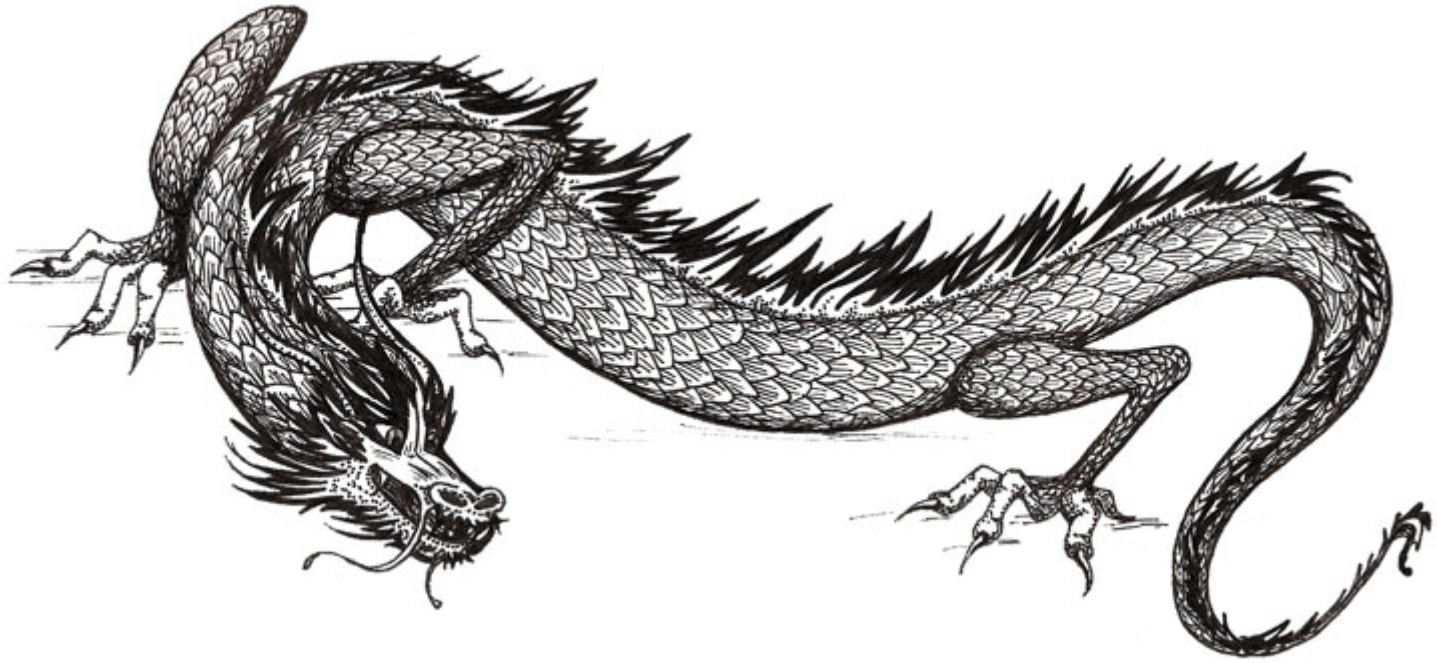
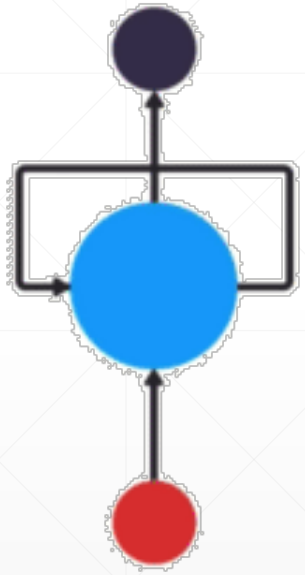


