Tensorizing Convolutional layer

Junyao Pu

October 2, 2021

1 Correlation

Given $I \in R^{P_1 \times \cdots \times P_k \times \cdots \times P_K}$, $F \in R^{Q_1 \times \cdots \times Q_j \times \cdots \times Q_J}$ and G = F $\circledast$ I where $\circledast$ is the correlation defined as Eq 1

$$G[i, j, ...] = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} \ldots \sum_{\infty}^{\infty} F[u, v, ...] I[i + u, j + v, ...] \tag{1}$$

The result $G \in R^{(P_1+Q_1-1) \times \cdots \times (P_k+Q_j-1) \times \cdots \times (P_K+Q_J-1)}$ with full correlation
or $G \in R^{(P_1-Q_1+1) \times \cdots \times (P_k-Q_j+1) \times \cdots \times (P_K-Q_J+1)}$ with fully overlapping correlation.

***In the following sections, we will assume our correlation is fully overlapping correlation**

2 Convolutional Layer

In Deep learning, the convolutional layer is usually operating the correlation between filter and image. By given a RGB image $I \in R^{P_1 \times P_2 \times 3}$ and a filter $F \in R^{Q_1 \times Q_2 \times 3}$, the output of the convolutional layer is $G \in R^{(P_1-Q_1+1) \times (P_2-Q_2+1) \times (1)}$ with fully overlapping correlation G = F $\circledast$ I.

2.1 Activation Map

The activation map is the output of the convolutional layer that has single image with multiple filters. Let's have M filters $F^{(1)} \ldots F^{(m)} \ldots F^{(M)}$, every filter $F^{(m)}$ with the single image I will have a output $G^{(m)} \in R^{(P_1-Q_1+1) \times (P_2-Q_2+1) \times (1)}$ with the fully overlapping correlation. The activation map is defined as the collection of $G^{(m)}$ or $G_{(MAP)} \in R^{(P_1-Q_1+1) \times (P_2-Q_2+1) \times (M)}$ where simply stacking $G^{(m)}$ along 3rd axis.

3 Partial (mode-n) correlation

The partial (mode-n) correlation is a tensor operation defined by $\boxed{\circledast}_n$. For given two tensors with N dimensions $F \in R^{Q_1 \times \cdots \times Q_n \times \cdots \times Q_N}$, $I \in R^{P_1 \times \cdots \times P_n \times \cdots \times P_N}$, the partial mode-n correlation $P = F \boxed{\circledast}_n I$ belongs to $\in R^{(P_1 Q_1) \times \cdots \times (P_n - Q_n + 1) \times \cdots \times (P_N Q_N)}$. The subtensor of $P(k_1, \ldots, :, \ldots, k_N) = F(i_1, \ldots, :, \ldots, i_N) \circledast I(j_1, \ldots, :, \ldots, j_N) \in R^{P_n - Q_n + 1}$, where $k = \overline{ij}$ is the multi-indices.

4 Convolutional layer with Tensor Train Decomposition

4.1 Multiple Filters with Single Image

Given a image $I \in R^{I_1 \times I_2 \times C}$ and M filters $F^{(m)} \in R^{J_1 \times J_2 \times C}$ where $m = \{1...M\}$. The activation map $\mathbf{G}_{(MAP)}$ of the convolutional layer belongs to $R^{(I_1-J_1+1) \times (I_2-J_2+1) \times M}$.

Let's tensorize above the convolutional layer with tensor train decomposition. First, we can rewrite I and F as two 4th order tensor $\mathbf{A} \in R^{1 \times I_1 \times I_2 \times C}$ and $\mathbf{B} \in R^{M \times J_1 \times J_2 \times C}$.
Tensor $\mathbf{A}$ can be represented by 4 tensor cores with TT-rank $\{R_1, R_2, R_3, R_4, R_5\}$:
$\alpha_1 \in R^{R_1 \times 1 \times R_2}$
$\alpha_2 \in R^{R_2 \times I_1 \times R_3}$
$\alpha_3 \in R^{R_3 \times I_2 \times R_4}$
$\alpha_4 \in R^{R_4 \times C \times R_5}$
Tensor $\mathbf{B}$ can be also represented by 4 tensor cores with TT-rank $\{Q_1, Q_2, Q_3, Q_4, Q_5\}$:
$\beta_1 \in R^{Q_1 \times M \times Q_2}$
$\beta_2 \in R^{Q_2 \times J_1 \times Q_3}$
$\beta_3 \in R^{Q_3 \times J_2 \times Q_4}$
$\beta_4 \in R^{Q_4 \times C \times Q_5}$

We want to compute the activation map with tensor operation such as partial (mode-n) correlation. Before that, let's see how can we represent every single filter from the tensor cores $\beta_1, \beta_2, \beta_3, \beta_4$ above.

The tensor $\mathbf{B} \in R^{M \times J_1 \times J_2 \times C}$ which is made by M filters $F^{(m)} \in R^{J_1 \times J_2 \times C}$ and represented by tensor cores $\beta_1, \beta_2, \beta_3, \beta_4$. So, the single filter $F^{(m)}$ can be represent by the mth fiber of mode-2 of the first tensor core with the rest of tensor cores, such as
mth fiber of mode-2 of $\beta_1 \in R^{Q_1 \times M \times Q_2}$
$\beta_2 \in R^{Q_2 \times J_1 \times Q_3}$
$\beta_3 \in R^{Q_3 \times J_2 \times Q_4}$
$\beta_4 \in R^{Q_4 \times C \times Q_5}$
or
$\beta_1^{(m)} \in R^{Q_1 \times 1 \times Q_2}$
$\beta_2 \in R^{Q_2 \times J_1 \times Q_3}$
$\beta_3 \in R^{Q_3 \times J_2 \times Q_4}$
$\beta_4 \in R^{Q_4 \times C \times Q_5}$

