Vanishing Gradient

İpek Erdoğan

Güray Baydur

Outline

What is Vanishing Gradient Problem?

How this problem occurs?

How to solve it (CNN Example)

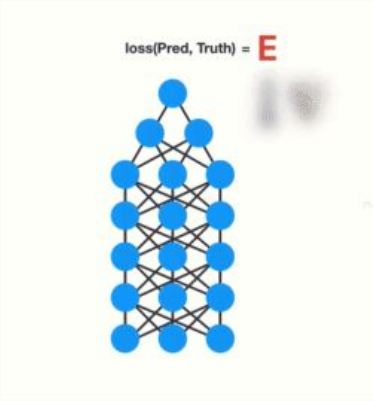
RNN Reminder (Forward and Backward Propagation)

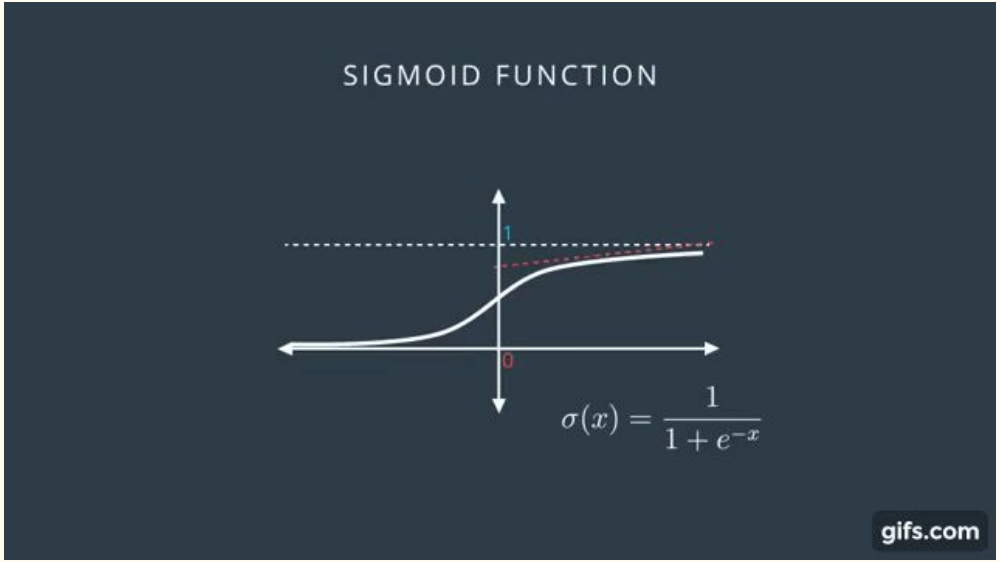
Vanishing Gradient Problem in RNN

How LSTM architecture is more robust to VG problem?

Vanishing Gradient Problem

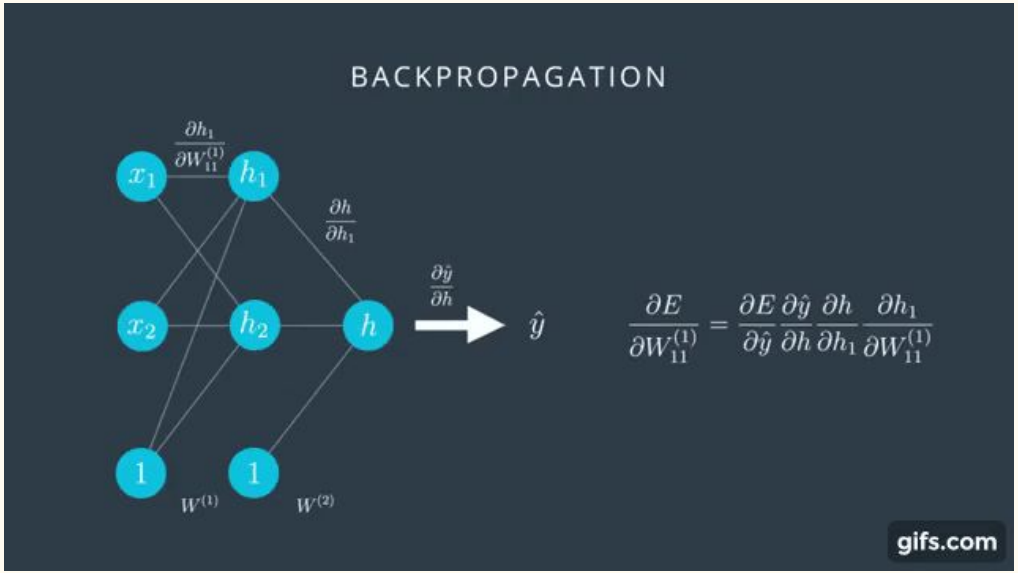
Gradients become so small (vanishingly small) that they are not powerful enough to update network's weights after some point.



