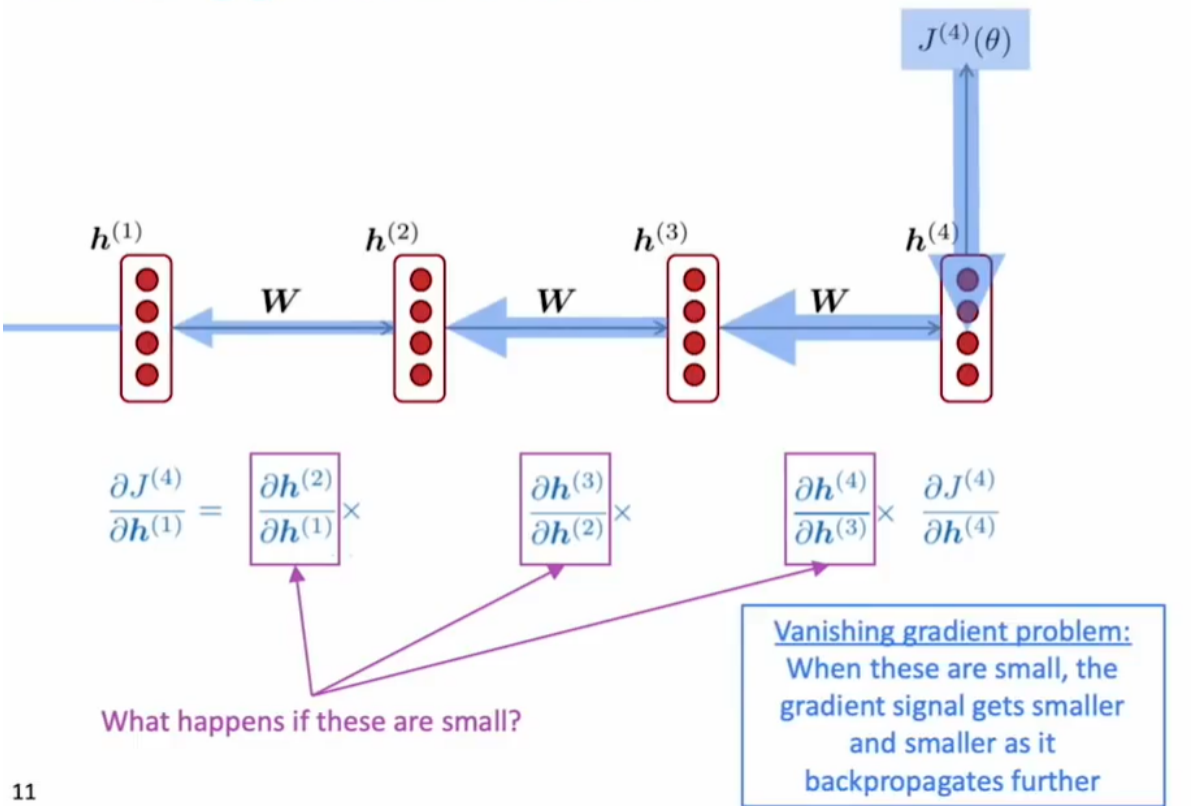
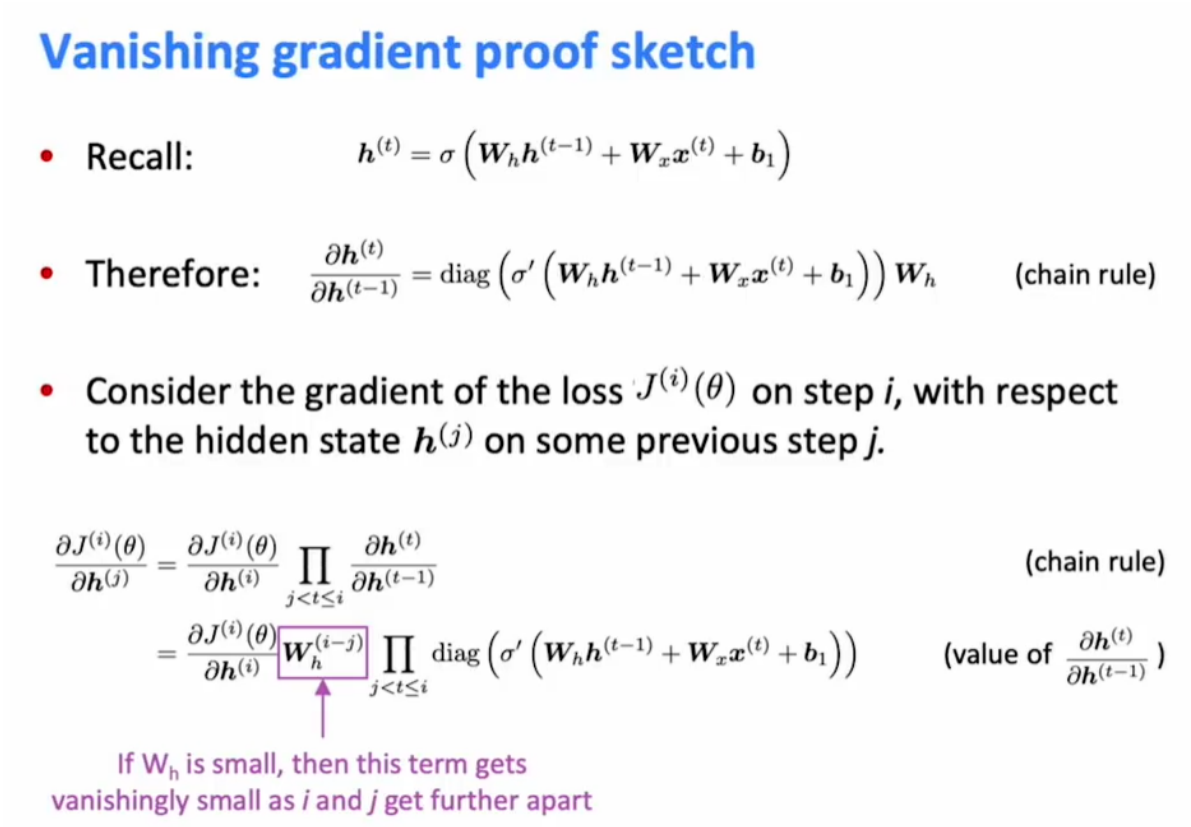
Vanishing gradient problem

