1、

$$\frac{\partial L}{\partial y_{pred}} = -\left(\frac{y_{gt}}{y_{pred}} - \frac{1-y_{gt}}{1-y_{pred}}\right)$$

$$\frac{\partial L}{\partial d} = \frac{\partial L}{\partial y_{pred}} \cdot \frac{\partial y_{pred}}{\partial o} \frac{\partial o}{\partial d} = -\left(\frac{y_{gt}}{y_{pred}} - \frac{1-y_{gt}}{1-y_{pred}}\right) \cdot y_{pred}(1-y_{pred}) \cdot 1 = -\left(y_{gt}(1-y_{pred}) - (1-y_{gt})y_{pred}\right)$$

$$= -(y_{gt} - y_{pred}) = y_{pred} - y_{gt}$$

$$o = Vh_t + d$$

$$\frac{\partial L}{\partial V} = \frac{\partial L}{\partial y_{pred}} \cdot \frac{\partial y_{pred}}{\partial o} \frac{\partial o}{\partial V} = (y_{pred} - y_{gt}) \frac{\partial o}{\partial V} = (y_{pred} - y_{gt}) h_t$$

要求 $\frac{\partial L}{\partial h_t}$ 先求 $\frac{\partial L}{\partial h_t}$

$$\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial y_{pred}} \cdot \frac{\partial y_{pred}}{\partial o} \cdot \frac{\partial o}{\partial h_t} = (y_{pred} - y_{gt}) \cdot V \cdot$$

$$\frac{\partial h_t}{\partial h_{t-1}} = \begin{cases} W & b + Wh_{t-1} + Ux_t \geq 0 \\ 0 & b + Wh_{t-1} + Ux_t < 0 \end{cases}$$

$$\frac{\partial h_t}{\partial h_t} = \frac{\partial h_t}{\partial h_{t-(t-t)}} = \begin{cases} W^{t-t} & b + Wh_i + Ux_{i+1} \geq 0 \quad i = t-1, t, \cdots t-1 \\ 0 & b + W h_i + Ux_{i+1} < 0 \quad \exists i \in \{t-1, t, \cdots t-1\} \end{cases}$$

$$\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial h_t} = (y_{pred} - y_{gt}) \cdot V \cdot W^{t-t}$$

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y_{pred}} \cdot \frac{\partial y_{pred}}{\partial o} \cdot \frac{\partial o}{\partial h_t} \cdot \frac{\partial h_t}{\partial W}$$

$$\frac{\partial h_t}{\partial W} = h_{t-1} + \frac{\partial h_t}{\partial h_{t-1}} \cdot \frac{\partial h_{t-1}}{\partial W} = I_n \otimes h_{t-1} + \begin{cases} W \\ 0 \end{cases} \cdot \frac{\partial h_{t-1}}{\partial W}$$

$$\frac{\partial h_{t-1}}{\partial W} = I_n \otimes h_{t-2} + \frac{\partial h_{t-1}}{\partial h_{t-2}} \cdot \frac{\partial h_{t-2}}{\partial W}$$

$$\frac{\partial h_t}{\partial W} = I_n \otimes h_{t-1} + W\left(I_n \otimes h_{t-2} + W\left(I_n \otimes h_{t-3} + W(\quad)\right)\right)$$ 如果没有 ReLU=0的时候