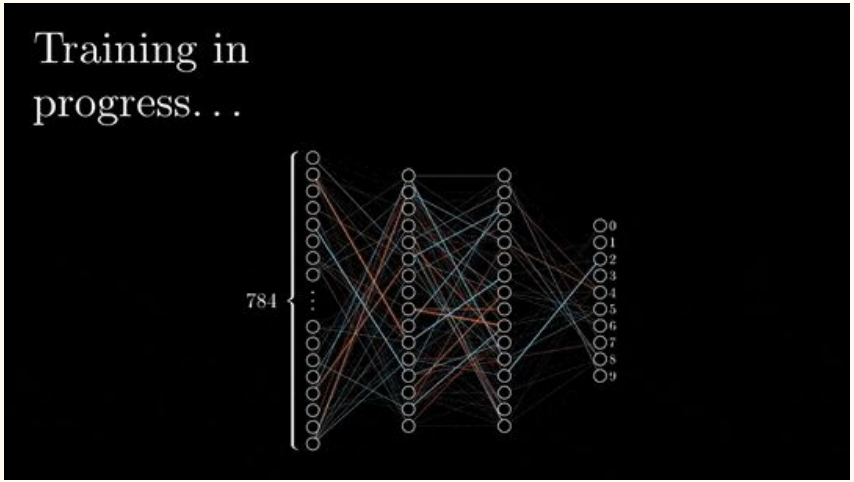


How This Problem Occurs?

The product of some numbers less than one is going to end up with an even smaller number!

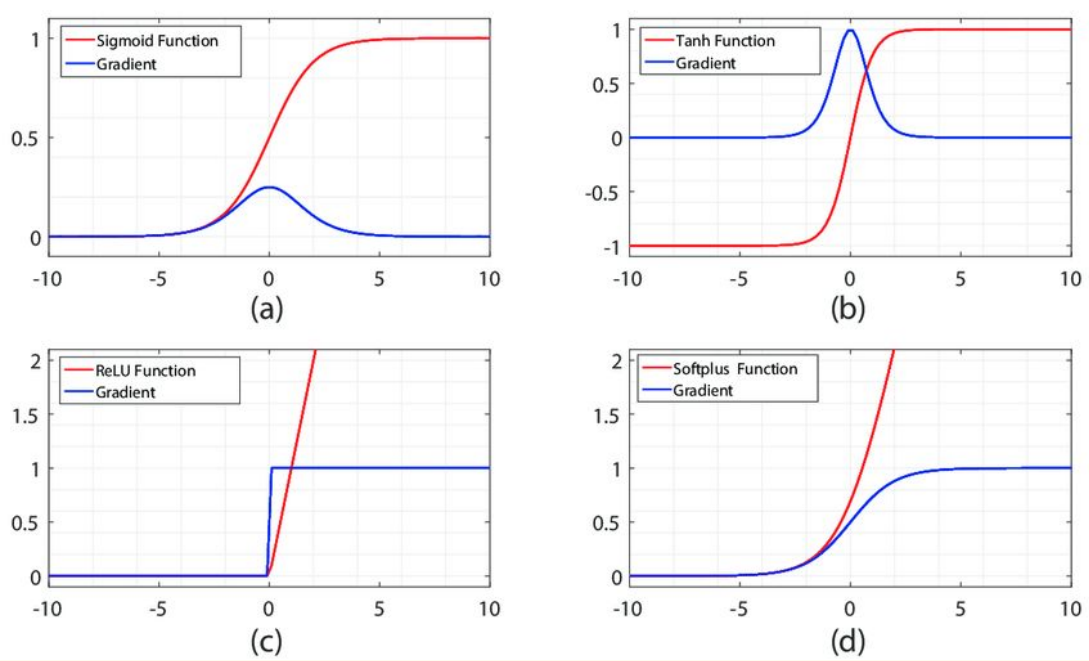


How This Problem Occurs? (CNN Example)



