以CPU平行運算/GPU與CUDA 加速機器學習之案例研究

資工四 b04902103 蔡昀達

**Experiment and Report:**

1. What’s the relationship between back propagation, inference and training over 4 different models?

🡺 training = back propagation + inference

1. What’s the relationship between batch size, performance and memory over 4 different models? (you can observe the performance by adjust batch\_size from 1, 2, 4, 8,...,32)

🡺 training mode

Resnet50

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Batch size | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| Images/sec | 39.85 | 63.48 | 98.66 | 134.67 | 172.12 | 197.54 | 213.25 |

Alexnet

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Batch size | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| Images/sec | 83.93 | 157.32 | 293.29 | 533.39 | 928.51 | 1443.51 | 1958.64 |

Vgg16

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Batch size | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| Images/sec | 27.11 | 45.31 | 68.68 | 91.62 | 116.88 | 131.25 | 134.71 |

Inception3

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Batch size | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| Images/sec | 28.26 | 47.42 | 68.76 | 98.45 | 122.21 | 132.42 | 139.06 |

🡺 inference mode

Resnet50

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Batch size | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| Images/sec | 125.77 | 239.76 | 356.62 | 472.87 | 565.15 | 637.60 | 684.55 |

Alexnet

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Batch size | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| Images/sec | 437.10 | 711.10 | 1319.31 | 1956.40 | 2687.85 | 4232.19 | 5372.66 |

Vgg16

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Batch size | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| Images/sec | 123.58 | 185.78 | 254.52 | 303.61 | 378.37 | 409.35 | 421.79 |

Inception3

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Batch size | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| Images/sec | 73.68 | 132.79 | 213.59 | 320.62 | 417.12 | 461.94 | 484.82 |

1. Please discuss performance difference between CPU and GPU

* In a words, CPUs are best at handling single, more complex calculations sequentially, while GPUs are better at handling multiple but simpler calculations in parallel.

GPU compute instances will typically cost 2-3x that of CPU compute instances, and 2-3x performance gains in your GPU-based training models.

**以下為我執行實驗的腳本:**

#!/bin/bash

### install

# virtualenv hpc

# source hpc/bin/activate

# pip install tensorflow-gpu==1.8.0

# git clone -b cnn\_tf\_v1.8\_compatible --single-branch https://github.com/tensorflow/benchmarks

# cd benchmarks/scripts/tf\_cnn\_benchmarks

# ./run\_all.sh ## run this script

**exec** 2**<&-**

file**=**"hpc.txt"

**rm** ${file}

cmd**=**"python tf\_cnn\_benchmarks.py --num\_gpus=1 --variable\_update=parameter\_server"

pipe**=**'| tail -n 2 | head -n 1 | cut -d" " -f 3 >> ${file}'

models**=(**resnet50 alexnet vgg16 inception3**)**

# training

**echo** "### training mode ###" **>>** ${file} **&** **cat** ${file} **|** **tail** -n 1

**for** M **in** ${models[@]}**;** **do**

**echo** ${M} **>>** ${file} **&** **cat** ${file} **|** **tail** -n 1

**for** **((** i**=**0**;** i **<** 7**;** i++ **));** **do**

bs**=$((**2**\*\***${i}**))**

**printf** "${bs} " **>>** ${file}

**eval** "${cmd} --batch\_size=${bs} --model=${M} ${pipe}"

**cat** ${file} **|** **tail** -n 1

**done**

**done**

# inference

**echo** "### inference mode ###" **>>** ${file} **&** **cat** ${file} **|** **tail** -n 1

**for** M **in** ${models[@]}**;** **do**

**echo** ${M} **>>** ${file} **&** **cat** ${file} **|** **tail** -n 1

**for** **((** i**=**0**;** i **<** 7**;** i++ **));** **do**

bs**=$((**2**\*\***${i}**))**

**printf** "${bs} " **>>** ${file}

**eval** "${cmd} --batch\_size=${bs} --model=${M} --forward\_only=True ${pipe}"

**cat** ${file} **|** **tail** -n 1

done

done

# training

# python tf\_cnn\_benchmarks.py --num\_gpus=1 --batch\_size=32 --model=resnet50 --variable\_update=parameter\_server

# inference

# python tf\_cnn\_benchmarks.py --num\_gpus=1 --batch\_size=32 --model=resnet50 --variable\_update=parameter\_server --forward\_only=True