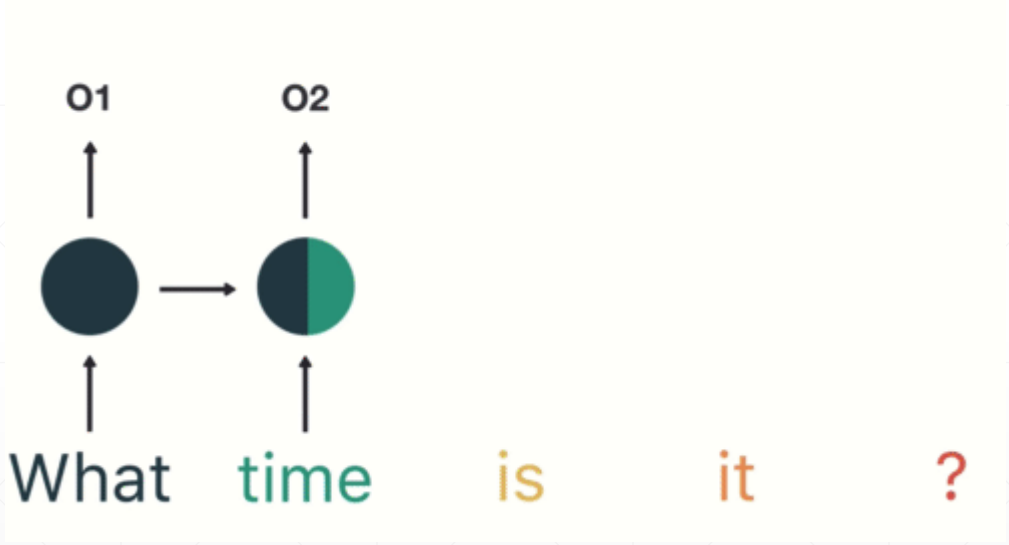
梯度弥散与梯度爆炸

主讲：龙良曲

Recap



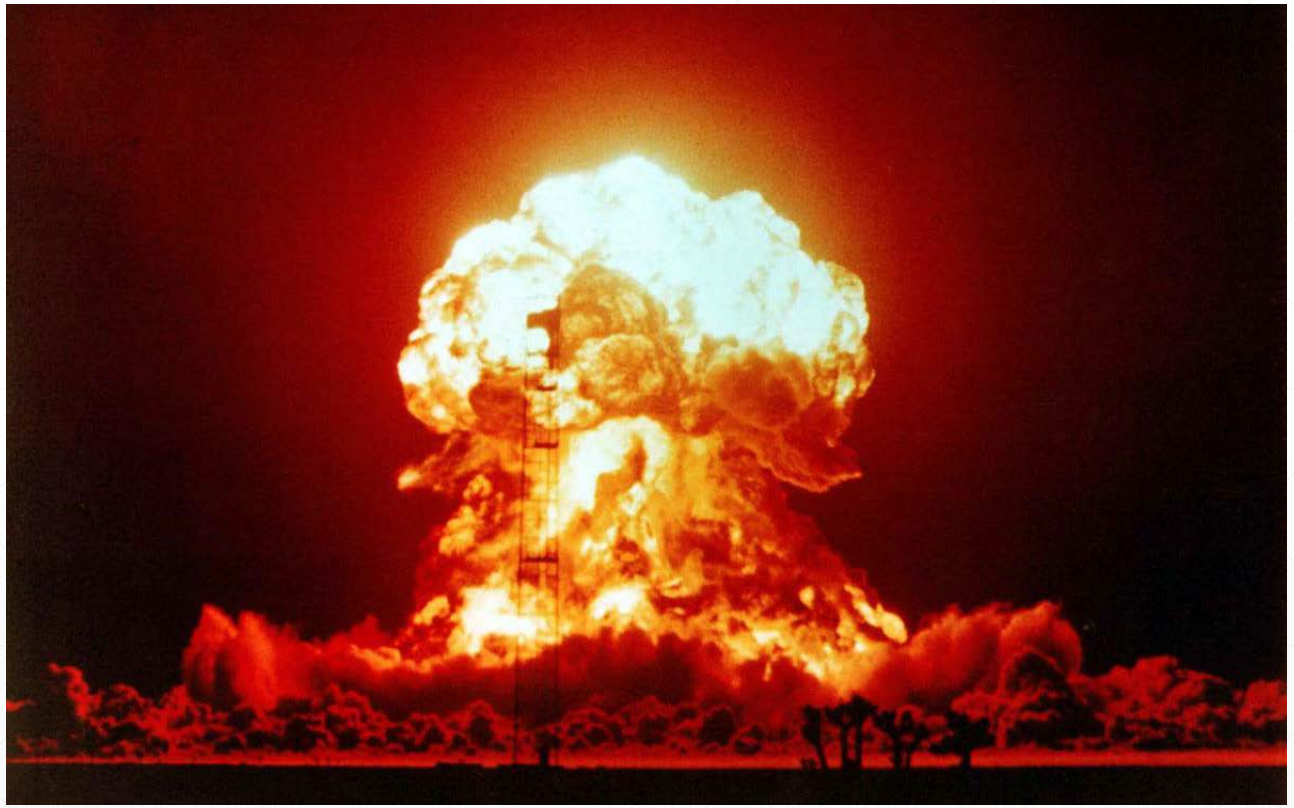
Unroll in time line



Simple Yet?

- Nothing is straightforward.

Gradient Exploding and Gradient Vanishing



Why

$$h_t = tanh(W_I x_t + W_R h_{t-1})$$
$$y_t = W_O h_t$$

$$\frac{\partial E_t}{\partial W_R} = \sum_{i=0}^{t} \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_t}{\partial h_i} \frac{\partial h_i}{\partial W_R}$$

$$\frac{\partial h_t}{\partial h_i} = \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \cdots \frac{\partial h_{i+1}}{\partial h_i} = \prod_{k=i}^{t-1} \frac{\partial h_{k+1}}{\partial h_k}$$

$$\frac{\partial h_{k+1}}{\partial h_k} = diag(f'(W_I x_i + W_R h_{i-1}))W_R$$

