$$
gates \begin{cases} i = \sigma(\mathbf{x}_t \mathbf{W}_{xi} + \mathbf{h}_{t-1}\mathbf{W}_{hi} + \mathbf{b}_i) \\ f = \sigma(\mathbf{x}_t \mathbf{W}_{xf} + \mathbf{h}_{t-1}\mathbf{W}_{hf} + \mathbf{b}_f) \\ o = \sigma(\mathbf{x}_t \mathbf{W}_{xo} + \mathbf{h}_{t-1}\mathbf{W}_{ho} + \mathbf{b}_o) \\ g = \tanh(\mathbf{x}_t \mathbf{W}_{xg} + \mathbf{h}_{t-1}\mathbf{W}_{hg} + \mathbf{b}_g) \end{cases}
\qquad
states \begin{cases} \mathbf{c}_t = f \odot \mathbf{c}_{t-1} + i \odot g \\ \mathbf{h}_t = o \odot \tanh(\mathbf{c}_t) \end{cases}
$$

$$
gates \begin{cases} i = \sigma(\mathbf{z}_i) = \sigma(\mathbf{x}_t \mathbf{W}_{xi} + \mathbf{h}_{t-1}\mathbf{W}_{hi} + \mathbf{b}_i) \\ f = \sigma(\mathbf{z}_f) = \sigma(\mathbf{x}_t \mathbf{W}_{xf} + \mathbf{h}_{t-1}\mathbf{W}_{hf} + \mathbf{b}_f) \\ o = \sigma(\mathbf{z}_o) = \sigma(\mathbf{x}_t \mathbf{W}_{xo} + \mathbf{h}_{t-1}\mathbf{W}_{ho} + \mathbf{b}_o) \\ g = \tanh(\mathbf{z}_g) = \tanh(\mathbf{x}_t \mathbf{W}_{xg} + \mathbf{h}_{t-1}\mathbf{W}_{hg} + \mathbf{b}_g) \end{cases}
$$

$$
states \begin{cases} \mathbf{c}_t = f \odot \mathbf{c}_{t-1} + i \odot g \\ \mathbf{h}_t = o \odot \tanh(\mathbf{c}_t) \end{cases}
$$

$$
\begin{aligned}
\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}_i} &= \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{c}_t}\frac{\partial \mathbf{c}_t}{\partial i}\frac{\partial i}{\partial \mathbf{z}_i} + \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{h}_t}\frac{\partial \mathbf{h}_t}{\partial \mathbf{c}_t}\frac{\partial \mathbf{c}_t}{\partial i}\frac{\partial i}{\partial \mathbf{z}_i} \\
&= g \odot \sigma'(\mathbf{z}_i)\left(\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{c}_t} + \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{h}_t} \odot o \odot \tanh'(\mathbf{c}_t)\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{b}} &= \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{b}} + \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{b}} \\
&= \sum_{k=1}^{n}\left(\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}_f} &= \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{c}_t}\frac{\partial \mathbf{c}_t}{\partial f}\frac{\partial f}{\partial \mathbf{z}_f} + \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{h}_t}\frac{\partial \mathbf{h}_t}{\partial \mathbf{c}_t}\frac{\partial \mathbf{c}_t}{\partial f}\frac{\partial f}{\partial \mathbf{z}_f} \\
&= \mathbf{c}_{t-1} \odot \sigma'(\mathbf{z}_f)\left(\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{c}_t} + \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{h}_t} \odot o \odot \tanh'(\mathbf{c}_t)\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{W}_h} &= \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{W}_h} + \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{W}_h} \\
&= \mathbf{h}_{t-1}^T\left(\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}_o} &= \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{h}_t}\frac{\partial \mathbf{h}_t}{\partial o}\frac{\partial o}{\partial \mathbf{z}_o} \\
&= \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{h}_t} \odot \tanh(\mathbf{c}_t) \odot \sigma'(\mathbf{z}_o)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{W}_x} &= \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{W}_x} + \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{W}_x} \\
&= \mathbf{x}_t^T\left(\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}_g} &= \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{c}_t}\frac{\partial \mathbf{c}_t}{\partial g}\frac{\partial g}{\partial \mathbf{z}_g} + \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{h}_t}\frac{\partial \mathbf{h}_t}{\partial \mathbf{c}_t}\frac{\partial \mathbf{c}_t}{\partial g}\frac{\partial g}{\partial \mathbf{z}_g} \\
&= i \odot \tanh'(\mathbf{z}_g)\left(\frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{c}_t} + \frac{\partial L_t(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{h}_t} \odot o \odot \tanh'(\mathbf{c}_t)\right)
\end{aligned}
$$