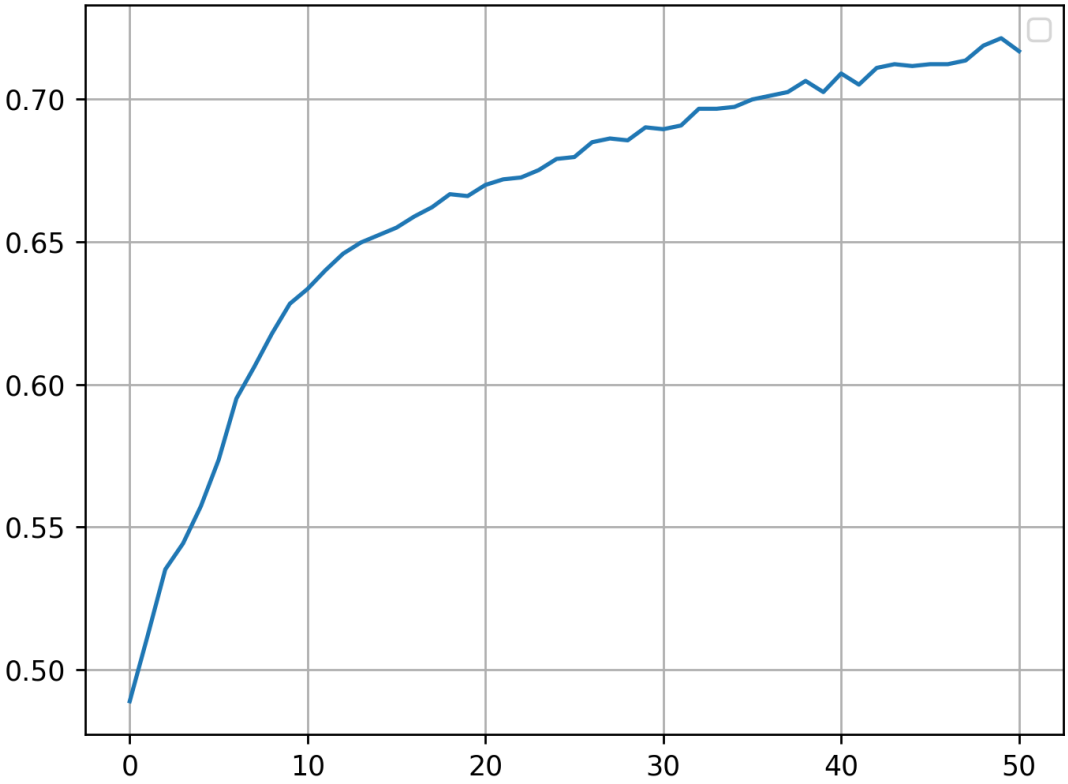
如果有 ReLU=0 的情况，例如 $ReLU(b + Wh_i + Ux_t) = 0$，则 $\frac{\partial h_{i+1}}{\partial W} = 0$，$\frac{\partial h_{i+2}}{\partial W} = I_n \otimes h_{i+1}$

代入 $\frac{\partial h_t}{\partial W}$ 即可

3 waimai_10k Mini分类

3.1词向量平均线性分类器

判定准则选择为：超过80个epoch之后，再看验证集上的准确率和测试集上的准确率
验证集上的准确率：



测试集上的准确率：

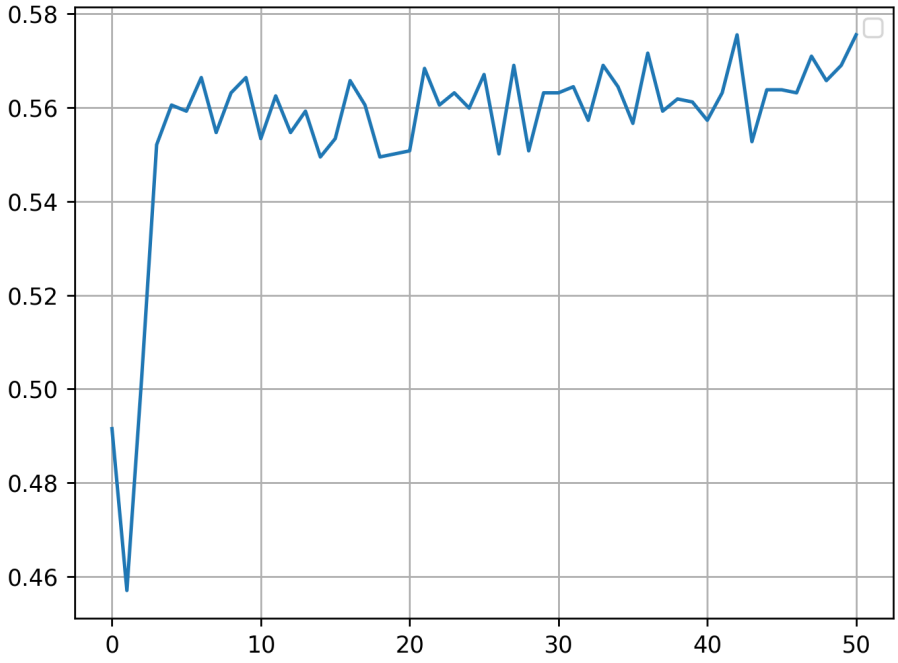
```
Avg Accuracy(): 0.7168
on testing set:
Avg Accuracy(): 0.6901
```