$$\frac{\partial h_k}{\partial h_1} = \prod_{i}^{k} diag(f'(W_I x_i + W_R h_{i-1}))W_R$$

https://weberna.github.io/blog/2017/11/15/LSTM-Vanishing-Gradients.html


THE MOST
MOTIVATIONAL POSTER EVER
$1.01^{365}=37.8$
$0.99^{365}=0.03$

Step 1. Gradient Exploding 2013

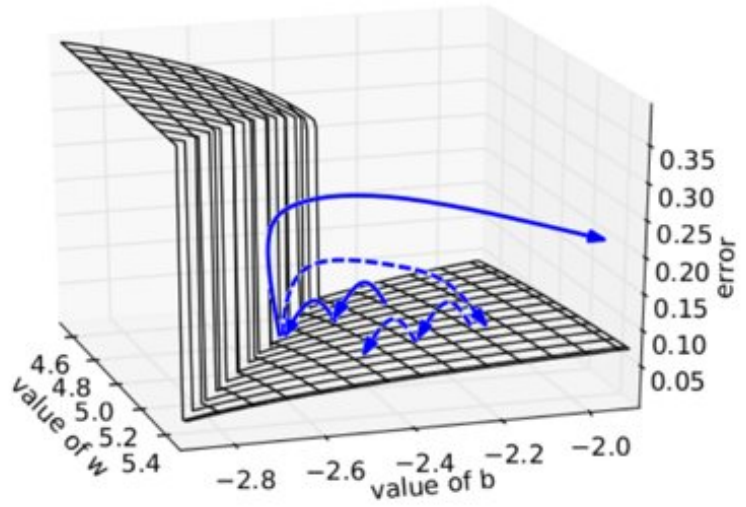
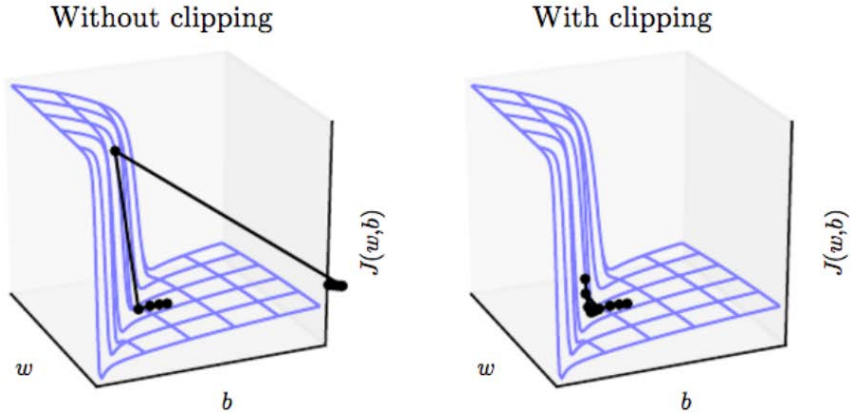


Figure 6. We plot the error surface of a single hidden unit recurrent network, highlighting the existence of high curvature walls. The solid lines depicts standard trajectories that gradient descent might follow. Using dashed arrow the diagram shows what would happen if the gradients is rescaled to a fixed size when its norm is above a threshold.

