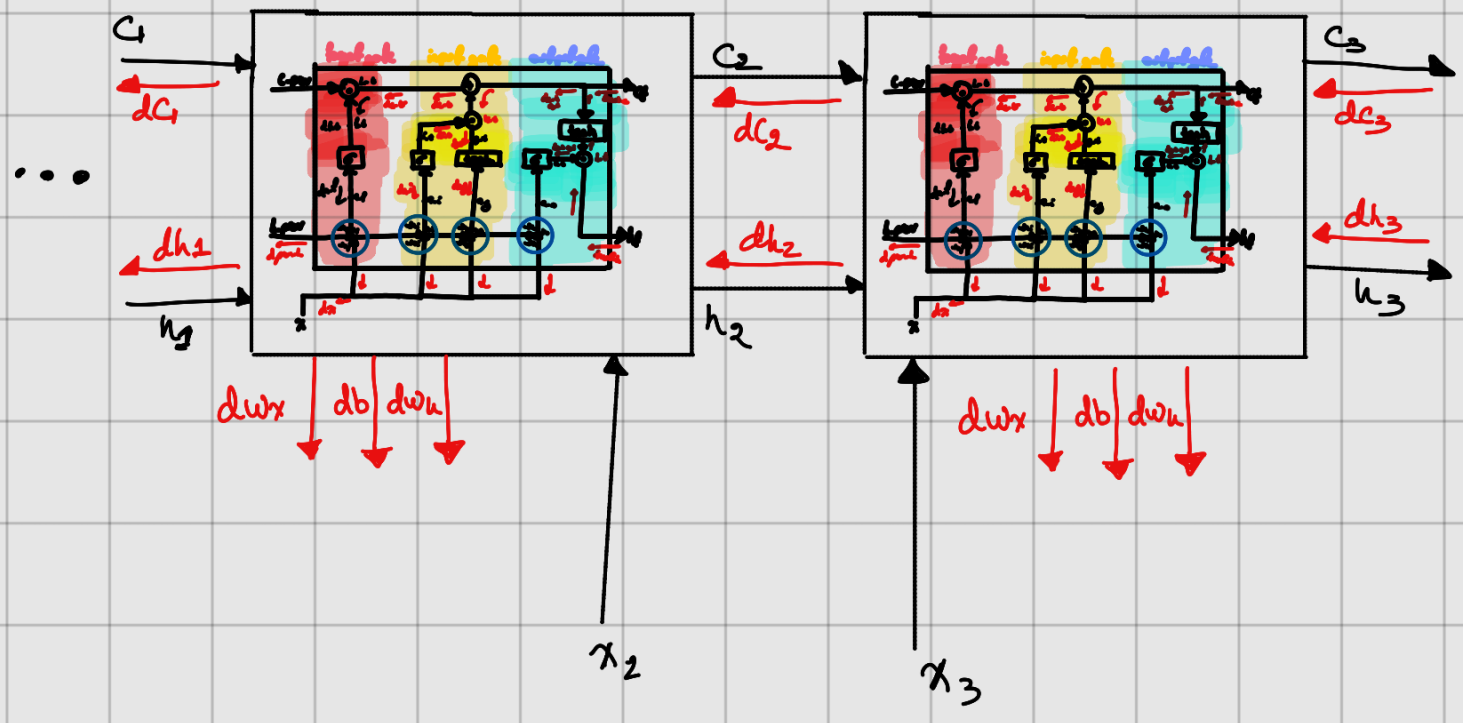

similar to the example of a RNN, gradients are back propagated through time.

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

$\therefore$ $dWx \mathrel{+}= dWx\_t$

$dWh \mathrel{+}= dWh\_t$

$db \mathrel{+}= db\_t$

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

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