3 个训练的超参数的理解

在训练代码里面经常见到的 3 个超参数：

```
learning_rate: float = 6e-4
weight_decay: float = 1e-1
grad_clip: float = 1.0
```

1. learning_rate
2. 为何需要 grad_clip

更新权重的 delta 为： learning_rate * grad

但是上面的 delta 还是会很大很小， 所以继续限制 grad

delta为： earning_rate * Norm_clip(grad)

另外：

“The problem is that gradient clipping interacts with adaptive methods like Adam. Since Adam is scale invariant, the absolute scale cannot matter. Rather gradient clipping changes something about the inter-batch variance.”

[Relationship between Learning Rate and Gradient Clipping]
(https://www.reddit.com/r/MachineLearning/comments/eea88q/d_relationship_between_learning_rate_and_gradient/)

Norm_clip:

```
if g >= threshold then
   g <- g*threshold/Norm(g)
```

3. 什么是 weight decay:

Loss = MSE(y_hat, y) + wd * sum(w^2)