How to Solve It? (CNN Example)

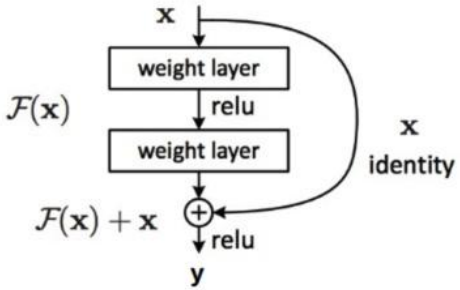
Trying different activation functions.



How to Solve It? (CNN Example)

Adding “skip connections” to the architecture.

DenseNet[3]



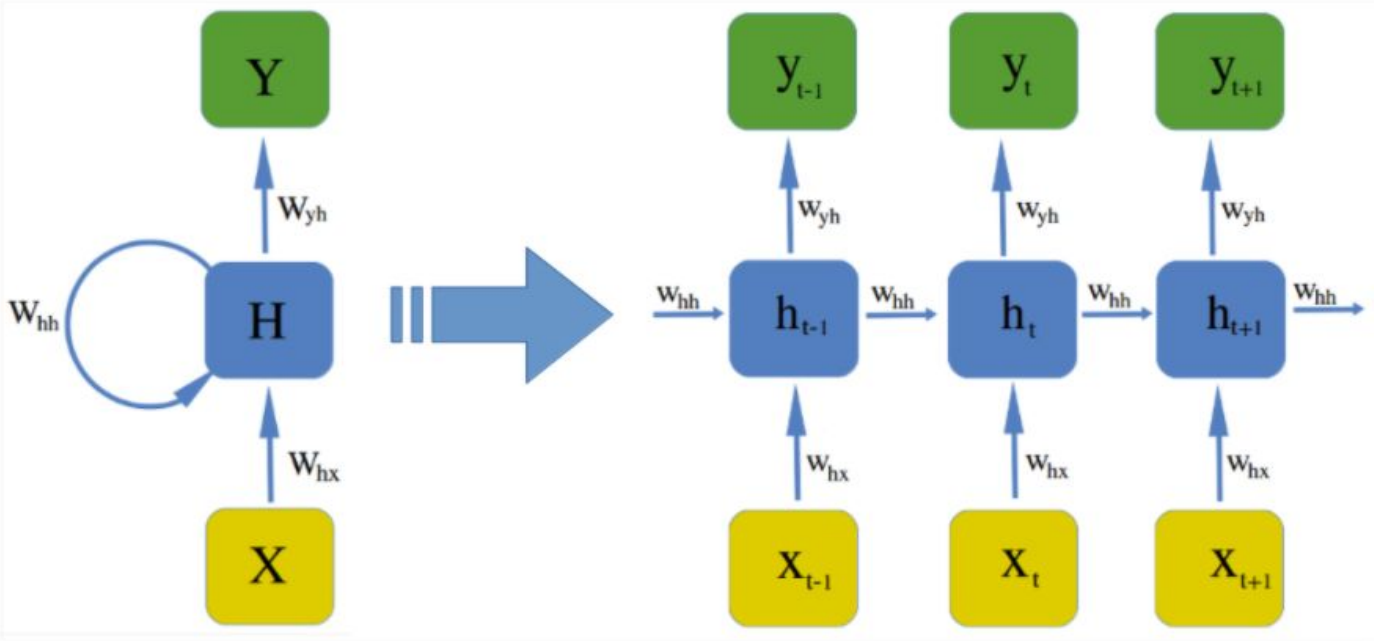
$$y = x + F(x)$$

$$\frac{\delta E}{\delta x} = \frac{\delta E}{\delta y} * \frac{\delta y}{\delta x} = \frac{\delta E}{\delta y} * (1 + F'(x))$$

$$= \frac{\delta E}{\delta y} + \frac{\delta E}{\delta y} * F'(x)$$

ResNet[2]