$$
gates\begin{cases} i = \sigma(\mathbf{z}_i) = \sigma(\mathbf{x}_t\mathbf{W}_{xi} + \mathbf{h}_{t-1}\mathbf{W}_{hi} + \mathbf{b}_i) \\ f = \sigma(\mathbf{z}_f) = \sigma(\mathbf{x}_t\mathbf{W}_{xf} + \mathbf{h}_{t-1}\mathbf{W}_{hf} + \mathbf{b}_f) \\ o = \sigma(\mathbf{z}_o) = \sigma(\mathbf{x}_t\mathbf{W}_{xo} + \mathbf{h}_{t-1}\mathbf{W}_{ho} + \mathbf{b}_o) \\ g = \tanh(\mathbf{z}_g) = \tanh(\mathbf{x}_t\mathbf{W}_{xg} + \mathbf{h}_{t-1}\mathbf{W}_{hg} + \mathbf{b}_g) \end{cases}
$$

$$
states\begin{cases} \mathbf{c}_t = f \odot \mathbf{c}_{t-1} + i \odot g \\ \mathbf{h}_t = o \odot \tanh(\mathbf{c}_t) \end{cases}
$$

$$
\begin{aligned}
\frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{c}_{t-1}} &= \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{c}_t}\frac{\partial \mathbf{c}_t}{\partial \mathbf{c}_{t-1}} + \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{h}_t}\frac{\partial \mathbf{h}_t}{\partial \mathbf{c}_t}\frac{\partial \mathbf{c}_t}{\partial \mathbf{c}_{t-1}} \\
&= \frac{\partial \mathbf{c}_t}{\partial \mathbf{c}_{t-1}}\left(\frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{c}_t} + \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{h}_t}\frac{\partial \mathbf{h}_t}{\partial \mathbf{c}_t}\right) \\
&= f \odot \mathbf{c}'_t
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{c}'_t &= \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{c}_t} + \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{h}_t}\frac{\partial \mathbf{h}_t}{\partial \mathbf{c}_t} \\
&= \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{c}_t} + \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{h}_t} \odot o \odot \tanh'(\mathbf{c}_t)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{h}_{t-1}} &= \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{c}_t}\frac{\partial \mathbf{c}_t}{\partial \mathbf{h}_{t-1}} + \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{h}_t}\left(\frac{\partial \mathbf{h}_t}{\partial \mathbf{c}_t}\frac{\partial \mathbf{c}_t}{\partial \mathbf{h}_{t-1}} + \frac{\partial \mathbf{h}_t}{\partial o}\frac{\partial o}{\partial \mathbf{h}_{t-1}}\right) \\
&= \left(\frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{c}_t} + \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{h}_t}\frac{\partial \mathbf{h}_t}{\partial \mathbf{c}_t}\right)\frac{\partial \mathbf{c}_t}{\partial \mathbf{h}_{t-1}} + \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{h}_t}\frac{\partial \mathbf{h}_t}{\partial o}\frac{\partial o}{\partial \mathbf{h}_{t-1}} \\
&= \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{c}_t}\left(\frac{\partial \mathbf{c}_t}{\partial f}\frac{\partial f}{\partial \mathbf{z}_f}\frac{\partial \mathbf{z}_f}{\partial \mathbf{h}_{t-1}} + \frac{\partial \mathbf{c}_t}{\partial i}\frac{\partial i}{\partial \mathbf{z}_i}\frac{\partial \mathbf{z}_i}{\partial \mathbf{h}_{t-1}} + \frac{\partial \mathbf{c}_t}{\partial g}\frac{\partial g}{\partial \mathbf{z}_g}\frac{\partial \mathbf{z}_g}{\partial \mathbf{h}_{t-1}}\right) + \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{h}_t}\frac{\partial \mathbf{h}_t}{\partial o}\frac{\partial o}{\partial \mathbf{z}_o}\frac{\partial \mathbf{z}_o}{\partial \mathbf{h}_{t-1}} \\
&= \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{z}_f}\frac{\partial \mathbf{z}_f}{\partial \mathbf{h}_{t-1}} + \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{z}_i}\frac{\partial \mathbf{z}_i}{\partial \mathbf{h}_{t-1}} + \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{z}_g}\frac{\partial \mathbf{z}_g}{\partial \mathbf{h}_{t-1}} + \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{z}_o}\frac{\partial \mathbf{z}_o}{\partial \mathbf{h}_{t-1}} \\
&= \frac{\partial L_t(\mathbf{y},\hat{\mathbf{y}})}{\partial \mathbf{z}}\mathbf{W}_h^T
\end{aligned}
$$