times N}$ and $V_{N \times d_v}$.

Hence, in self-attention, each attention, i.e. weight, between each element in the sequential input to the other element in the same sequential input was weighted by the Boltzmann distribution $\text{softmax}\left(\frac{(QK^T)_i}{\sqrt{d_k}}\right)_{i=1}^{N}$.

Furthermore, because of the random initiation of each weight matrices, if this self-attention, i.e. weights assigning process were run in parallel, then the model can learn from more than one different perspectives. In [2], this notion is defined as multi-head attention. Each self-attention function $\text{Attention}(Q, K, V)_i$ is called head_i, where i is from 1 to H. The output of all H heads are concatenated to a matrix denoted by $\text{Concat}(\text{head}_1, ..., \text{head}_H)$, then defined Multihead(Q,K,V) := $\text{Concat}(\text{head}_1, ..., \text{head}_H) \circ W^O$, where $W^O \in \mathbb{R}^{H \cdot d_v} \times \mathbb{R}^{d_{\text{out}}}$.

Let $\epsilon > 0$ be given. Because of the mathematical form of the Boltzmann distribution, if the input is $z_i + \epsilon$, then $\text{Softmax}(\beta z_i + \epsilon) = \text{Softmax}(\beta z_i)$. However, a naive but natural question to ask is what if the Boltzmann distribution, or the self-attention function attend at β^ϵ or $\beta \cdot \epsilon$? It is necessary to ask this question, since we want to be able to justify why and when it makes sense to use the scaling factor $\beta = \frac{1}{\sqrt{d_k}}$ in the self-attention function. Is it sufficient to ask this question? That is the question this paper is going to answer.

2 Algorithm


Given a set of group generators that generates a discrete group $G = \langle g_1, g_2, ..., g_n \rangle$, $G < \mathrm{Isom}^+(\mathbb{B}^2)$ is a geometrically finite non-elementary Kleinian group
Is G non-elementary (Def. **??**), without sharing a common fixed point (Thm. **??**), and elements in its generator set are not all elliptic generators?
no
$\delta(G) = 0$, G is elementary, but G is not discrete, and can have common fixed points between generators
yes
Find the largest Schottky subgroup G' in G. Use Definition **??**, *well-ordering configuration*, and a method demonstated in Theorem **??** and Example **??** to reach to $\Theta_M = \Theta_m$, and $\kappa_{sd}(G') = 0$, if all imaginary part of diagonal entries of group generators of G' vanish and $G' = G$, apply Theorem **??** to derive $\dim_H(L(G'))$; else apply Theorem **??** to obaint an least upper bound on $\delta(G')$, and for all cases, apply Proposition **??** to derive an lower bound
$\dim_H(L(G')) \leq \dim_H(L(G))$, equality holds if $G = G'$, and can be computed if all imaginary part of diagonal entries of group generators of G' vanish.

3 Experiments

4 Discussion

5 Conclusion

6 Appendix: Deriving the scaling factor $\frac{1}{\sqrt{d_k}}$ used in [2]

Let X and Y be random variables.

Firstly, recall the definition of variance[1]:

Definition 2.

$$\sigma^2(X) := \inf_{a \in \mathbb{R}} \mathbb{E}\left[(X-a)^2\right].$$

Hence if $X \notin L^2$, then $\mathrm{Var}[X] = \infty$. If $X \in L^2$, then $\mathbb{E}\left[(X-a)^2\right] = \mathbb{E}(X^2) - 2a\mathbb{E}(X) + a^2$ has a minimum at $a = \mathbb{E}(X)$, and the expression of the variance of X can be rewritten using the minimum:

$$\mathrm{Var}[X] = \mathbb{E}\left[(X-a)^2\right] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

Proposition 1.

$$Var[X] = \mathbb{E}(X^2) + (\mathbb{E}[X])^2 - 2(\mathbb{E}[x])^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2.$$

Proof.

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2 + (\mathbb{E}X)^2 - 2X\mathbb{E}X\right]$$

$$= \mathbb{E}(X^2) + (\mathbb{E}[X])^2 - 2(\mathbb{E}[x])^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2.$$

□

Secondly,

Definition 3. *the covariance of X and Y is:*

$$Cov[X, Y] := \mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right].$$

Proposition 2.

$$Cov[X, Y] = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y.$$

Proof.

$$\mathrm{Cov}[X, Y] = \mathbb{E}\left[XY - X\mathbb{E}Y - Y\mathbb{E}X + \mathbb{E}X\mathbb{E}Y\right]$$

$$= \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y - \mathbb{E}X\mathbb{E}Y + \mathbb{E}X\mathbb{E}Y$$

$$= \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y.$$

$\square$

Then, the expectation value of X^2Y^2 can be written in the following expression:

$$\mathbb{E}X^2Y^2 = \mathrm{Cov}[X^2, Y^2] + \mathbb{E}X^2\mathbb{E}Y^2$$

$$= \mathrm{Cov}[X^2, Y^2] + (\mathrm{Var}X + (\mathbb{E}X)^2)(\mathrm{Var}Y + (\mathbb{E}Y)^2).$$

