modern era

Data Set	Type	Task	Samples	Memory
MNIST (1998)	Image	Classification	60K	12MB
CIFAR-10,100 (2009)	Image	Classification	60K	160MB
ImageNet-1k/21k	Image	Classification	1.4M/14M	0.13TB/1.3TB
Pile	Text	Language modeling	∼186B	0.8TB
LAION-400M/5B	Image-Text	Multi-Modal Learning	400M/5B	11TB/230TB
LLaMA/RedPajama	Text	Language modeling	∼1.2T	5TB