链式法则可能导致梯度消失。



数学细节推导，造成梯度消失的是W_h的幂次:



当W_h的最大特征值大于1，梯度就会以指数速度消失;大于1，梯度以指数速度爆炸:

Vanishing gradient proof sketch

- Consider matrix L2 norms:

$$\left\|\frac{\partial J^{(i)}(\theta)}{\partial \boldsymbol{h}^{(j)}}\right\| \leq \left\|\frac{\partial J^{(i)}(\theta)}{\partial \boldsymbol{h}^{(i)}}\right\| \|\boldsymbol{W}_h\|^{(i-j)} \prod_{j<t\leq i} \left\|\operatorname{diag}\left(\sigma'\left(\boldsymbol{W}_h \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_x \boldsymbol{x}^{(t)} + \boldsymbol{b}_1\right)\right)\right\|$$

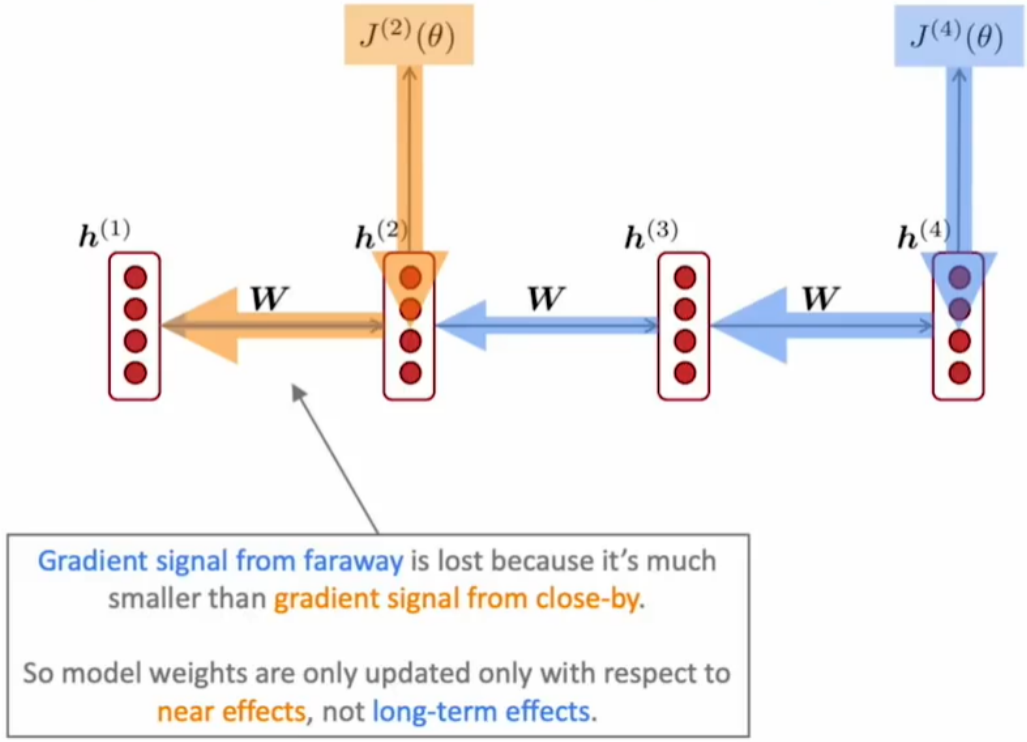
- Pascanu et al showed that that if the largest eigenvalue of W_h is less than 1, then the gradient $\left\|\frac{\partial J^{(i)}(\theta)}{\partial \boldsymbol{h}^{(j)}}\right\|$ will shrink exponentially
  - Here the bound is 1 because we have sigmoid nonlinearity
- There's a similar proof relating a largest eigenvalue >1 to exploding gradients

13 **Source**: "On the difficulty of training recurrent neural networks", Pascanu et al, 2013. http://proceedings.mlr.press/v28/pascanu13.pdf

梯度消失有什么问题?

会导致$h^{(i)}$受$J^{(i)}$的影响远大于J^{i+j}，即$h^{(i)}$对long-term的影响在梯度下降时考虑很少。

Why is vanishing gradient a problem?



