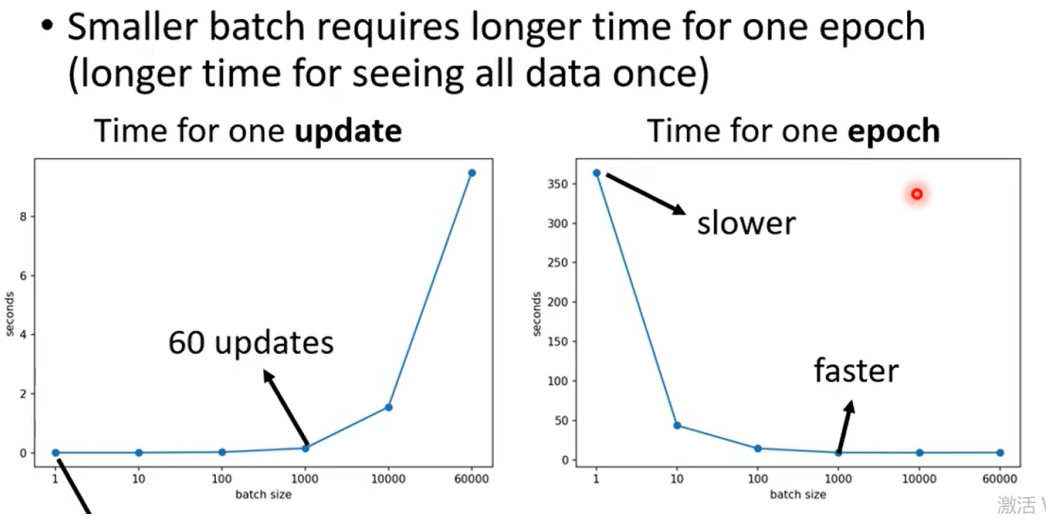
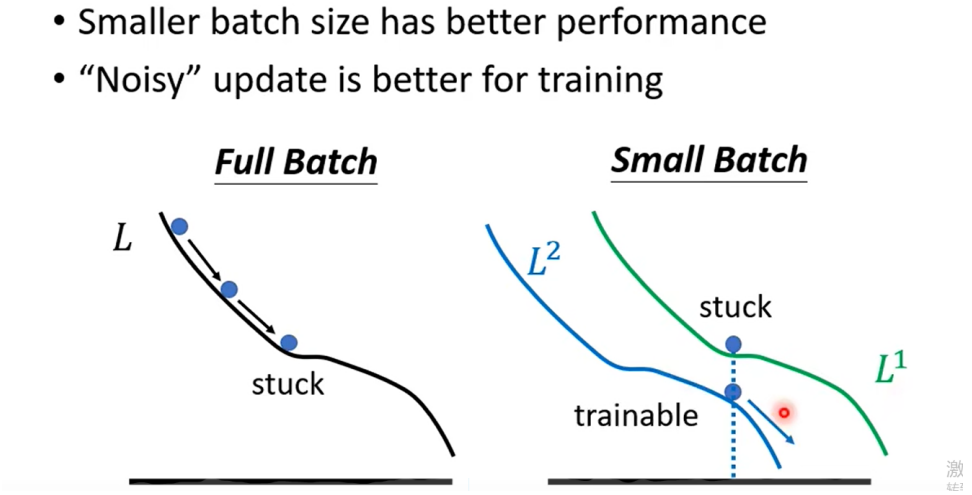
Batch

