6.7 DataLoader IMPORTANT PARAMETERS

03 September 2025 01:11 PM Avinash Yadav

The DataLoader class in PyTorch comes with several parameters that allow you to customize how data is loaded, batched, and pre-processed.
Some of the most commonly used and important parameters include:

✓ 1. **dataset (mandatory):**
- The Dataset from which the DataLoader will pull data.
- Must be a subclass of ***torch.utils.data.Dataset*** that implements ***__getitem__*** and ***__len__***.

✓ 2. **batch_size:**
- How many samples per batch to load.
- Default is 1.
- Larger batch sizes can speed up training on GPUs but require more memory.

✓ 3. **shuffle:**
- If True, the DataLoader will shuffle the dataset indices each epoch.
- Helpful to avoid the model becoming too dependent on the order of samples.

✓ 4. **num_workers:**
- The number of worker processes used to load data in parallel.
- Setting ***num_workers > 0*** can speed up data loading by leveraging multiple CPU cores, especially if I/O or preprocessing is a bottleneck.

✓ 5. **pin_memory:**
- If True, the DataLoader will copy tensors into pinned (page-locked) memory before returning them.
- This can improve GPU transfer speed and thus overall training throughput, particularly on CUDA systems.

✓ 6. **drop_last:**
- If True, the DataLoader will drop the last incomplete batch if the total number of samples is not divisible by the batch size.
- Useful when exact batch sizes are required (for example, in some batch normalization scenarios).

✓ 7. **collate_fn:**
- A callable that processes a list of samples into a batch (the default simply stacks tensors).
- Custom ***collate_fn*** can handle variable-length sequences, perform custom batching logic, or handle complex data structures.

✓ 8. **sampler:**
- sampler defines the strategy for drawing samples (e.g., for handling imbalanced classes, or custom sampling strategies).
- ***batch_sampler*** works at the batch level, controlling how batches are formed.
- Typically, you don't need to specify these if you are using ***batch_size*** and shuffle.
- However, they provide lower-level control if you have advanced requirements.