14

几个梯度消失的例子

Effect of vanishing gradient on RNN-LM

- **LM task:** *When she tried to print her tickets, she found that the printer was out of toner. She went to the stationery store to buy more toner. It was very overpriced. After installing the toner into the printer, she finally printed her _____*
- To learn from this training example, the RNN-LM needs to model the dependency between *"tickets"* on the 7th step and the target word *"tickets"* at the end.
- But if gradient is small, the model can't learn this dependency
  - So the model is unable to predict similar long-distance dependencies at test time

Effect of vanishing gradient on RNN-LM

- **LM task:** *The writer of the books ___* → *is* / *are*
- **Correct answer**: *The writer of the books is planning a sequel*
- **Syntactic recency:** *The writer of the books is* (correct)
- **Sequential recency:** *The writer of the books are* (incorrect)
- Due to vanishing gradient, RNN-LMs are better at learning from sequential recency than syntactic recency, so they make this type of error more often than we'd like [Linzen et al 2016]

梯度爆炸有什么问题?

每次梯度下降幅度过大，甚至可能出现NaN和Inf。

解决方法

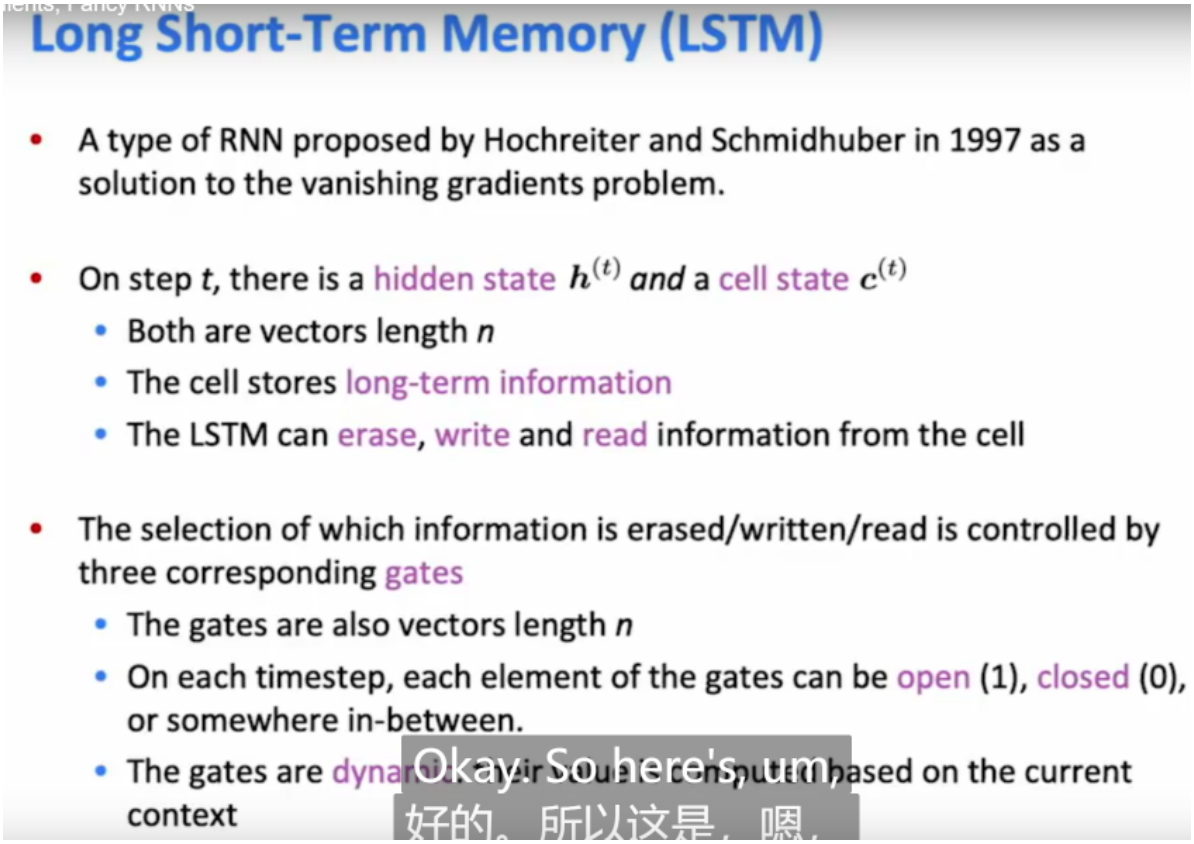
- gradient clipping

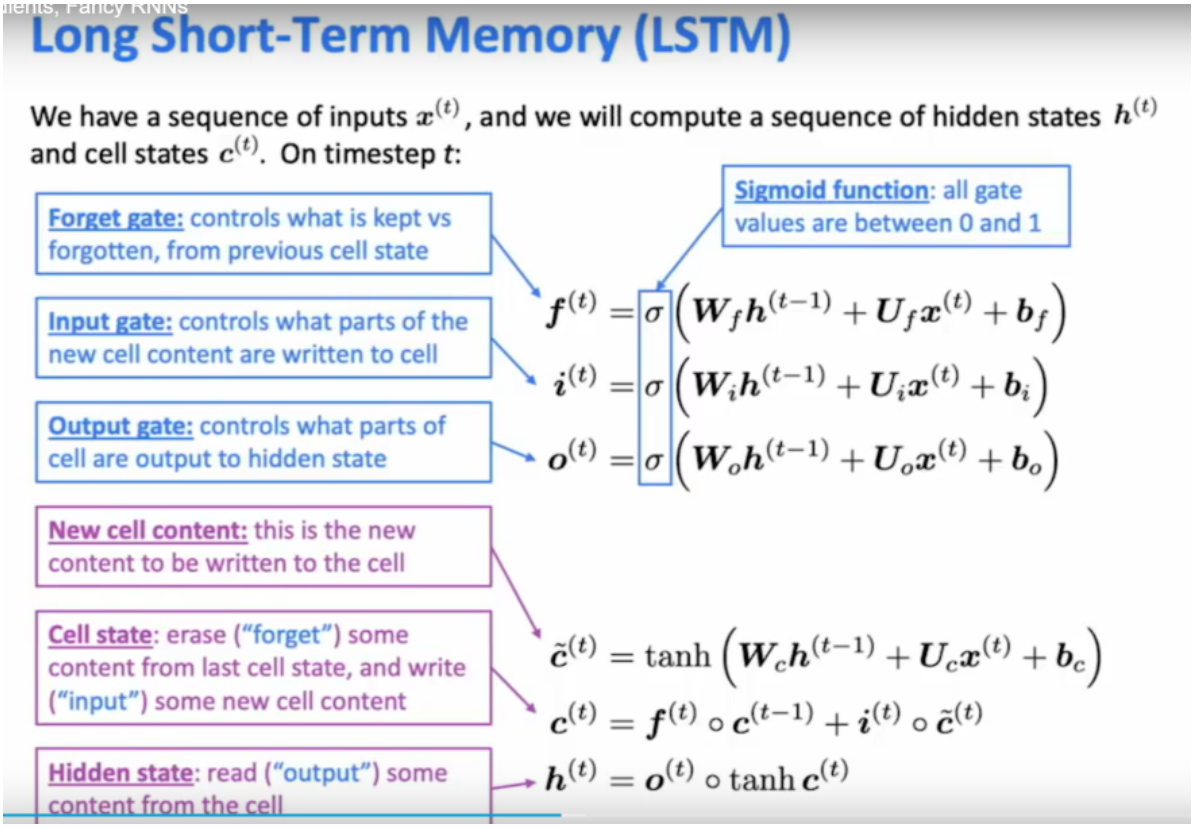
Algorithm 1 Pseudo-code for norm clipping

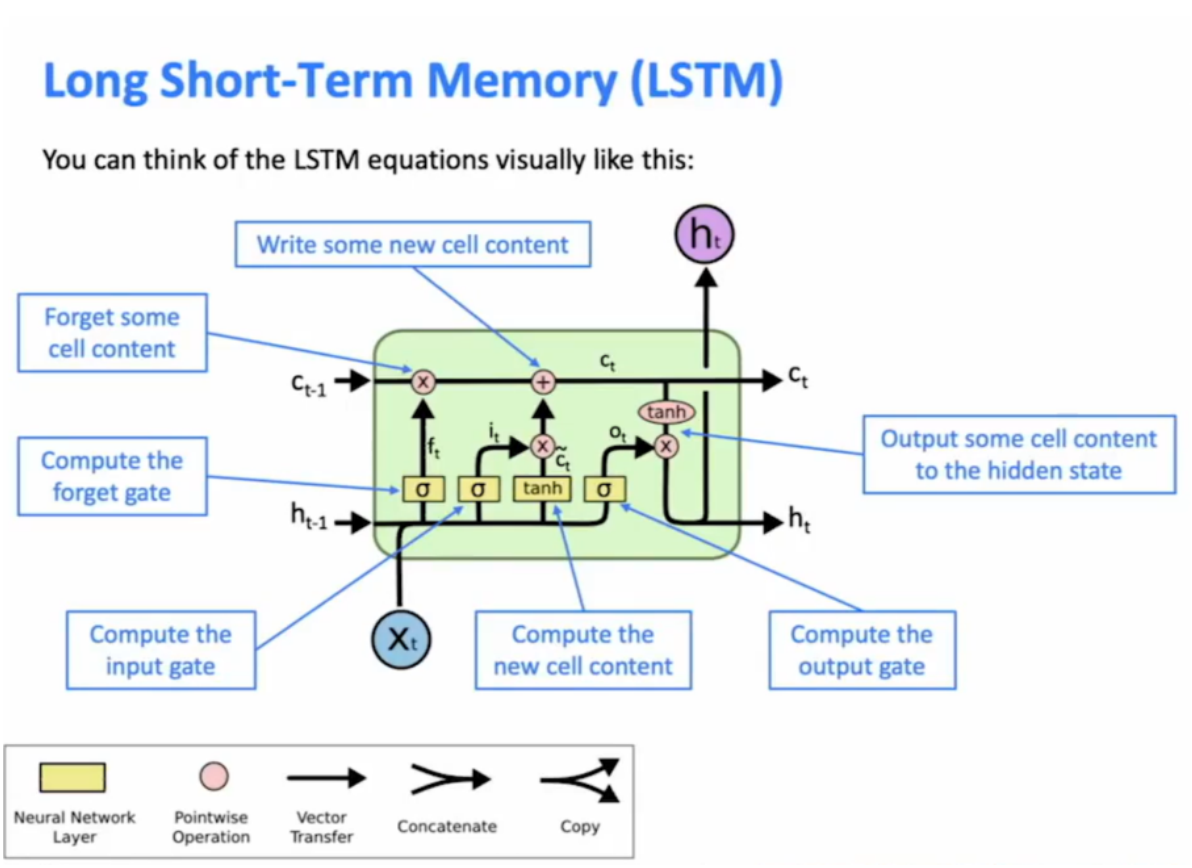
$\hat{\mathbf{g}} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$
if $\|\hat{\mathbf{g}}\| \geq threshold$ **then**
  $\hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|}\hat{\mathbf{g}}$
end if