RNN



RNN Forward Propagation

$$\mathbf{h}_t = f_{\mathbf{W}}\left(\mathbf{h}_{t-1}, \mathbf{x}_t\right) \tag{1}$$

$$\mathbf{h}_t = f\left(\mathbf{W}_{hx}x_t + \mathbf{W}_{hh}h_{t-1} + \mathbf{b}_h\right) \tag{2a}$$

$$\mathbf{h}_t = \tanh\left(\mathbf{W}_{hx}x_t + \mathbf{W}_{hh}h_{t-1} + \mathbf{b}_h\right) \tag{2b}$$

$$\hat{\mathbf{y}}_t = \mathrm{softmax}\left(\mathbf{W}_{yh}h_t + \mathbf{b}_y\right) \tag{3}$$

RNN Backward Propagation through time

$$\frac{\partial \mathbf{E}}{\partial \mathbf{W}} = \sum_{t=1}^{T} \frac{\partial \mathbf{E}_t}{\partial \mathbf{W}}$$

$$\frac{\partial \mathbf{E}}{\partial \mathbf{W}} = \sum_{t=1}^{T} \frac{\partial \mathbf{E}}{\partial \mathbf{y}_t} \frac{\partial \mathbf{y}_t}{\partial \mathbf{h}_t} \overbrace{\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_k}}^{\bigstar} \frac{\partial \mathbf{h}_k}{\partial \mathbf{W}}$$

$$\bigstar \quad \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_k} = \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} \frac{\partial \mathbf{h}_{t-1}}{\partial \mathbf{h}_{t-2}} \cdots \frac{\partial \mathbf{h}_{k+1}}{\partial \mathbf{h}_k} = \prod_{i=k+1}^{t} \frac{\partial \mathbf{h}_i}{\partial \mathbf{h}_{i-1}}$$

Vanishing Gradient in RNN

$$\phi_j = W_{hx} x_j + W_{hh} h_{j-1}$$

$$h_j = \tanh(\phi_j)$$

$$\frac{\partial h_j}{\partial h_{j-1}} = \frac{\partial h_j}{\partial \phi_j} \frac{\partial \phi_j}{\partial h_{j-1}}$$

Vanishing Gradient in RNN

$$\frac{\partial h_j}{\partial \phi_j} = \begin{bmatrix} \frac{\partial h_1}{\partial \phi_1} & \cdots & \frac{\partial h_{j-1}}{\partial \phi_1} & \frac{\partial h_j}{\partial \phi_1} \\ \frac{\partial h_1}{\partial \phi_2} & \ddots & \frac{\partial h_{j-1}}{\partial \phi_2} & \frac{\partial h_j}{\partial \phi_1} \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\partial h_1}{\partial \phi_j} & \cdots & \frac{\partial h_{j-1}}{\partial \phi_j} & \frac{\partial h_j}{\partial \phi_j} \end{bmatrix}_{j \times j}$$

Vanishing Gradient in RNN

$$\frac{\partial h_j}{\partial \phi_j} = \begin{bmatrix} \frac{\partial h_1}{\partial \phi_1} & \dots & 0 & 0 \\ 0 & \frac{\partial h_2}{\partial \phi_2} & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{\partial h_j}{\partial \phi_j} \end{bmatrix}_{j \times j} = diag(\tanh^1(\phi_j))$$

$$\phi_j = W_{hx} x_j + W_{hh} h_{j-1}$$

$$\frac{\partial \phi_j}{\partial h_{j-1}} = W_{hh}$$

Vanishing Gradient in RNN

$$\frac{\partial h_j}{\partial \phi_j} = \begin{bmatrix} \frac{\partial h_1}{\partial \phi_1} & \dots & 0 & 0 \\ 0 & \frac{\partial h_2}{\partial \phi_2} & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{\partial h_j}{\partial \phi_j} \end{bmatrix}_{j \times j} = diag(\tanh^{1}(\phi_j)) \qquad \frac{\partial \phi_j}{\partial h_{j-1}} = W_{hh}$$

$$\frac{\partial h_j}{\partial h_{j-1}} = \frac{\partial h_j}{\partial \phi_j} \frac{\partial \phi_j}{\partial h_{j-1}} \qquad \left\| \frac{\partial h_j}{\partial h_{j-1}} \right\| = \| diag(\tanh^{1}(\phi_j)) W_{hh} \|$$

