From Weighted Mean to FlashAttention

liuzhenhai93@outlook.com

June 3, 2025

In this note, self-attention is approached from the view of weighted mean. Then from this insight, flash attention and ring attention are deciphered in a unified way.

1 The Self-Attention

The computation of self attention in it's simplest form can be described as:

$$O = softmax(QK^T)V \tag{1}$$

which can be further factorized to 3 stage:

$$S = QK^T \tag{2}$$
$$P = softmax(S) \tag{3}$$
$$O = PV \tag{4}$$

For the convenience of subsequent discussions, two extra matrices are defines beforehand.

$$W = exp(S) = P * rowsum(S) \tag{5}$$
$$W_{safe} = exp(S - rowmax(S)) = W * exp(-rowmax(S)) \tag{6}$$

2 Weighted Mean

Formally, the weighted mean of a non-empty finite tuple of data $(x_1, x_2, \cdots, x_n)$, with the corresponding non-negative weights $(w_1, w_2, \cdots, w_n)$ is

$$\bar{x} = \frac{\sum_1^n w_k * x_k}{\sum_1^n w_k} \tag{7}$$

2.1 Online Weighted Mean

Weighted mean can be calculated in a streaming way, with the data and weight feed one by one (or block by block). And there are two algorithms (1, 2).

3 Weighted Mean and Self-Attention

Since the sum of each row of P is 1. Each row of O can be considered as a weighted mean of the rows of V, with the corresponding row of P as weights.

Mathematically, If the weights are scaled by a common factor, the weighted mean remains the same. Hence, each row of O can also be considered as a weighted mean of the rows of V, with the corresponding row of W or W_{safe} as weights.

4 Flash Attention: Online Weighted Mean

With numerical stability taken into consideration, each row of O can be regarded as a weighted mean of the rows of V , with the corresponding row of W_{safe} as weights. Resembling the two algorithms for online weighted mean, there are two algorithms (3, 4) for computing O, corresponding to flash attention 1 and flash attention 2 respectively.

5 Ring Attention: Hierarchy Weighted Mean

If we define

$$\bar{x}_{(i,j)} = \frac{\sum_i^j w_k * x_k}{\sum_i^j w_k} \tag{8}$$

$$d_{(i,j)} = \sum_i^j w_k \tag{9}$$

then the global weighted mean can be computed as a weighted mean of two local weighted mean

$$\bar{x} = \frac{d_{(1,m)}}{d_{(1,n)}} \bar{x}_{(1,m)} + \frac{d_{(m+1,n)}}{d_{(1,n)}} \bar{x}_{(m+1,n)} \tag{10}$$

The equation above assumes that the elements are split into two groups. However, it can be extrapolated to any group, the global weighted mean is a weighted mean of local weighted mean:

$$\bar{x} = \frac{d_{(1,m_1)}}{d_{(1,n)}} \bar{x}_{(1,m_1)} + \frac{d_{(m_1+1,m_2+1)}}{d_{(1,n)}} \bar{x}_{(m_1+1,m_2+1)} + \cdots + \frac{d_{(m_k+1,n)}}{d_{(1,n)}} \bar{x}_{(m_k+1,n)} \tag{11}$$

This is the mathematics behind ring attention: K and V are split into blocks; each block attention results in a partial result, and the final result is a weighted mean of the partial results.

Algorithm 1 one pass weighted mean 1

$\bar{x}_0 = 0$
$d_0 = 0$
for $i = 1, 2, \ldots, N$ **do**
  $d_i = d_{i-1} + w_i$
  $\bar{x}_i = \frac{d_{i-1}}{d_i}\bar{x}_{i-1} + \frac{w_i}{d_i}x_i$
end for
$\bar{x} = \bar{x}_N$

Algorithm 2 one pass weighted mean 2

$\bar{x}_0 = 0$
$d_0 = 0$
for $i = 1, 2, \ldots, N$ **do**
  $d_i = d_{i-1} + w_i$
  $\bar{x}_i = \bar{x}_{i-1} + w_i x_i$
end for
$\bar{x} = \frac{\bar{x}_N}{d_N}$

Algorithm 3 flash attention 1

$m_0 = -\infty$
$d'_0 = 0$
$o'_0 =$ row vector of 0
for $i = 1, 2, \ldots, N$ **do**
  $x_i = Q[k,:]K^T[:,i]$
  $m_i = max(m_{i-1}, x_i)$
  $d'_i = d'_{i-1}e^{m_{i-1}-m_i} + e^{x_i - m_i}$
  $o'_i = \frac{d'_{i-1}e^{m_{i-1}-m_i}}{d'_i}o'_{i-1} + \frac{e^{x_i-m_i}}{d'_i}V[i,:]$
end for
$O[k,:] = o'_N$

Algorithm 4 flash attention 2

$m_0 = -\infty$
$d'_0 = 0$
$o'_0 =$ row vector of 0
for $i = 1, 2, \ldots, N$ **do**
  $x_i = Q[k,:]K^T[:,i]$
  $m_i = max(m_{i-1}, x_i)$
  $d'_i = d'_{i-1}e^{m_{i-1}-m_i} + e^{x_i - m_i}$
  $o'_i = o'_{i-1}e^{m_{i-1}-m_i} + e^{x_i-m_i}V[i:]$
end for
$O[k,:] = \frac{o'_N}{d'_N}$