LSTM and GRU

Long Short-Term Memory (LSTM)







注：

- LSTM中通常把h视作输出，类似RNN的y，因为c是对外不可见的。
- 其实一切只和c有关，h只是c经过一些操作得到的。
- 用tanh只是加上激活函数，类似cnn的线性层后面会加上一个relu。

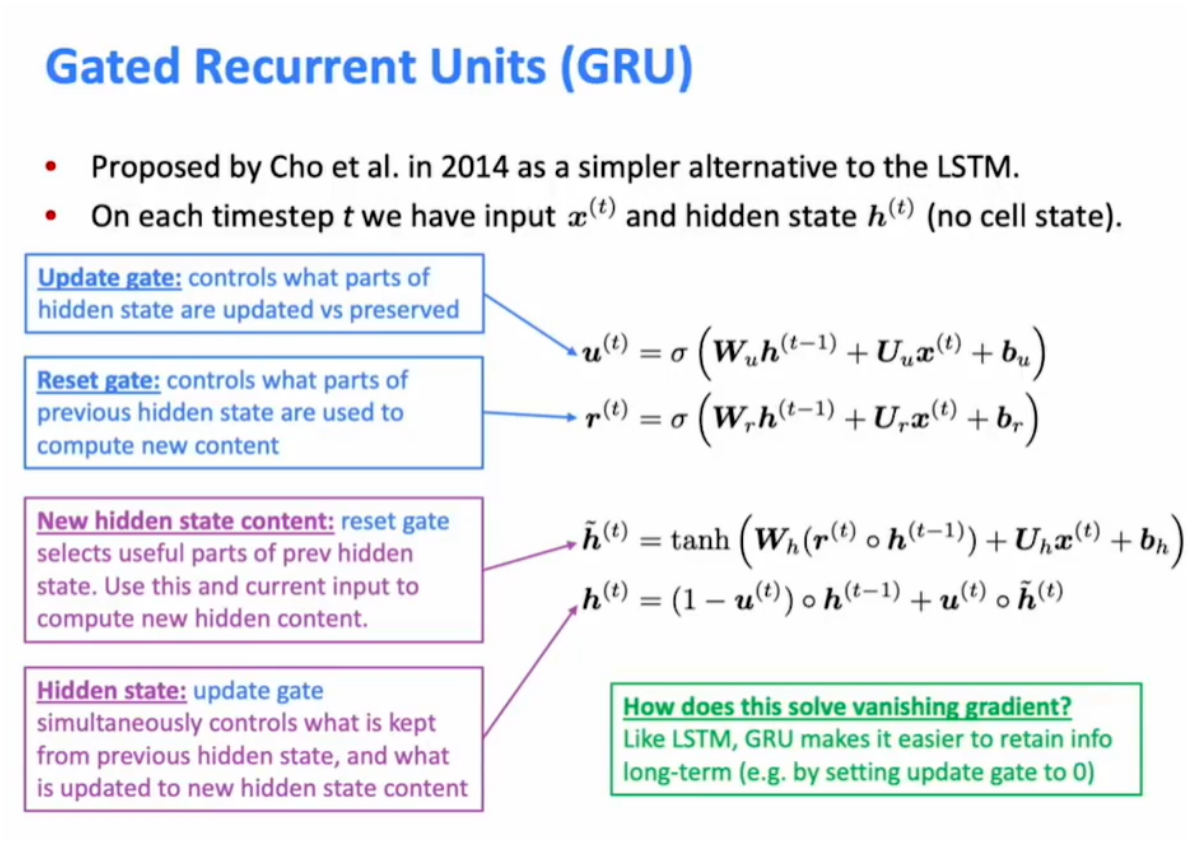
LSTM如何解决vanishing gradients?

RNN的vanishing gradients问题会导致“长期的影响丢失”，即远处传来的梯度很小。

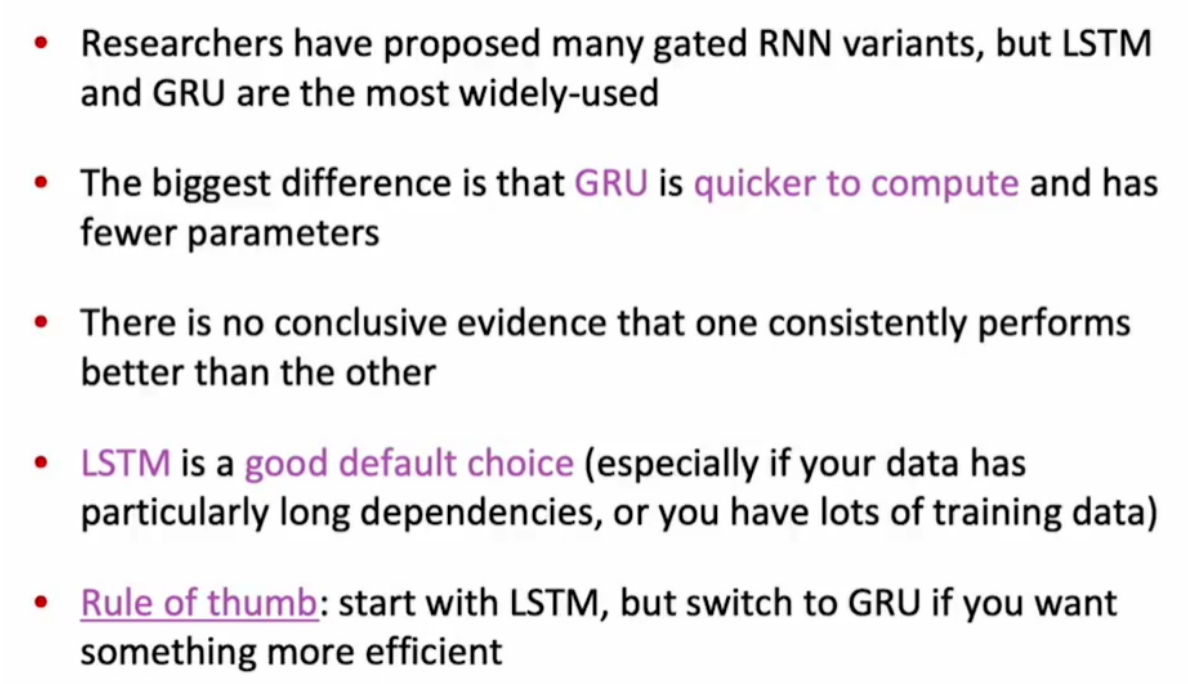
LSTM的解决方式是通过f(遗忘门)来控制保留以前的信息，如果f全1，就会全部保留。换句话说，LSTM并不是从数学上(反向传播)上来避免vanishing gradients，而是减轻了它的后果，更加robust。