Vanishing Gradient in RNN

$$\left\| \frac{\partial h_j}{\partial h_{j-1}} \right\| = \| diag(\tanh^1(\phi_j)) W_{hh} \|$$

$$\|AB\| \leq \|A\| \|B\|$$

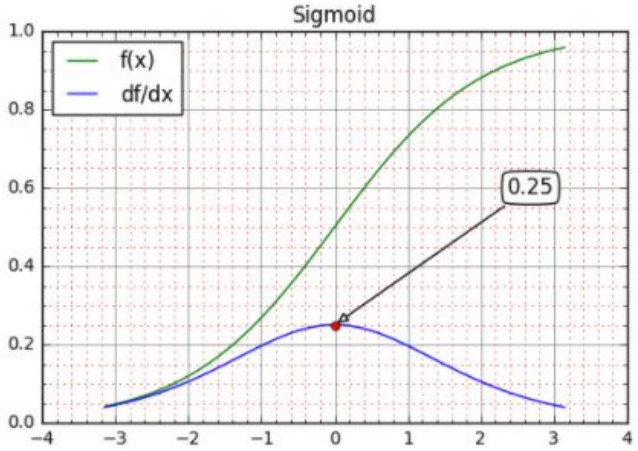
$$\left\| \frac{\partial h_j}{\partial h_{j-1}} \right\| \leq \| diag(\tanh^1(\phi_j)) \| \| W_{hh} \|$$

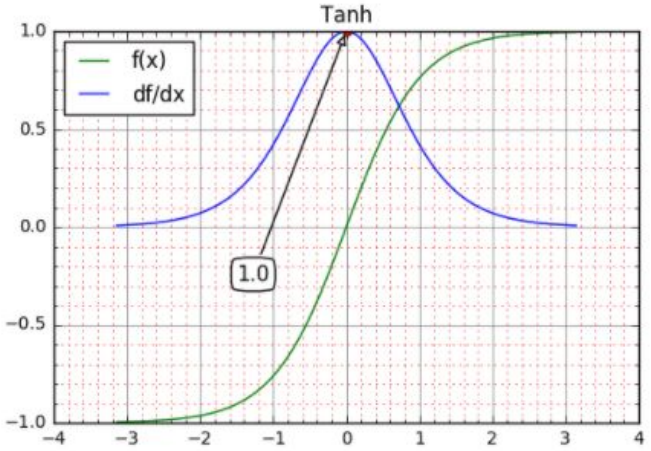
Vanishing Gradient in RNN

$$\left\| \frac{\partial h_j}{\partial h_{j-1}} \right\| \leq \|diag(\tanh^1(\phi_j))\| \|W_{hh}\|$$

$$\alpha = \|diag(\tanh^1(\phi_j))\|$$

$$\beta = \|W_{hh}\|$$





Vanishing Gradient in RNN

$$\left\| \frac{\partial h_j}{\partial h_{j-1}} \right\| \leq \|diag(\tanh^{1}(\phi_j))\| \|W_{hh}\|$$

$$\alpha = \|diag(\tanh^{1}(\phi_j))\|$$

$$\beta = \|W_{hh}\|$$

Pascanu et al [1] showed that that if the largest eigenvalue of Whh is less than 1, then the gradient will shrink exponentially

Vanishing Gradient in RNN

$$\left\| \frac{\partial h_j}{\partial h_{j-1}} \right\| \leq \|diag(\tanh^{1}(\phi_j))\| \|W_{hh}\|$$

$$\left\| \prod_{m=k+1}^{i} \frac{\partial h_m}{\partial h_{m-1}} \right\| \leq \prod_{m=k+1}^{i} \alpha\beta$$

$$\alpha = \|diag(\tanh^{1}(\phi_j))\|$$

$$\beta = \|W_{hh}\|$$

$$\left\| \prod_{m=k+1}^{i} \frac{\partial h_m}{\partial h_{m-1}} \right\| \leq (\alpha\beta)^{(i-k)}$$

