GPU选型调研！3090依旧是性价比之王

原创 louwill 机器学习实验室 2022-03-04 15:52

最近算力不够，一些加Transfomer的3D图像分割，现有的显卡显存都带不动，或者是一个实验要跑一周以上时间。所以近期又专门花时间调研了下GPU选型。

现有两张3090显卡，因为是公版，卡外形比较大，dell的服务器只能塞下两张卡。原先设想是做8卡的3090，但咨询了Dell的供应商，说是现在都不太做8卡了，一般只做到4卡，个别型号可以做到6卡。但3090显存只有24G，要体验大batch条件下的3D图像分割计算，这个显存还不够。所以就把目光从消费级显卡投向了专业计算卡。

Nvidia显卡型号看似眼花缭乱，但结合具体使用需求来看，符合的显卡也就那么几款。Nvidia主流的几款GPU型号简介如下表所示。

型号	显存	单精 (FP32)	半精 (FP32)
TITAN Xp	12GB	12.15 T	12.15 T
1080 Ti	11GB	11.34 T	11.34 T
2080Ti	11GB	13.45 T	53.8 T
V100	16/32GB	15.7 T	125 T
3060	12GB	12.74 T	约24T
A4000	16GB	19.17 T	约76T
3080Ti	12GB	34.10 T	约70T
A5000	24GB	27.77T	约117T
3090	24GB	35.58 T	约71T
A40	48GB	37.42 T	149.7 T
A100 SMX4	40/80GB	19.5 T	312 T

从表中可以看到，除了A系列和V系列的专业计算卡之外，其余都是消费级显卡。**其中TITAN Xp、1080Ti和3060都可以作为入门选手使用**，显存不是那么大但作为入门跑跑中小模型还是没问题的。作为进阶的话，2080Ti、A4000、A5000、3080Ti和3090都很合适，**尤其是3090，可以算是性价比之王**，因为其比较大的显存带宽，虽然单精、半精都弱于A40专业计算卡，但到大多数算法上的实测速度都不差于A40。至于A40，可以视作是扩了显存版本的3090，像笔者目前这样对显存有一定要求的，A40就是一个不错的选择。V100是老一代专业计算卡王，而A100则是新一代专业计算卡王，这类级别的显卡，除了贵，没其他缺点了。

Nvidia RTX 3090

关于更具体的GPU参数信息，可参考这个地址：
https://www.techpowerup.com/gpu-specs/

以下是3090和A40在ResNet50和ViT上性能实测。
3090：

```
>>> ResNet50
Namespace(device=0, model='resnet50', precision='float16', train=False)
Iteration 0, 2294.06 images/s in 0.837s.
Iteration 1, 2391.29 images/s in 0.803s.
Iteration 2, 2396.06 images/s in 0.801s.
Iteration 3, 2394.62 images/s in 0.802s.
Iteration 4, 2402.61 images/s in 0.799s.
Namespace(device=0, model='resnet50', precision='float32', train=False)
Iteration 0, 1453.34 images/s in 1.321s.
Iteration 1, 1490.90 images/s in 1.288s.
Iteration 2, 1491.79 images/s in 1.287s.
Iteration 3, 1493.76 images/s in 1.285s.
Iteration 4, 1494.50 images/s in 1.285s.

>>> ViT Transformer
Namespace(device=0, model='vit_base_patch16_224', precision='float16', train=False)
Iteration 0, 1044.44 images/s in 1.838s.
Iteration 1, 1047.37 images/s in 1.833s.
Iteration 2, 1046.37 images/s in 1.835s.
Iteration 3, 1044.68 images/s in 1.838s.
Iteration 4, 1043.91 images/s in 1.839s.
Namespace(device=0, model='vit_base_patch16_224', precision='float32', train=False)
Iteration 0, 596.59 images/s in 3.218s.
Iteration 1, 599.41 images/s in 3.203s.
Iteration 2, 598.86 images/s in 3.206s.
Iteration 3, 597.92 images/s in 3.211s.
Iteration 4, 597.46 images/s in 3.214s.
```

A40：

```
>>> ResNet50
Namespace(device=0, model='resnet50', precision='float16', train=False)
Iteration 0, 1837.41 images/s in 1.045s.
Iteration 1, 1892.04 images/s in 1.015s.
Iteration 2, 1893.29 images/s in 1.014s.
Iteration 3, 1892.99 images/s in 1.014s.
Iteration 4, 1892.73 images/s in 1.014s.
Namespace(device=0, model='resnet50', precision='float32', train=False)
Iteration 0, 1102.49 images/s in 1.742s.
Iteration 1, 1115.45 images/s in 1.721s.
Iteration 2, 1118.49 images/s in 1.717s.
Iteration 3, 1117.32 images/s in 1.718s.
Iteration 4, 1117.80 images/s in 1.718s.

>>> ViT Transformer
Namespace(device=0, model='vit_base_patch16_224', precision='float16', train=False)
Iteration 0, 1155.09 images/s in 1.662s.
Iteration 1, 1153.70 images/s in 1.664s.
Iteration 2, 1152.89 images/s in 1.665s.
Iteration 3, 1150.99 images/s in 1.668s.
Iteration 4, 1150.53 images/s in 1.669s.
Namespace(device=0, model='vit_base_patch16_224', precision='float32', train=False)
Iteration 0, 675.17 images/s in 2.844s.
Iteration 1, 680.69 images/s in 2.821s.
Iteration 2, 679.15 images/s in 2.827s.
Iteration 3, 678.90 images/s in 2.828s.
Iteration 4, 678.21 images/s in 2.831s.
```

可见，虽然A40是专业计算卡内存大，并且单精半精都强于3090，但因其显存带宽的劣势，模型实测性能可能还不如3090。

所以，总结起来就是，买显卡尽量买3090！

参考资料：

https://www.autodl.com/docs/gpu_perf/

往期精彩：

《机器学习 公式推导与代码实现》随书PPT示例
时隔一年！深度学习语义分割理论与代码实践指南.pdf第二版来了！
新书首发 | 《机器学习 公式推导与代码实现》正式出版！
《机器学习公式推导与代码实现》将会配套PPT和视频讲解！
2021，我读了32本书！