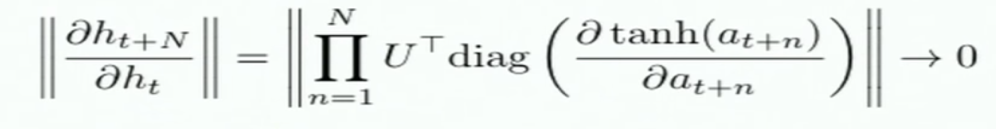
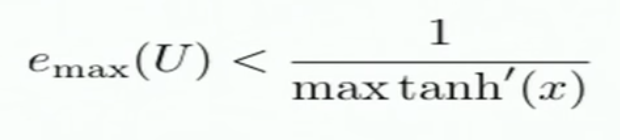
Lecture 11 | Gated Recurrent Units and Further Topics in NMT

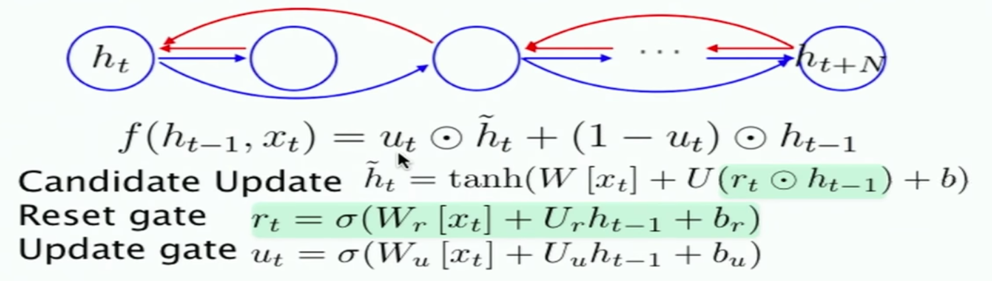
* For GRUs, vanishing gradient is super-problematic
  + When we only observe:

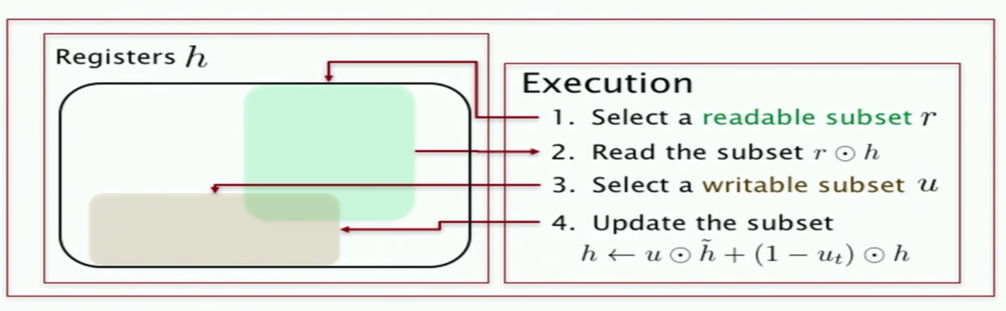


* + - We cannot tell whether:
      * No dependency between t and t+n in data
      * Wrong configuration or parameters:

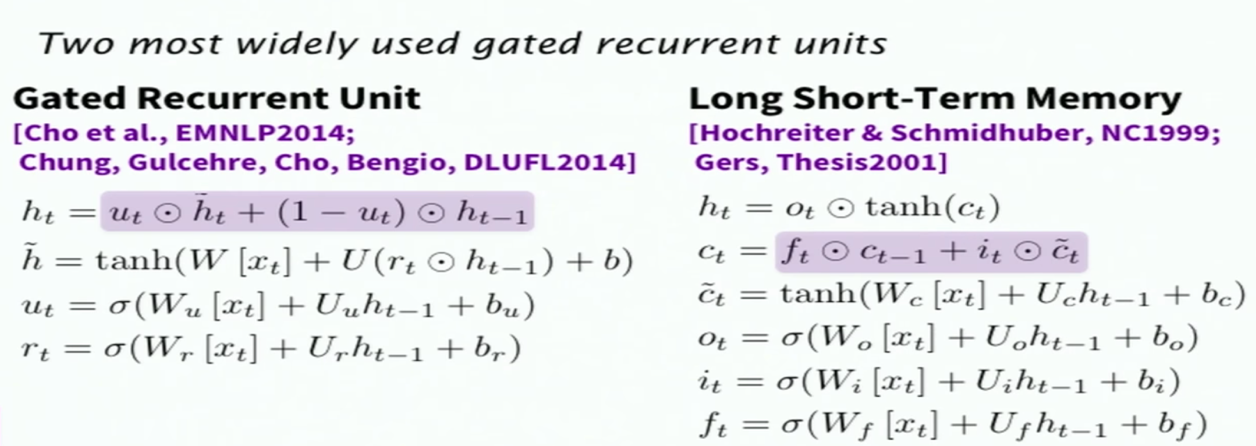


* + We want to be able to learn how much effect things have on our latest decisions by backpropagating through this whole sequence of intermediate nodes.
    - We would like to get more direct evidence of the effect of early time steps on much later time steps without having to do the long sequence matrix multiples – **adaptive** **shortcut connections**, therefore learn about long-term dependencies





* + - Reset gate decides which part of the hidden state to read to update the hidden state
    - If you keep doing matrix multiplication along the RNN, we will lose the notion of long term memory. Therefore, sometimes it’s good to just carry from the previous hidden state rather than calculating a new hidden state



* + The h\_t of GRU is behaving very similarly as the c\_t of LSTM
  + f\_t is ‘don’t forget’ gate
* Training a gated RNN
  + Use an LSTM or GRU
  + Initialise recurrent matrices to be orthogonal
  + Initialise other matrices with a sensible (small) scale
  + Initialise forget gate bias to 1 (default to remembering)
  + Use adaptive learning rate algo (Adam, AdaDelta,…)
  + Clip the norm of the gradient (1-5 seems to be a reasonable threshold)
  + Either only dropout vertically or learn how to do it horizontally
  + Ensembling: train 8 – 10 nets and average their predictions
* MT evaluation
  + Manual evaluation
    - SSER (Subjective sentence error rate)
    - Adequacy and Fluency
    - Comparative ranking of translations
  + BLEU (Bilingual Evaluation Understudy) score
    - You have a **reference translation** (and we know that there are many possible ways that something can be translated)
    - Machine translation of the same text
    - **Therefore, what percent of machine n-grams can be found in the reference translation**
    - For each n-gram size, you are not allowed to match identical portion of reference translation more than once
    - **Brevity penalty** – can’t just type out single word ‘the’ (precision 1.0)
    - Commonly n-grams up to 4