Vanishing Gradient in RNN

if $\alpha\beta > 1$, series explode

else $\alpha\beta \leq 1$, series vanish

$$\left\| \prod_{m=k+1}^{i} \frac{\partial h_m}{\partial h_{m-1}} \right\| \leq (\alpha\beta)^{(i-k)}$$

$$\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_k} = \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} \frac{\partial \mathbf{h}_{t-1}}{\partial \mathbf{h}_{t-2}} \cdots \frac{\partial \mathbf{h}_{k+1}}{\partial \mathbf{h}_k} = \prod_{i=k+1}^{t} \frac{\partial \mathbf{h}_i}{\partial \mathbf{h}_{i-1}}$$

$$\frac{\partial \mathbf{E}}{\partial \mathbf{W}} = \sum_{t=1}^{T} \frac{\partial \mathbf{E}}{\partial \mathbf{y}_t} \frac{\partial \mathbf{y}_t}{\partial \mathbf{h}_t} \overbrace{\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_k}}^{\bigstar} \frac{\partial \mathbf{h}_k}{\partial \mathbf{W}}$$

Solutions for Vanishing Gradient Problem

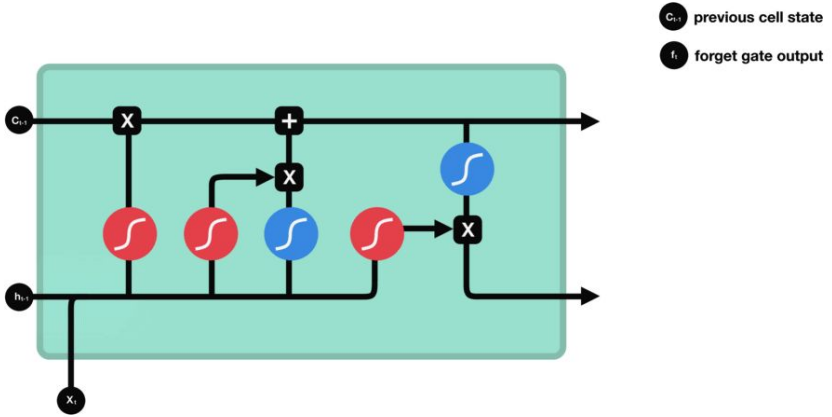
Vanishing Gradient Regularization

$$\frac{\partial \Omega}{\mathbf{W}_{hh}} = \sum_k \frac{\partial \Omega_k}{\mathbf{W}_{hh}}$$

$$= \sum_k \frac{\partial \left(\frac{\left\| \frac{\partial \mathbf{E}}{\partial \mathbf{h}_{k+1}} \mathbf{W}_{hh}{}^{\top} \mathbf{diag}\left(f'(\mathbf{h}_k)\right) \right\|}{\left\| \frac{\partial \mathbf{E}}{\partial \mathbf{h}_{k+1}} \right\|} - 1 \right)^2}{\partial \mathbf{W}_{hh}}$$

How LSTM[4] is Better in terms of Vanishing Gradient?

$$
\begin{aligned}
f_t &= \sigma(W_f[h_{t-1}, x_t]) \\
i_t &= \sigma(W_i[h_{t-1}, x_t]) \\
o_t &= \sigma(W_o[h_{t-1}, x_t]) \\
\widetilde{C}_t &= tanh(W_C[h_{t-1}, x_t]) \\
C_t &= f_t C_{t-1} + i_t \widetilde{C}_t \\
h_t &= o_t tanh(C_t)
\end{aligned}
$$



$$
\frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial C_t}{\partial f_t}\frac{\partial f_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial C_{t-1}} + \frac{\partial C_t}{\partial i_t}\frac{\partial i_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial C_{t-1}} + \frac{\partial C_t}{\partial \widetilde{C}_t}\frac{\partial \widetilde{C}_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial C_{t-1}} + \frac{\partial C_t}{\partial C_{t-1}}
$$

$$
\begin{aligned}
\frac{\partial C_t}{\partial C_{t-1}} &= C_{t-1}\sigma'(\cdot)W_f * o_{t-1}tanh'(C_{t-1}) \\
&+ \widetilde{C}_t\sigma'(\cdot)W_i * o_{t-1}tanh'(C_{t-1}) \\
&+ i_t \tanh'(\cdot)W_C * o_{t-1}tanh'(C_{t-1}) \\
&+ f_t
\end{aligned}
$$

References

[1] Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks." ICML (3) 28 (2013): 1310-1318.

[2] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[3] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[4] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.