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



— Goodfellow et al., *Deep Learning*

http://proceedings.mlr.press/v28/pascanu13.pdf

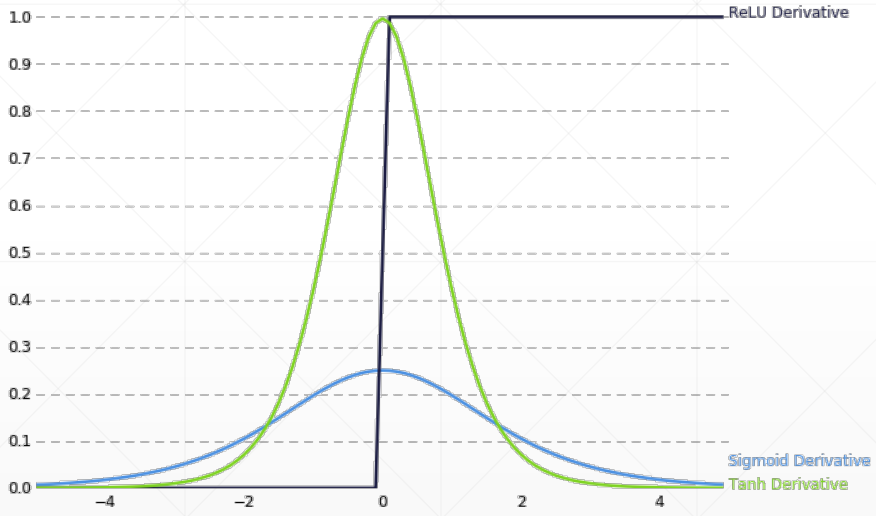
Gradient Clipping

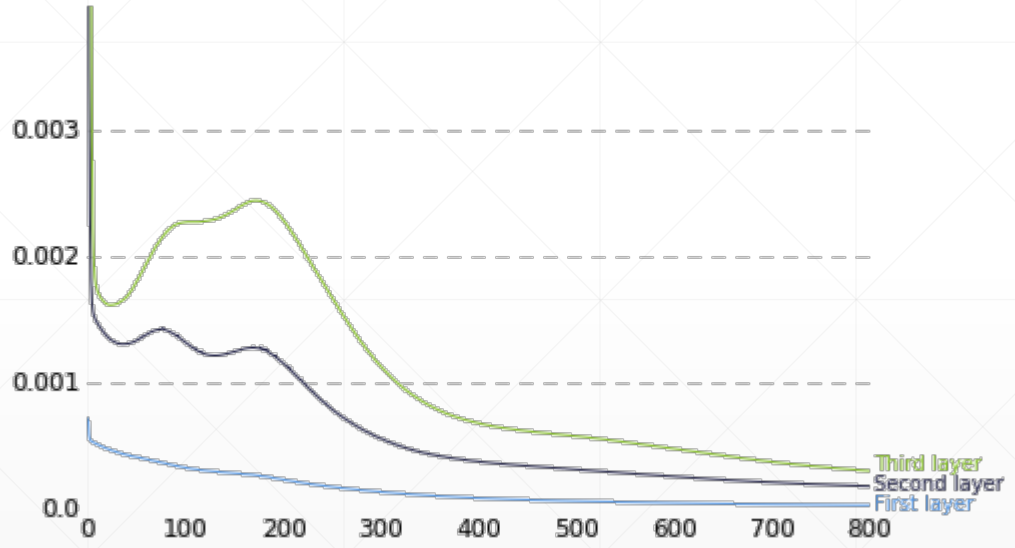
```
with tf.GradientTape() as tape:
    logits = model(x)
    loss = criteon(y, logits)

grads = tape.gradient(loss, model.trainable_variables)
# MUST clip gradient here or it will disconverge!

optimizer.apply_gradients(zip(grads, model.trainable_variables))
```