Now, we can compute the activation map naturally with partial mode-2 correlation base on following algorithm.

for *i in number of filters* **do**
  for *k in number of tensor cores* **do**
    if *k = 1* **then**
      $\gamma_1 = \alpha_1 \boxed{\circledast}_2 \beta_1^{(i)} \in R^{(R_1 Q_1) \times (1-1+1) \times (R_2 Q_2)}$.
    else
      $\gamma_k = \alpha_k \boxed{\circledast}_2 \beta_k$.
    end
  end
  We will end with following tensor cores γ_k
  $\gamma_1 \in R^{R_1 Q_1 \times (1-1+1) \times R_2 Q_2}$
  $\gamma_2 \in R^{R_2 Q_2 \times (I_1 - J_1 + 1) \times R_3 Q_3}$
  $\gamma_3 \in R^{R_3 Q_3 \times (I_2 - J_2 + 1) \times R_4 Q_4}$
  $\gamma_4 \in R^{R_4 Q_4 \times (C - C + 1) \times R_5 Q_5}$
  By reconstructing the activation map slice with tensor cores γ_k
  The activation map slice belongs to $\in R^{1 \times (I_1 - J_1 + 1) \times (I_2 - J_2 + 1) \times 1}$
end
Stacking all slices along 4th axis to form the activation map:
$G_{(MAP)} \in R^{1 \times (I_1 - J_1 + 1) \times (I_2 - J_2 + 1) \times M}$

Algorithm 1: Activation map computation by tensor operation

Given inputs: a tensor $\mathbf{A} \in R^{1 \times I_1 \times I_2 \times C}$ and tensor $\mathbf{B} \in R^{M \times J_1 \times J_2 \times C}$. We have the output: the activation map tensor $\mathbf{G}_{(MAP)} \in R^{1 \times (I_1 - J_1 + 1) \times (I_2 - J_2 + 1) \times M}$ for the single image.

4.2 Multiple Filters with Multiple Images

Given N images $I^{(n)} \in R^{I_1 \times I_2 \times C}$ where $n = \{1...N\}$ and M filters $F^{(m)} \in R^{J_1 \times J_2 \times C}$ where $m = \{1...M\}$. The activation map $\mathbf{G}_{(MAP)}$ of the convolutional layer will belongs to $R^{N \times (I_1 - J_1 + 1) \times (I_2 - J_2 + 1) \times M}$.

We can rewrite I and F as two 4th order tensor $\mathbf{A} \in R^{N \times I_1 \times I_2 \times C}$ and $\mathbf{B} \in R^{M \times J_1 \times J_2 \times C}$.

Tensor **A** can be represented by 4 tensor cores with TT-rank $\{R_1, R_2, R_3, R_4, R_5\}$:
$\alpha_1 \in R^{R_1 \times N \times R_2}$
$\alpha_2 \in R^{R_2 \times I_1 \times R_3}$
$\alpha_3 \in R^{R_3 \times I_2 \times R_4}$
$\alpha_4 \in R^{R_4 \times C \times R_5}$
Tensor **B** can be also represented by 4 tensor cores with TT-rank $\{Q_1, Q_2, Q_3, Q_4, Q_5\}$:
$\beta_1 \in R^{Q_1 \times M \times Q_2}$
$\beta_2 \in R^{Q_2 \times J_1 \times Q_3}$
$\beta_3 \in R^{Q_3 \times J_2 \times Q_4}$
$\beta_4 \in R^{Q_4 \times C \times Q_5}$

Now, we can compute the activation map with partial mode-2 correlation with following algorithm.

for *i in number of filters* **do**
  for *k in number of tensor cores* **do**
    if *k = 1* **then**
      $\gamma_1 = \alpha_1 \boxed{\circledast}_2 \beta_1^{(i)} \in R^{(R_1 Q_1) \times (N-1+1) \times (R_2 Q_2)}$.
    else
      $\gamma_k = \alpha_k \boxed{\circledast}_2 \beta_k$.
    end
  end
  We will end with following tensor cores γ_k
  $\gamma_1 \in R^{R_1 Q_1 \times (N-1+1) \times R_2 Q_2}$
  $\gamma_2 \in R^{R_2 Q_2 \times (I_1 - J_1 + 1) \times R_3 Q_3}$
  $\gamma_3 \in R^{R_3 Q_3 \times (I_2 - J_2 + 1) \times R_4 Q_4}$
  $\gamma_4 \in R^{R_4 Q_4 \times (C - C + 1) \times R_5 Q_5}$
  By reconstruct the 4th order tensor with tensor cores γ_k belongs to
  $\in R^{N \times (I_1 - J_1 + 1) \times (I_2 - J_2 + 1) \times 1}$
end
Stacking all 4th order tensor along 4th axis to form the activation map:
$\mathbf{G}_{(MAP)} \in R^{N \times (I_1 - J_1 + 1) \times (I_2 - J_2 + 1) \times M}$

Algorithm 2: Activation map computation by tensor operation

Given inputs: tensor $\mathbf{A} \in R^{N \times I_1 \times I_2 \times C}$ and tensor $\mathbf{B} \in R^{M \times J_1 \times J_2 \times C}$. We have the output: the activation map tensor $\mathbf{G}_{(MAP)} \in R^{N \times (I_1 - J_1 + 1) \times (I_2 - J_2 + 1) \times M}$ where $\mathbf{G}_{(MAP)}(n, :, :, m)$ the fully overlapping correlation result between n-th image and m-th filter.