Gated Recurrent Units (GRU)

GRU的提出是为了简化LSTM，同时保留LSTM的优点。



LSTM vs GRU



其他网络解决vanishing gradients

ResNet

aka skip-connections。有效是因为identity connection可以至少保持恒等映射。

DenseNet

所有层有加上skip-connections。

HighwayNet

借鉴了LSTM的gate

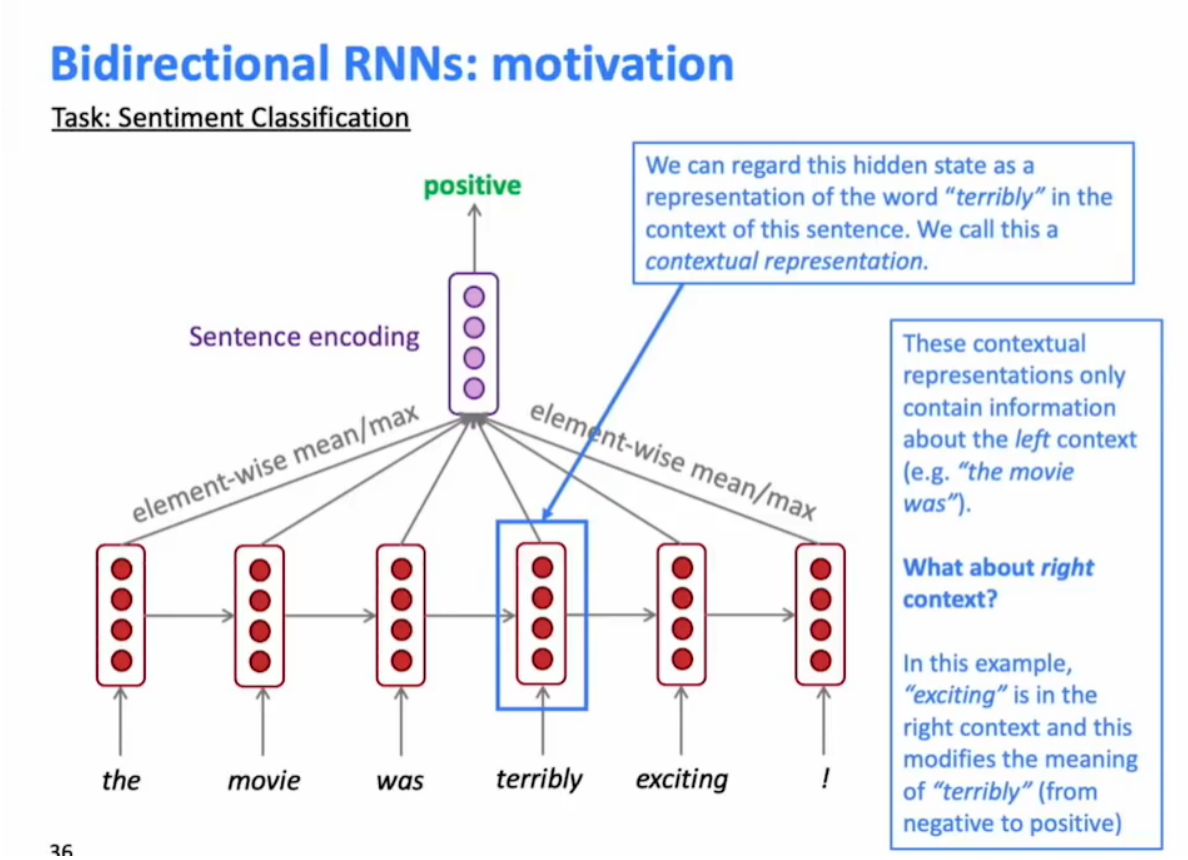
总结：梯度消失不只是RNN的问题，但在RNN里尤其严重，因为重复使用相同的矩阵。