3.2RNN

使用LSTM的

判定准则选择为：超过50个epoch之后，再看验证集上的准确率和测试集上的准确率

验证集上的准确率：

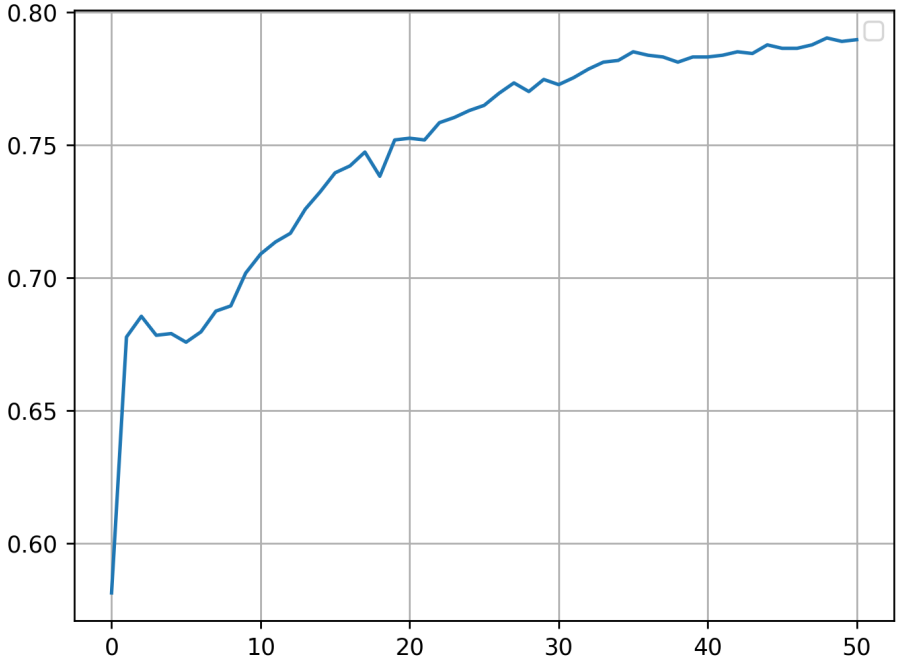


测试集上的准确率：

```
Avg Accuracy(): 0.5755
on testing set:
Avg Accuracy(): 0.6406
```

使用普通RNN的

判定准则：同LSTM



