轻墨
2018年08月28日 阅读 73

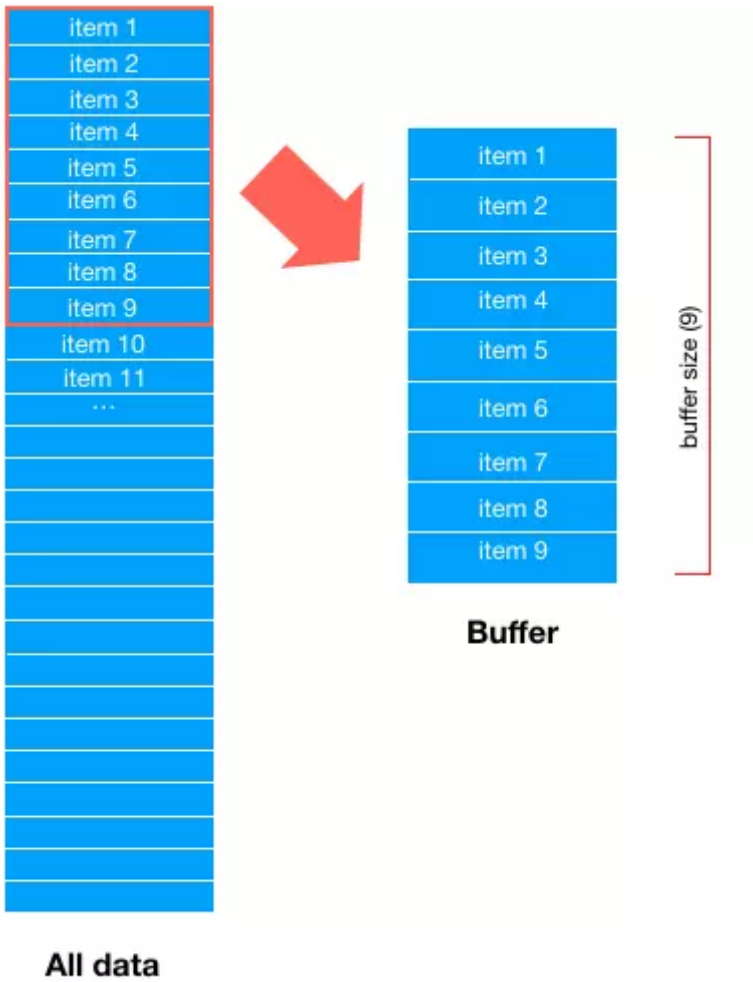
tf.data.Dataset.shuffle(buffer_size)中buffer_size的理解

`tensorflow` 中的数据集类 `Dataset` 有一个 `shuffle` 方法，用来打乱数据集中数据顺序，训练时非常常用。其中 `shuffle` 方法有一个参数 `buffer_size`，非常令人费解，文档的解释如下：

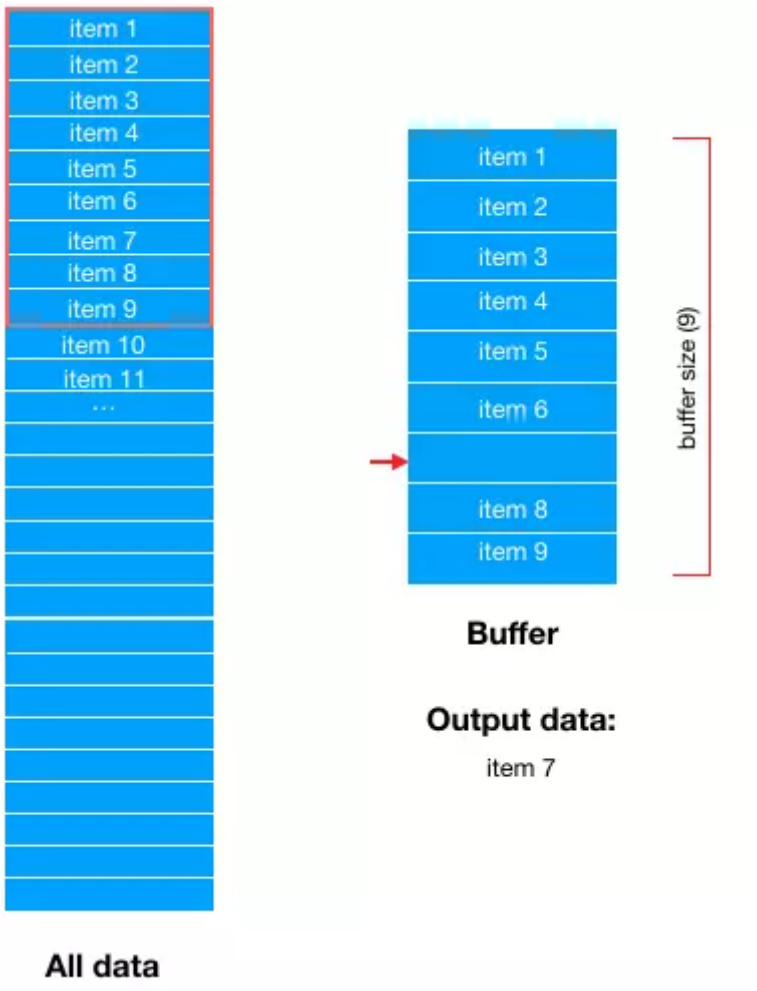
> buffer_size: A tf.int64 scalar tf.Tensor, representing the number of elements from this dataset from which the new dataset will sample.

