A WORKLOADS

Table 2 is the full list of workloads and their batch sizes we used in our evaluation.

Figure 15 is the same peak and average GPU memory usage measurement done in PyTorch, except `overfeat`, which we could not find a working implementation.

Model	Type	Batch Sizes
`alexnet`	Classification	25, 50, 100
`googlenet`	Classification	25, 50, 100
`inception3`	Classification	25, 50, 100
`inception4`	Classification	25, 50, 75
`overfeat`	Classification	25, 50, 100
`resnet50`	Classification	25, 50, 75
`resnet101`	Classification	25, 50, 75
`resnet152`	Classification	25, 50, 75
`vgg11`	Classification	25, 50, 100
`vgg16`	Classification	25, 50, 100
`vgg19`	Classification	25, 50, 100
`vae`	Auto Encoder	64, 128, 256
`superres`	Super Resolution	32, 64, 128
`speech`	NLP	25, 50, 75
`seq2seq`	NLP	Small, Medium, Large

Table 2. DL models, their types, and the batch sizes we used. Note that the entire network must reside in GPU memory when it is running. This restricts the maximum batch size we can use for each network.

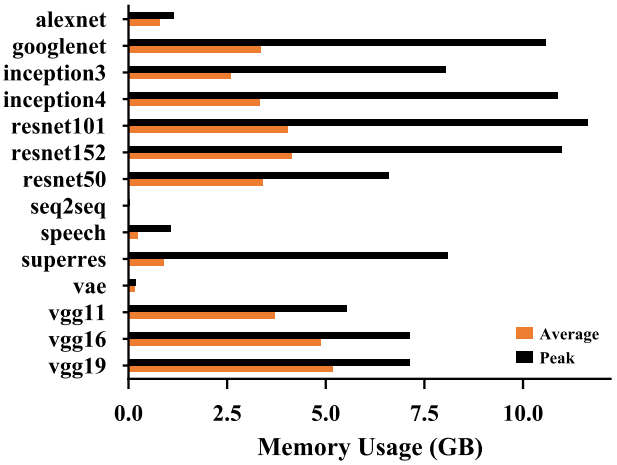

Figure 15. Average and peak GPU memory usage per workload, measured in PyTorch and running on NVIDIA P100 with 16 GB memory. The average and peak usage for vae is 156 MB, 185 MB, which are too small to show in the figure.