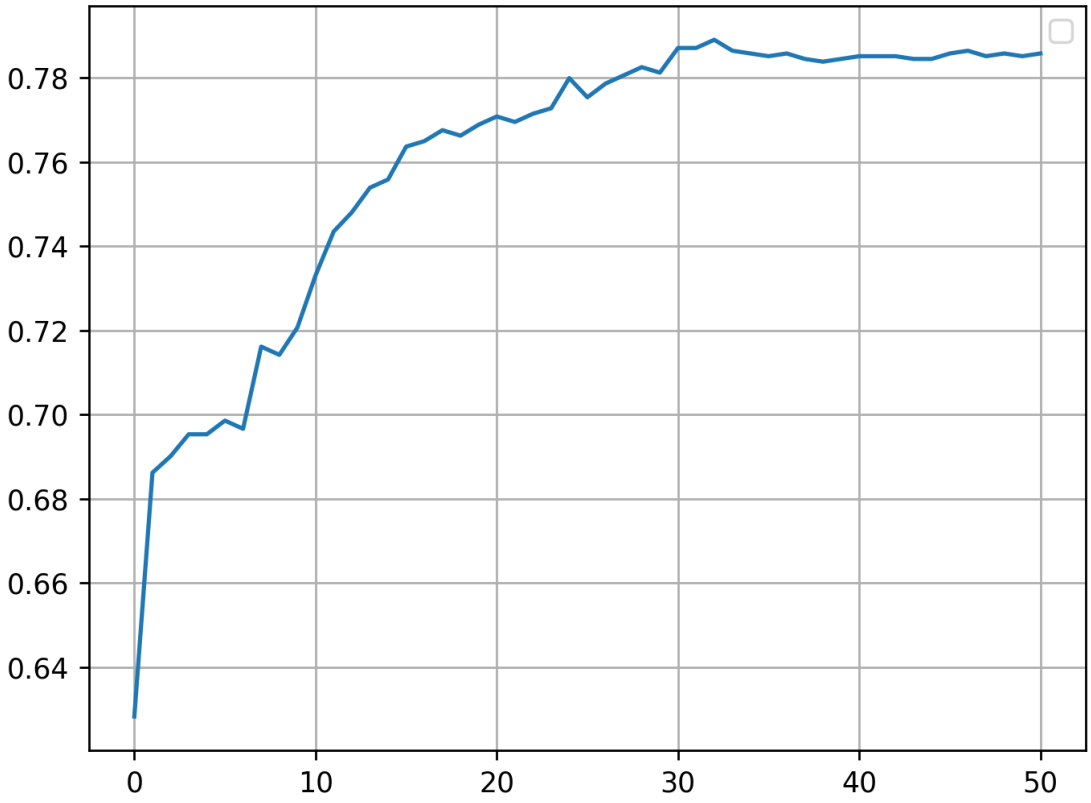
测试集上的准确率：

```
Avg Accuracy(): 0.7897
on testing set:
Avg Accuracy(): 0.7839
```

其他尝试：

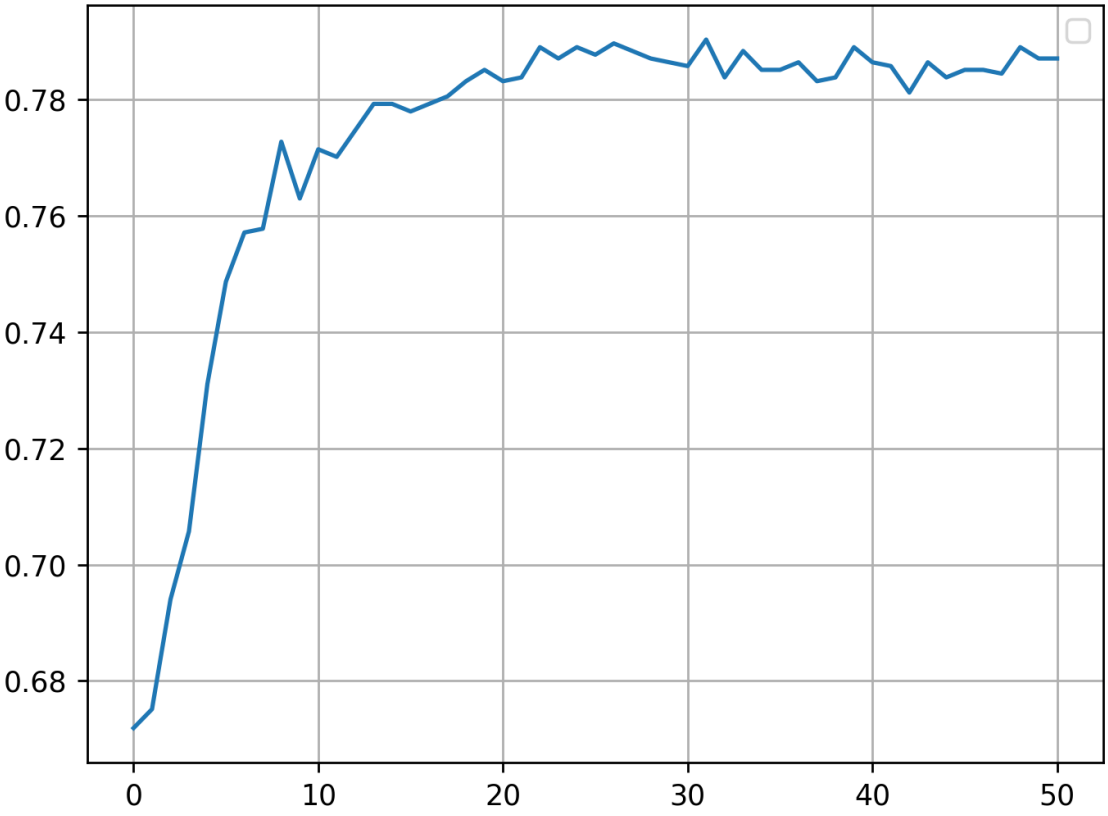
使用多层、双向 RNN，至少报告 1 层双向， 2 层单向、 2 层双向三种情况。