Proposition 3. *If X and Y are independent and both have zero means, then*

$$Var[XY] = VarX\, VarY.$$

Proof. By applying the above results, the variance of the product of X and Y can be derived:

$$\mathrm{Var}[XY] = \mathbb{E}X^2Y^2 - (\mathbb{E}XY)^2$$

$$= \mathbb{E}X^2Y^2 - (\mathbb{E}Y\mathbb{E}X + \mathrm{Cov}[X, Y])^2$$

$$= \mathrm{Cov}[X^2, Y^2] + (\mathrm{Var}X + (\mathbb{E}X)^2)(\mathrm{Var}Y + (\mathbb{E}Y)^2) - (\mathbb{E}Y\mathbb{E}X + \mathrm{Cov}[X, Y])^2.$$

Since X and Y are independent, then $\mathrm{Cov}[X^2, Y^2] = 0$, and

$$\mathrm{Var}[XY] = (\mathrm{Var}X + (\mathbb{E}X)^2)(\mathrm{Var}Y + (\mathbb{E}Y)^2) - (\mathbb{E}Y\mathbb{E}X + \mathrm{Cov}[X, Y])^2.$$

Since X and Y are independent and both have zero means, i.e. $\mathbb{E}X = 0 = \mathbb{E}Y$, then

$$\mathrm{Var}[XY] = \mathrm{Var}X\mathrm{Var}Y.$$

$\square$

Proposition 4. *Let $X_i, i \in \{1, ..., n\}$ be n independent variables.*

$$Var\left(\sum_i X_i\right) = \sum_i VarX_i.$$

Proof. Then

$$\begin{aligned}
\mathrm{Var}\left(\sum_i X_i\right) &= \mathbb{E}\left(\left(\sum_i X_i\right)^2\right) - \left(\mathbb{E}\sum X_i\right)^2 \\
&= \mathbb{E}\left(\sum_i\sum_j X_iX_j\right) - \left(\sum_i \mathbb{E}X_i\right)^2 \\
&= \sum_i\sum_j \left(\mathbb{E}X_iX_j - \mathbb{E}X_i\mathbb{E}X_j\right) \\
&= \sum_i\sum_j \mathrm{Cov}\left(X_i, X_j\right).
\end{aligned}$$

Since X_i are independent, then $\mathrm{Cov}\left(X_i, X_j\right) = 0, \forall i \neq j$, and

$$\begin{aligned}
\mathrm{Var}\left(\sum_i X_i\right) &= \sum_i \mathrm{Cov}\left(X_i, X_i\right) \\
&= \sum_i \mathrm{Var}X_i.
\end{aligned}$$

$\square$

Then we can derive the assumption in [2] for the variance of QK^T written in the hidden dimension d_k:

Proposition 5. *Let q_i and k_i be random variables with zero means, i.e. $\mathbb{E}q_i = 0 = \mathbb{E}k_i, \forall i$, and unit variances, i.e $Varq_i = 1 = Vark_i$, where $i \in \{1, ..., d_k\}$, and they are all independent in the following ways:*

$$Cov\left(q_i, k_i\right) = 0, \forall i,$$

$$Cov\left(q_i, q_j\right) = 0, \forall i \neq j, \text{ and}$$

$$Cov\left(k_i, k_j\right) = 0, \forall i \neq j.$$

Then

$$Var\left(QK^T\right) = d_k.$$

Proof.

$$\begin{gathered}
\operatorname{Var}\left(QK^T\right) \\
= \operatorname{Var}\left(\sum_{i=1}^{d_k} q_i k_i\right) \\
= \sum_i \operatorname{Var}\left(q_i k_i\right) \\
= \sum_i \operatorname{Var}\left(q_i\right) \operatorname{Var}\left(k_i\right) \\
= \sum_i 1 \cdot 1 \\
= d_k.
\end{gathered}$$

$\square$

The first equality is because in [2], Q and K are row vectors, and QK^T is taking their scalar product. The second equality is because $\operatorname{Cov}\left(q_i, q_j\right) = 0, \forall i \neq j$, and $\operatorname{Cov}\left(k_i, k_j\right) = 0, \forall i \neq j$. Then, the variance of the sum is the sum of variances. The third equality is because of the conditions $\mathbb{E}q_i = 0 = \mathbb{E}k_i, \forall i$ and $\operatorname{Cov}\left(q_i, k_i\right) = 0, \forall i$, hence we have the variance of the product is the product of variances.

The authors in [2] want to scale the result of QK^T using the standard deviation of the scalar product, so QK^T is multiplied by a factor

$$\frac{1}{\sqrt{\mathrm{Var}\left(QK^T\right)}} = \frac{1}{\sqrt{d_k}}.$$

References

[1] Gerald B Folland, *Real analysis: modern techniques and their applications*, vol. 40, John Wiley & Sons, 1999.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, Advances in neural information processing systems **30** (2017).