Deep Learning Performance Checklist

| Preparation Store as TFRecords Reduce size of input data Use TensorFlow Datasets |
|--|
| Reading Use tf.data Prefetch data Parallelize CPU processing Parallelize I/O and processing Enable nondeterministic ordering Cache data Turn on experimental optimizations Autotune parameter values |
| Augmentation Use GPU for augmentation |
| Use automatic mixed precision Use larger batch size Use multiples of eight Find the optimal learning rate Use tf.function Overtrain, then generalize Progressive sampling Progressive augmentation Progressive resizing Install an optimized stack for the hardware Optimize number of parallel CPU threads Use better hardware Distribute training Examine industry benchmarks |
| Use an efficient model Quantize the model Prune the model Use fused operations Enable GPU persistence |

From Chapter 6 of *Practical Deep Learning for Cloud, Mobile & Edge* by Anirudh Koul, Siddha Ganju, Meher Kasam https://PracticalDeepLearning.ai/