[[1508.03790] Depth-Gated LSTM](https://arxiv.org/abs/1508.03790)

**Video Summary**

None

**Prerequisite Stuff (Roots)**

[LSTM Stuff](https://docs.google.com/document/d/1sPZQ8wtE_RbSOcb23FBHaUJM3GkBbd7eAA6SjMsS83A/edit)

Stacked LSTM, htL is used as input xtL + 1 for the upper layer LSTM

**c**t is the output of "cell" portion of an LSTM cell

Memory cell has linear dependence of current and past activity and input, forget, and output gates to modulate activity

SimpleRNN equation:



**Notes**

Depth gate connecting memory cell ctL + 1 to ctL controlling how much flow from the lower memory cell directly to the upper layer memory cell

Gate function is a logistic function:



Wxd is the weight matrix to relate depth gate to the input of this layer, and both the past memory cell (at time t - 1) and lower layer memory cell are related through their own weight matrices/vec

Memory cell at layer L + 1 is computed as follows, using the above gate function



Depth gate to extract info from the lower layer, forget gate to choose what to keep from previous cell memory, and input gate to decide how to incorporate input

The idea of gated linear dependence can also be used to connect the first memory layer cell **c**t1 with the input xt0

The depth gate is calculated as:



Memory cell is computed as:



**Branches**

In the transformer paper, **c**t of the encoder and decoder are obtained differently by using hidden states of the encoder (as K and V) and hidden states of the decoder (as Q) to compute compatibility, which makes them better than the depth-gated LSTM for calculating context

Hidden state and context have same dimensionality, so matrix multiplication and dimensionality checks out