你看懂了吗？反正我反复看了这说明十几次，仍然不知所指。

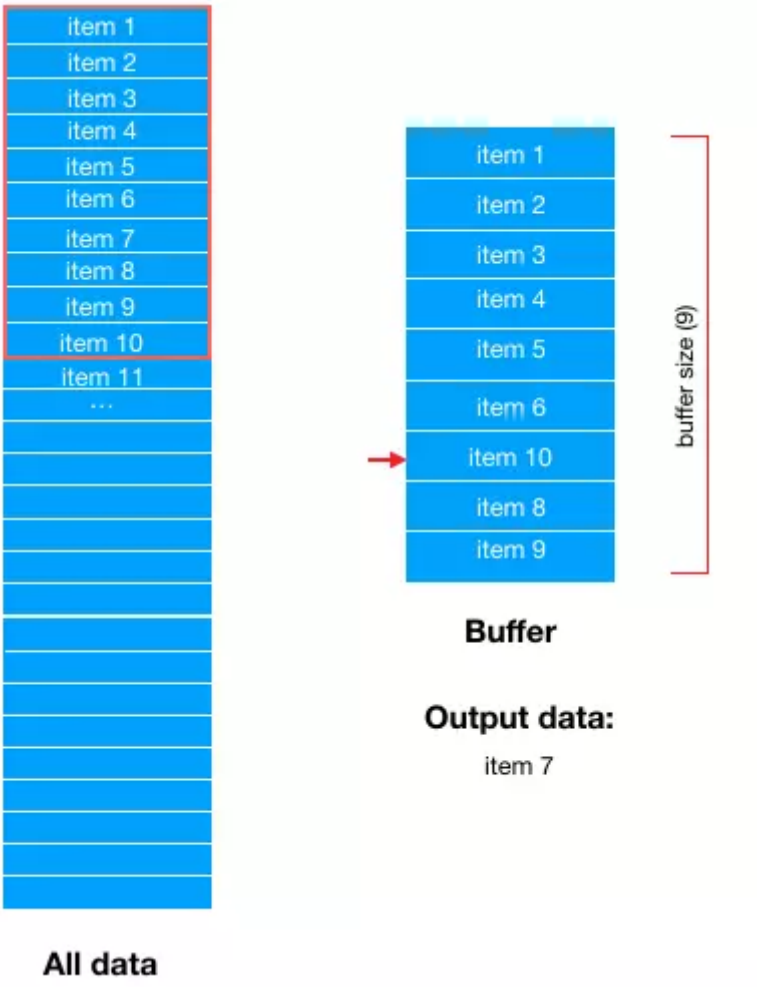
首先，`Dataset` 会取所有数据的前 `buffer_size` 数据项，填充 `buffer`，如下图



然后，从 `buffer` 中随机选择一条数据输出，比如这里随机选中了 `item 7`，那么 `buffer` 中 `item 7` 对应的位置就空出来了



然后，从 `Dataset` 中顺序选择最新的一条数据填充到 `buffer` 中，这里是 `item 10`



然后在从Buffer中随机选择下一条数据输出。

需要说明的是，这里的数据项item，并不只是单单一条真实数据，如果有`batch size`，则一条数据项item包含了`batch size`条真实数据。

shuffle是防止数据过拟合的重要手段，然而不当的buffer size，会导致shuffle无意义，具体可以参考这篇Importance of buffer_size in shuffle()