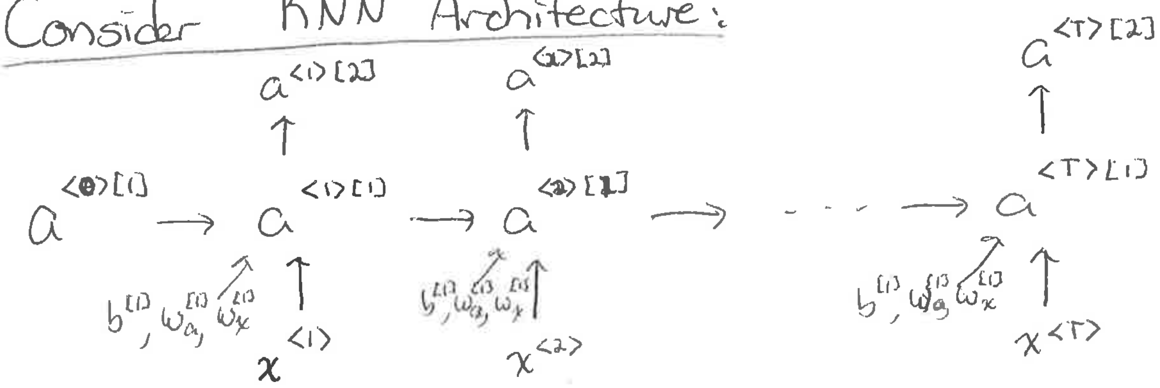
Motivation for GRU & LSTM:

Consider RNN Architecture:



$$z^{<t>[1]} = b^{[1]} + (W_x^{[1]})^T \cdot x^{<t>} + (W_a^{[1]})^T \cdot a^{<t-1>[1]}$$

$$a^{<t>[1]} = \tanh(z^{<t>[1]})$$

Back propagation is complicated.

One of the terms involved looks like:

$$\frac{\partial J}{\partial a^{<T>[2]}} \cdot \frac{\partial a^{<T>[2]}}{\partial a^{<T>[1]}} \cdot \underbrace{\frac{\partial a^{<T>[1]}}{\partial a^{<T-1>[1]}} \cdot \frac{\partial a^{<T-1>[1]}}{\partial a^{<T-2>[1]}} \cdots \frac{\partial a^{<2>[1]}}{\partial a^{<1>[1]}}} \cdot \frac{\partial a^{<1>[1]}}{\partial W_x^{[1]}}$$

- term for each time step
- Potential problem if ~~these~~ terms are close to 0 (vanishing gradient) or very large (exploding gradient)!!
- But no big deal if $\frac{\partial a^{<t>[1]}}{\partial a^{<t-1>[1]}} = 1$

↳ no big deal if $a^{<t>[1]} = a^{<t-1>[1]}$.

Main idea of (simplified) GRU: make it so sometimes, $a^{<t>[1]} \approx a^{<t-1>[1]}$.

- temporarily, rename to $c^{<t>}$ instead of $a^{<t>}$
  ↳ c for "memory cell"
- Generate a candidate for the updated memory cell in the usual way:

$$\tilde{c}^{<t>[1]} = \tanh(b_c^{[1]} + (W_{cc}^{[1]})^T c^{<t-1>[1]} + (W_{cx}^{[1]})^T x^{<t>})$$

- Generate a update gate of same shape as $c^{<t>}$ where #'s are approx. 0 or 1:

$$\Gamma_u^{<t>[1]} = \sigma(b_u^{[1]} + (W_{uc}^{[1]})^T \cdot c^{<t-1>[1]} + (W_{ux}^{[1]})^T x^{<t>})$$

- for entries where update gate is 1, use new value in $\tilde{c}^{<t>}$.
  " " " " " 0, use old value in $c^{<t-1>}$

$$c^{<t>[1]} = \Gamma_u * \tilde{c}^{<t>[1]} + (1 - \Gamma_u) * c^{<t-1>[1]}$$

elementwise products

GRU (Gated Recurrence Unit)

- Add one more thing:

  a gate Γ_r (for relevence) saying which elements ~~kinds of items~~ of $c^{<t-1>}$ are used for calculating $c^{<t>}$ (which are relevant?)

update gate: $\Gamma_u^{<t>[l]} = \sigma\left(b_u^{[l]} + (W_{uc}^{[l]})^T \cdot c^{<t-1>[l]} + (W_{ux}^{[l]})^T x^{<t>}\right)$

relevance gate: $\Gamma_r^{<t>[l]} = \sigma\left(b_r^{[l]} + (W_{rc}^{[l]})^T \cdot c^{<t-1>[l]} + (W_{rx}^{[l]})^T x^{<t>}\right)$

memory cell proposal: $\tilde{c}^{<t>[l]} = \tanh\left(b_c^{[l]} + (W_{cc}^{[l]})^T \cdot (\Gamma_r^{<t>[l]} * c^{<t-1>}) + (W_{cx}^{[l]})^T x^{<t>}\right)$

memory cell: $c^{<t>[l]} = \Gamma_u^{<t>[l]} * \tilde{c}^{<t>[l]} + (1 - \Gamma_u^{<t>}) * c^{<t-1>[l]}$

activation is same as memory cell: $a^{<t>[l]} = c^{<t>[l]}$

all of this happens within 1 cell/circle instead of just $z^{<t>}$ and $a^{<t>}$!

LSTM (Long Short Term Memory)

- Same basic set up as GRU
- Different configuration of Gates
  - − no relevance gate
  - + add a "forget" gate used in place of $(1-\Gamma_u)$
  - + add an "output" gate used to get $a^{<t>}$ from $c^{<t>}$

update gate: $\Gamma_u = \dots$

forget gate: $\Gamma_f = \dots$

output gate: $\Gamma_o = \dots$

no relevence gate

memory cell proposal: $\tilde{c}^{<t>[l]} = \tanh\left(b_c^{[l]} + (W_{cc}^{[l]})^T \cdot c^{<t-1>[l]} + (W_{cx}^{[l]})^T x^{<t>}\right)$

memory cell: $c^{<t>[l]} = \Gamma_u^{<t>[l]} * \tilde{c}^{<t>[l]} + \Gamma_f * c^{<t-1>[l]}$

Activation output: $a^{<t>[l]} = \Gamma_o^{<t>[l]} * c^{<t>[l]}$

instead of $(1-\Gamma_u)$

"forget" previous memory cell entries where $\Gamma_f = 0$, keep ones where $\Gamma_f = 1$.