一层双向：



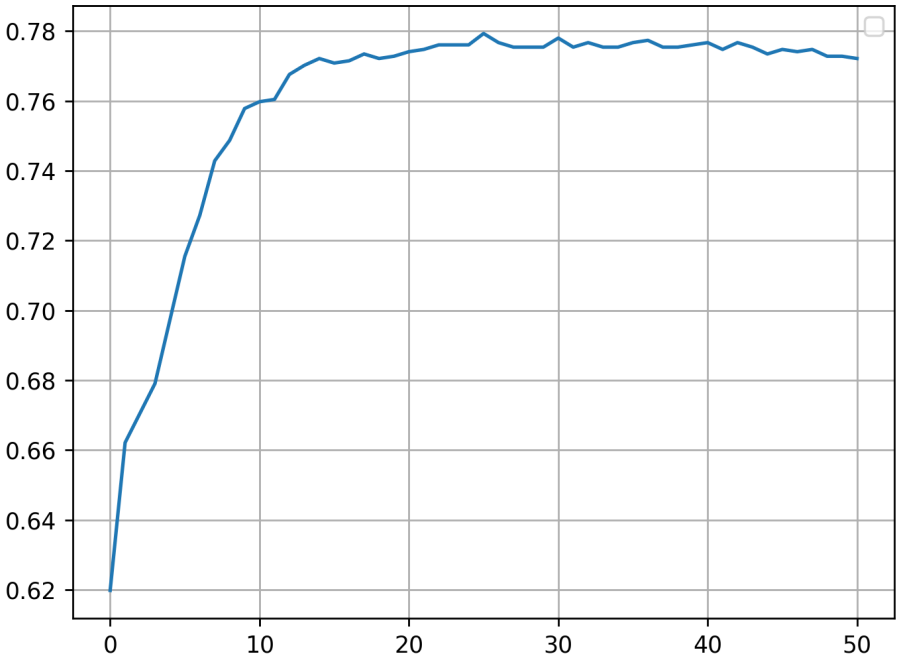

```
Avg Accuracy(): 0.7858
on testing set:
Avg Accuracy(): 0.7826
```

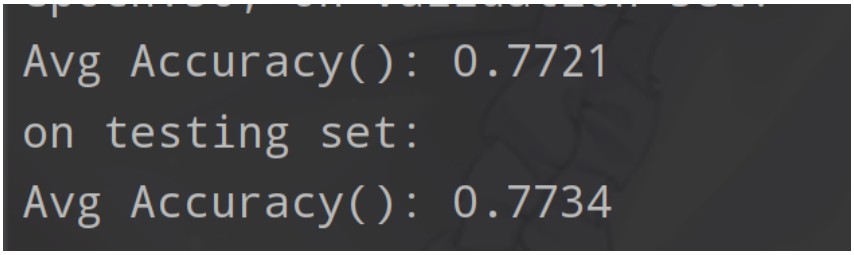
双向RNN，两层




```
Avg Accuracy(): 0.7871
on testing set:
Avg Accuracy(): 0.7734
```

单向RNN，两层





三种情况下的准确率差别不大，主要是数据集太简单了，所以差距不大。我认为是这是我预先padding了的原因。预先padding之后，很多信息就被掩盖和稀释了，所以RNN的效果就不明显了。