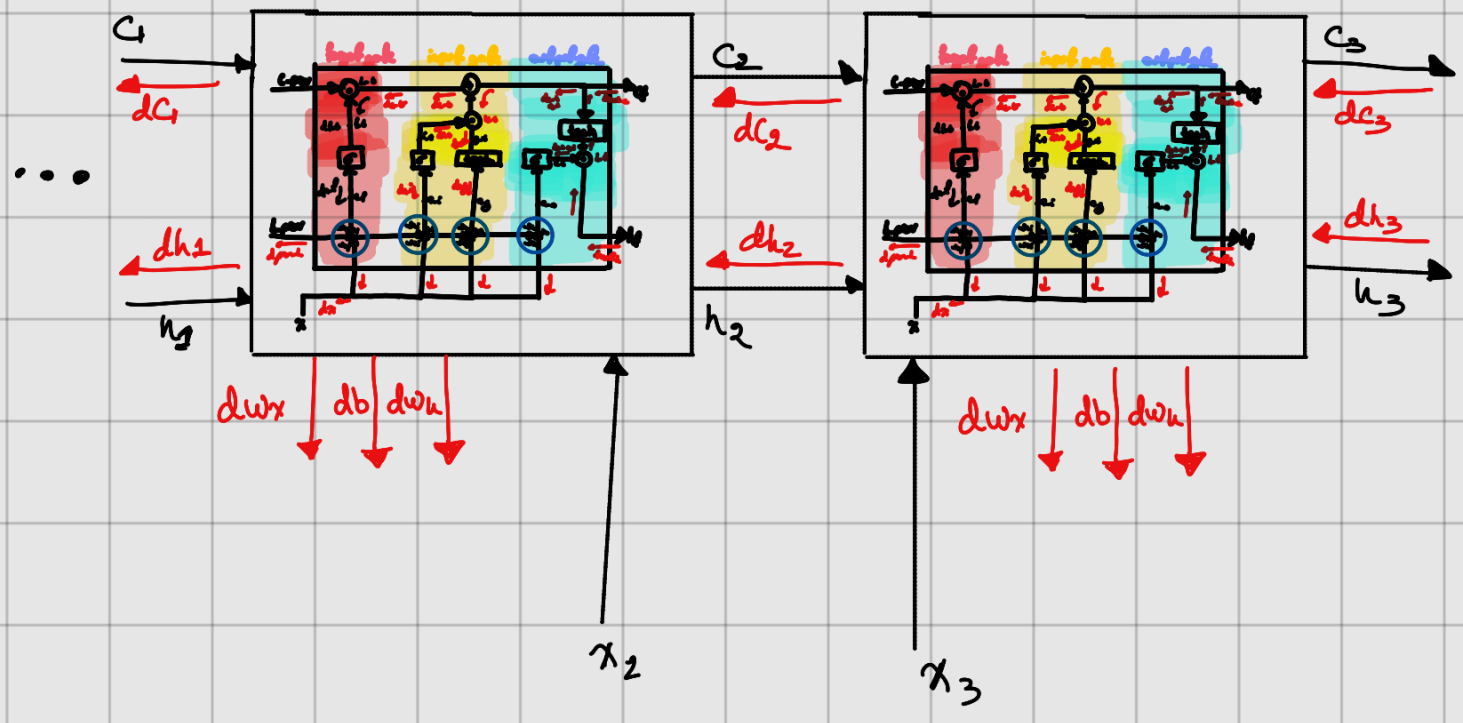


similar to the example of a RNN, gradients are back propagated through time.

- dWx, db, and dWh are internal state variables shared across the LSTM cells in time, so the gradients get summed

$\therefore$ $dWx \mathrel{+}= dWx\_t$ $(D, 4H)$

$dWh \mathrel{+}= dWh\_t$ $(H, 4H)$

$db \mathrel{+}= db\_t$ $(4H,)$

- dx is just returned from the step backprop and stored in a variable. We don't accumulate them because each timestep input is unique.

$\therefore$ $dx[:,T,:] = dx\_T \rightarrow (N,D)$

$(N,T,D)$ $(N,H)$

- dprev_h is returned from the step backprop and accumulated. dh0 is the accumulation at the first timestep.