Other Variants

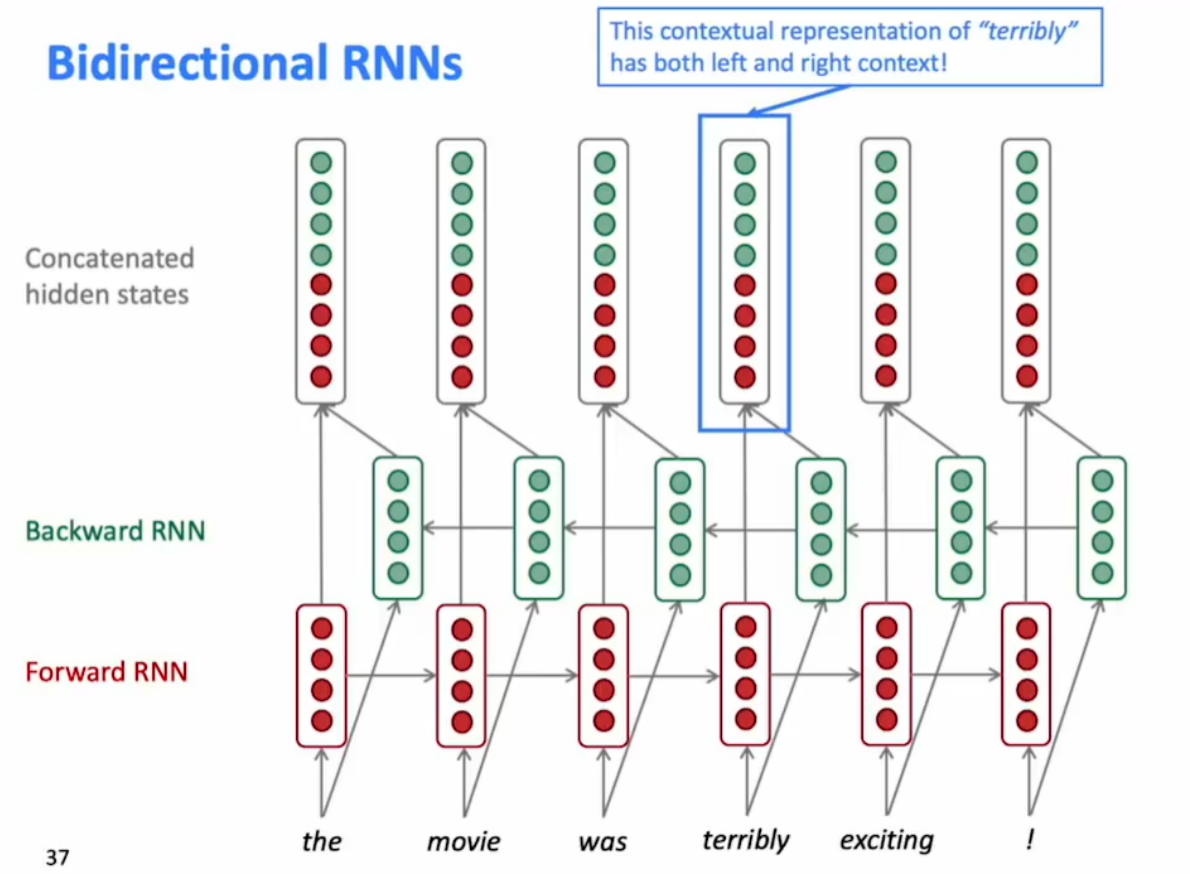
Bidirectional RNNS

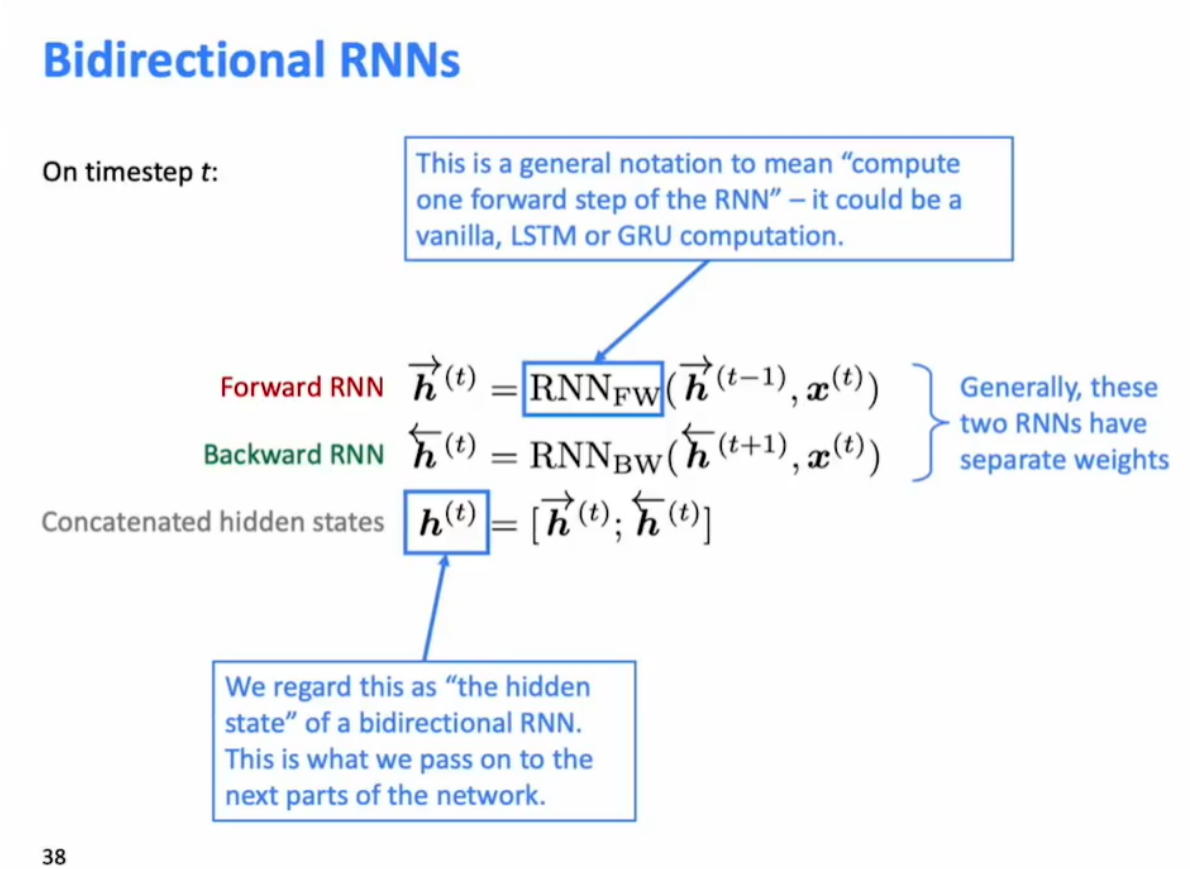
在sentiment classification任务中，每个t的h都作为当前t的特征。用RNN只能包含left context。