1.较大batch的优点：用较大的batch能够利用GPU的平行计算能力，减少运算时间。

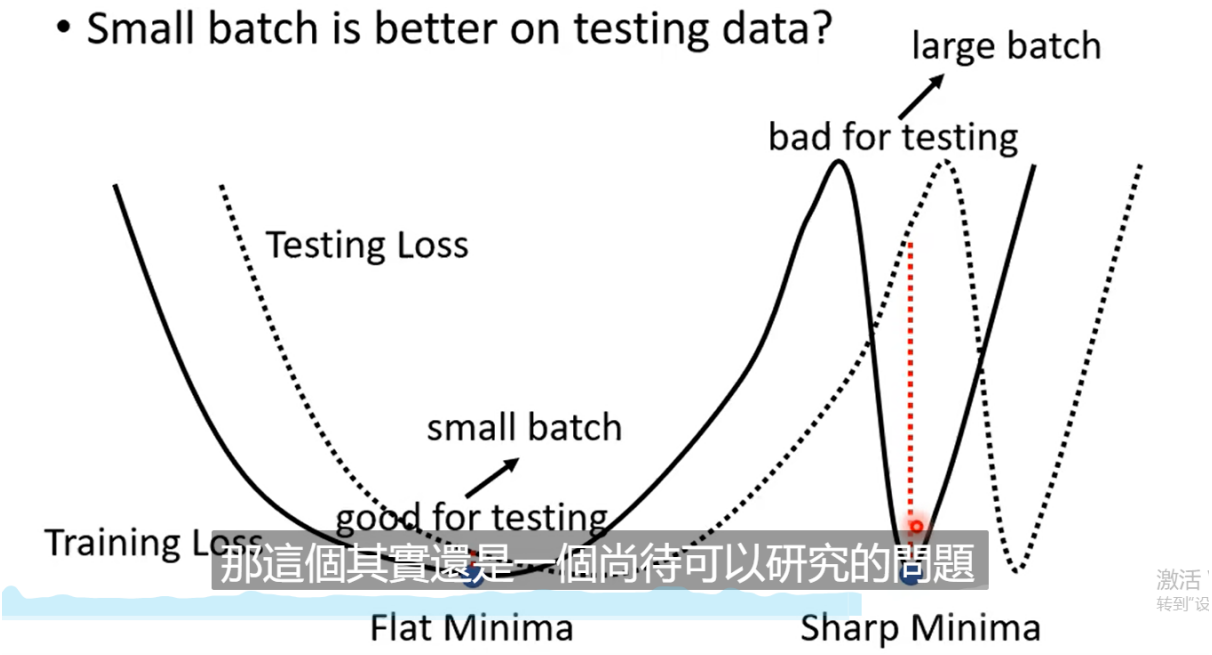


2.较小batch的优点：但是较小的batch可以增加模型的训练效果，其中一个假说如下两图所示：



假说：用FUll Batch时，模型容易陷入saddle point 或者local minimum，而采用更小的batch能够在不同的batch上训练时改变Loss function的值，从而减少模型被stuck的概率。

3.较小batch的优点：同时，在相同的traninig data准确率的情况下，采用小的batch训练出的模型在testing data上有比较好的结果。也就是大的batch会出现overfitting现象。



假说：小的batch更容易让模型跳出loss fuction的sharp minima，停留在flat minima，从而让模型变得在各种不同的testing data上表现稳定。（待研究）

4.batch size 对训练的时间、效果影响表格：

	Small	Large
Speed for one update (no parallel)	Faster	Slower
Speed for one update (with parallel)	Same	Same (not too large)
Time for one epoch	Slower	Faster WIN
Gradient	Noisy	Stable
Optimization	Better WIN	Worse
Generalization	Better WIN	Worse

- Large Batch Optimization for Deep Learning: Training BERT in 76 minutes (https://arxiv.org/abs/1904.00962)
- Extremely Large Minibatch SGD: Training ResNet-50 on ImageNet in 15 Minutes (https://arxiv.org/abs/1711.04325)
- Stochastic Weight Averaging in Parallel: Large-Batch Training That Generalizes Well (https://arxiv.org/abs/2001.02312)
- Large Batch Training of Convolutional Networks (https://arxiv.org/abs/1708.03888)
- Accurate, large minibatch sgd: Training imagenet in 1 hour (https://arxiv.org/abs/1706.02677)

采用较大的batch size增加训练速度的论文，同时采用不同的方法提升trainning的效果。