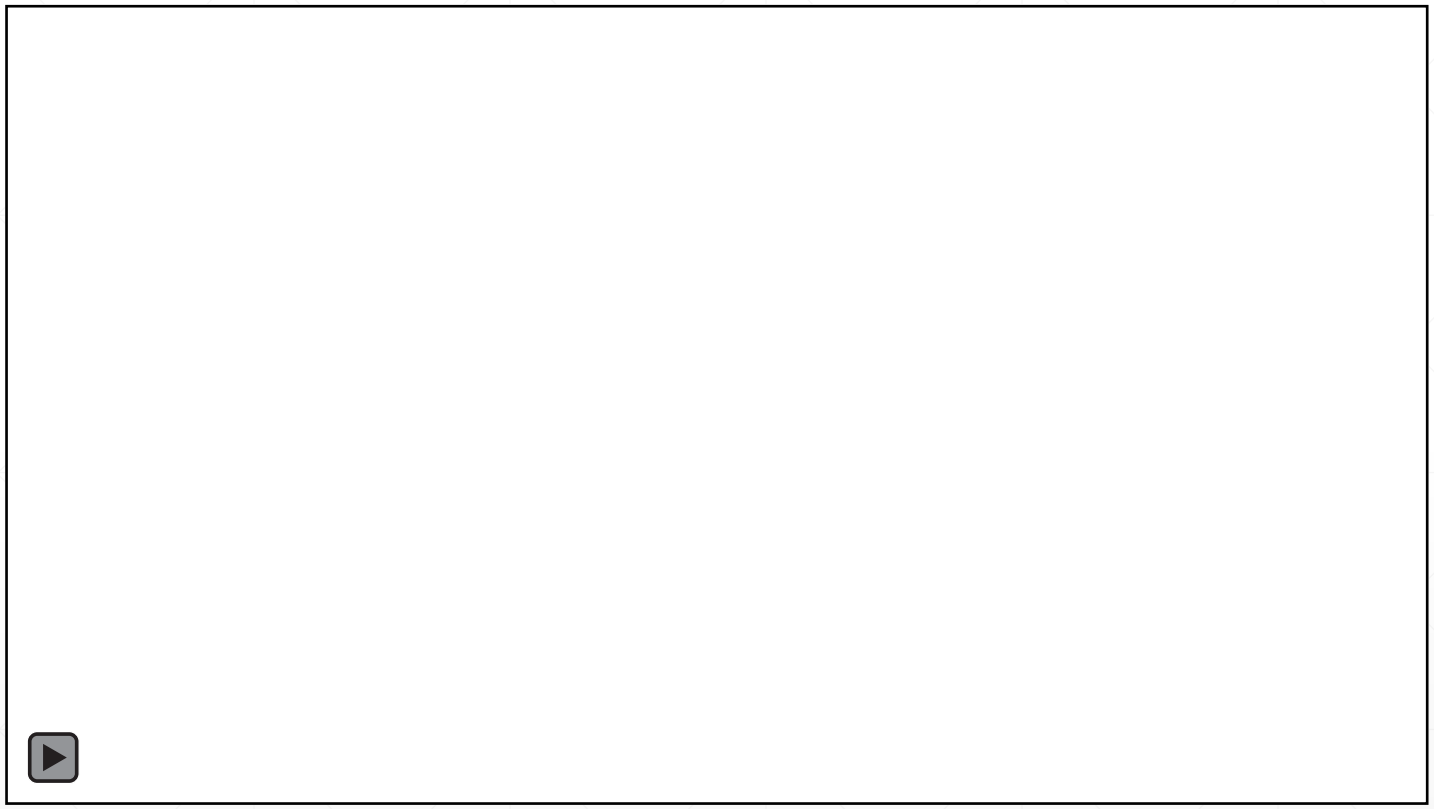
Step 2. Gradient Vanishing: 1997





http://harinisuresh.com/2016/10/09/lstms/

RNN V.S. LSTM Gradient Visualization



https://imgur.com/gallery/vaNahKE

下一课时

Thank You.