在下面的例子中和，希望能包含right context，因为right context正确地解释了terribly。



双向RNN的结构如下：





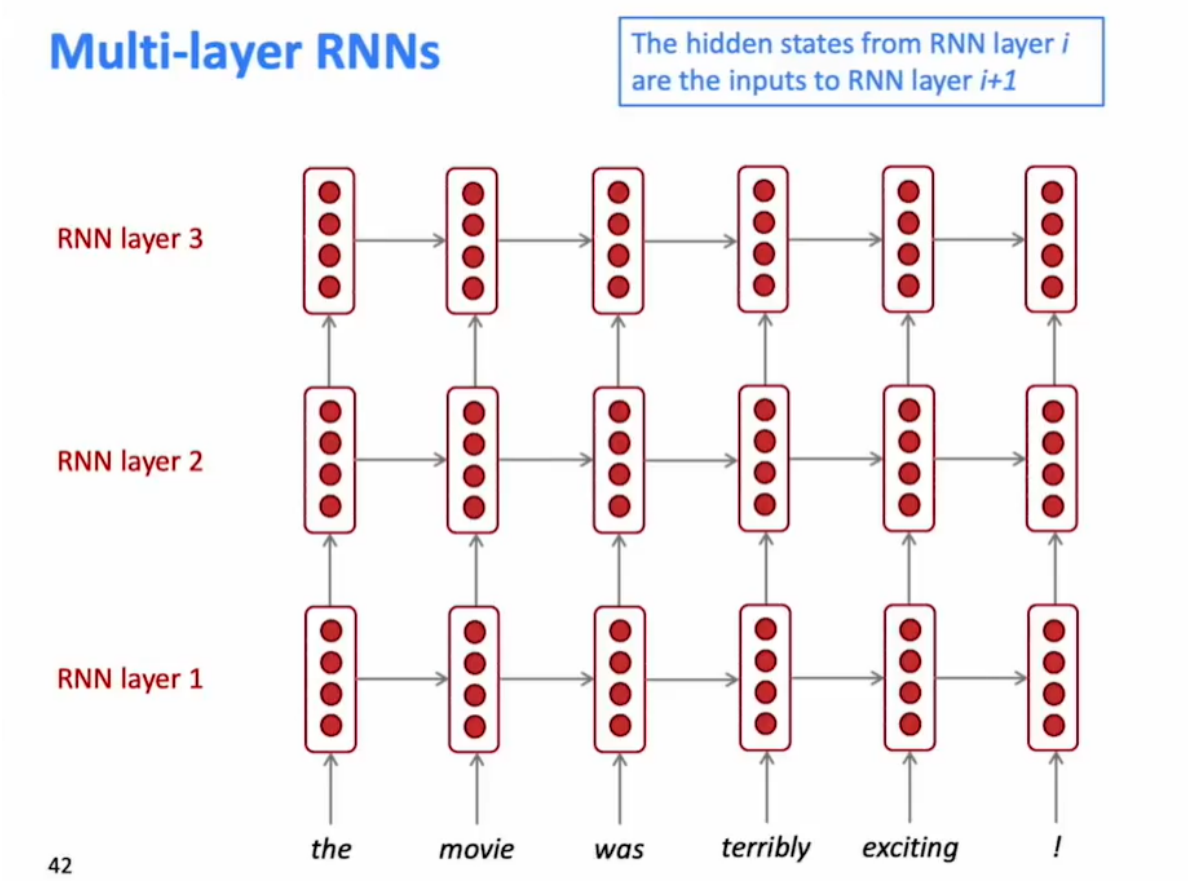
注:

- 双向RNN不能用于LM，只能用于已知完整的sentence。
- 在有完整sentence的情况下，bidirection很有用！

Multi-layer RNNs

在RNN的宽度上扩展，又称stacked RNN。

多层RNN的结构如下:



注:

- Pytorch中会先算第一层，再算第二层，以此类推。但是逻辑上可以先算第一个单词，再算第二个。
- High-performing RNN通常都是多层的，但一般2-4层即可，加上skip-connections可能8层(BERT有24层)。
- 没有几百层的原因是RNN计算是串行的，代价过高。
- 当RNN层数很多时，通常需要加上skip-connections。