Discussion on the Project of Running Stochastic Gradient

Last updated: April 6, 2021

Comments I

- In writing any statement you need some scientific support. Otherwise don't write it.
- Some reports are well written, but some are not

  Try to pay attention on writing your results. In particular, the organization and the workflow of your report
- It is not good to start a report with results

  Some explanation/introduction should be given first
- If you show timing results, probably you should give details of your computing environments

Comments II

- Some reported that the running time per epoch of MNIST is slightly longer than CIFAR10, but some reported the opposite

  Besides the first layer, we use the same network architevture on them.

  Thus we have that the input image of CIFAR10 is larger than MNIST, but the number of data in CIFAR10 is smaller. Later we will analyze their complexities in detail.

  Other issues (e.g., OS, etc.) may also affect the running time

Comments III

- You shouldn't have something called "checkpoint with the highest validation accuracy"

  Note that we don't have a separate validation set
- The issue of seed:
  - The page said that the default seed is 0
  - But in fact the program doesn't set a default seed

Time Complexity I

- MNIST image sizes at different stages of the convolutional layer:

Layers	Beginning $a^m \times b^m \times d^m$	Padding $a^m_{\text{pad}} \times b^m_{\text{pad}} \times d^m_{\text{pad}}$	Convolution $a^m_{\text{conv}} \times b^m_{\text{conv}} \times d^m_{\text{conv}}$
$m = 1$	$28 \times 28 \times 1$	$32 \times 32 \times 1$	$28 \times 28 \times 32$
$m = 2$	$14 \times 14 \times 32$	$16 \times 16 \times 32$	$14 \times 14 \times 64$
$m = 3$	$7 \times 7 \times 64$	$9 \times 9 \times 64$	$7 \times 7 \times 64$

Time Complexity II

- The most time-consuming part in foward pass is multiplication operations:

$$S^{m,i} = W^m \phi(\text{pad}(Z^{m,i})) \tag{1}$$

$$\boldsymbol{s}^{m,i} = W^m \boldsymbol{z}^{m,i} \tag{2}$$

By (1) the complexity in convolutional layers is

$$\mathcal{O}(I \times d^{m+1} h^m h^m d^m a^m_{\text{conv}} b^m_{\text{conv}}) \tag{3}$$

By (2) the complexity in fully-connected layers is

$$\mathcal{O}(I \times n_{m+1} n_m) \tag{4}$$

Time Complexity III

- MNIST: In the forward pass, by (3) the complexity in convolutional layers is

$$
\begin{aligned}
&\mathcal{O}\left(I \times \begin{pmatrix} d^2 & h^1 h^1 d^1 & a^1_{\text{conv}} b^1_{\text{conv}} \\ +d^3 & h^2 h^2 d^2 & a^2_{\text{conv}} b^2_{\text{conv}} \\ +d^4 & h^3 h^3 d^3 & a^3_{\text{conv}} b^3_{\text{conv}} \end{pmatrix}\right) \\
=&\mathcal{O}\left(60,000 \times \begin{pmatrix} 32 \times 5 \times 5 \times 1 \times 28 \times 28 \\ +64 \times 3 \times 3 \times 32 \times 14 \times 14 \\ +64 \times 3 \times 3 \times 64 \times 7 \times 7 \end{pmatrix}\right) \\
=&\mathcal{O}(60,000 \times (627,200 + 3,612,672 + 1,806,336))
\end{aligned} \tag{5}
$$

Time Complexity IV

- In fully-connected layers, the input features is:

$$n_4 = 3 \times 3 \times 64 = 576$$

The output features $n_5 = 10$, so by (4) the complexity is:

$$\begin{aligned} &\mathcal{O}(I \times n_{m+1} n_m) \\ =&\mathcal{O}(60,000 \times 5,760) \end{aligned}$$

- In our CNN_4layers architecture, the complexity in fully-connected layers can be omitted as it is much smaller than convolutional layers.

Time Complexity V

- CIFAR10 image sizes at different stages of the convolutional layer:

Layers	Beginning $a^m \times b^m \times d^m$	Padding $a^m_{\text{pad}} \times b^m_{\text{pad}} \times d^m_{\text{pad}}$	Convolution $a^m_{\text{conv}} \times b^m_{\text{conv}} \times d^m_{\text{conv}}$
$m = 1$	$32 \times 32 \times 3$	$36 \times 36 \times 3$	$32 \times 32 \times 32$
$m = 2$	$16 \times 16 \times 32$	$18 \times 18 \times 32$	$16 \times 16 \times 64$
$m = 3$	$8 \times 8 \times 64$	$10 \times 10 \times 64$	$8 \times 8 \times 64$

Time Complexity VI

- CIFAR10: In the forward pass, by (3) the complexity in convolutional layers is

$$
\begin{aligned}
&\mathcal{O}\left(I \times \begin{pmatrix} d^2 & h^1 h^1 d^1 & a^1_{\text{conv}} b^1_{\text{conv}} \\ +d^3 & h^2 h^2 d^2 & a^2_{\text{conv}} b^2_{\text{conv}} \\ +d^4 & h^3 h^3 d^3 & a^3_{\text{conv}} b^3_{\text{conv}} \end{pmatrix}\right) \\
=&\mathcal{O}\left(50,000 \times \begin{pmatrix} 32 \times 5 \times 5 \times 3 \times 32 \times 32 \\ +64 \times 3 \times 3 \times 32 \times 16 \times 16 \\ +64 \times 3 \times 3 \times 64 \times 8 \times 8 \end{pmatrix}\right) \\
=&\mathcal{O}(50,000 \times (2,457,600 + 4,718,592 + 2,359,296))
\end{aligned}
\tag{6}
$$

Time Complexity VII

- By (5) the complexity of MNIST is

$$\begin{aligned} &\mathcal{O}\left(\begin{array}{r} 37,632,000,000 \\ +216,760,320,000 \\ +108,380,160,000 \end{array}\right) \\ =&\mathcal{O}(362,772,480,000) \end{aligned}$$

Time Complexity VIII

- By (6) the complexity of CIFAR10 is

$$\begin{aligned}&\mathcal{O}\begin{pmatrix}122,880,000,000\\+235,929,600,000\\+117,964,800,000\end{pmatrix}\\=&\mathcal{O}(476,774,400,000)\end{aligned}$$

- Comparing their complexity:

$$\frac{\text{CIFAR10 Complexity}}{\text{MNIST Complexity}}=\frac{\mathcal{O}(476,774,400,000)}{\mathcal{O}(362,772,480,000)}\approx 1.31$$

Time Complexity IX

- Theoretically the running time of CIFAR10 is around 1.3 times slower than MNIST.
- So if your experiments show that CIFAR10 runs faster than MNIST, please check if you have set the option –dim 28 28 1 for the MNIST data.

Grading

- Typically your score is between 70 and 90
- If yours is outside this range, either you are very good or need lots of improvements.
- I may write some comments on your report. A copy of the gradded report will be emailed to you.