Modern RNN Units

- unit = building block of RNN

LSTM: Long Short-Term Memory

GRU: Gated Recurrent Unit (simplified LSTM)

Problem with RNNs

- RNN is nested composite fn (due to feedback)
- x_1 or x_t is most deeply nested.
- For chain rule, derivative of composite fn = multiplication
- RNN susceptible to vanishing / exploding gradients.
- Simple RNNs cannot learn well for long time sequences.
  - i.e forgets info from earlier due to vanishing gradient.
- LSTMs concept was created in 1997, GRU in 2014 (simpler same principle)

GRU: Gated Recurrent Unit

- same design as simple RNN

- uses gating mechanisms to remember / forget h_{t-1}

update gate ↓ candidate hidden state ↓

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

Update gate → $z_t = \sigma(W_{xz}^T x_t + W_{hz}^T \cdot h_{t-1} + \beta_z)$

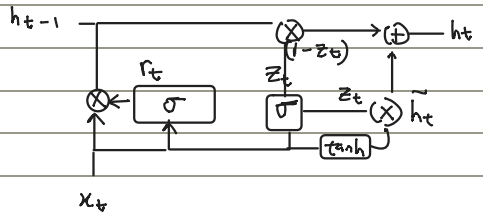
Candidate hidden → $\tilde{h}_t = \tanh(W_{xh}^T x_t + W_{hh}^T (r_t \odot h_{t-1}) + \beta_h)$

Reset gate → $r_t = \sigma(W_{xr}^T x_t + W_{hr}^T h_{t-1} + \beta_r)$ ← reset gate

z_t, h_t, r_t are vectors of size M

└ No. of hidden units

GRU Diagram



⊙ : Hadamard product (element wise product) : $\begin{bmatrix}2\\4\end{bmatrix} \odot \begin{bmatrix}3\\5\end{bmatrix} = \begin{bmatrix}6\\20\end{bmatrix}$

Shapes

From x_t → Gate : D x M

From h_t → Gate : M x M (feedback)

All Bias → M.

Purpose of z_t, $\tilde{h}_t$, r_t

1) Update gate z_t :

└ inside h_t eqn : Take new $\tilde{h}_t$ or keep old h_{t-1}

z_t output is sigmoid, ∴ P(t) & 1 - P(t)

if $z_t \to 0$, $1 - z_t \to 1$ (i.e remember old value)

if $z_t \to 1$, $1 - z_t \to 0$ (i.e take new value)

* z_t activation should always be sigmoid.

└ behaves like a binary classifier → Take new h_t / keep old h_{t-1}

* Not exactly kept or discarded, both components still present :

* "Gate analogy" → how big gate opened

2) Reset gate r_t :

- By itself is just another simple RNN neuron
- used as: $(r_t \odot h_{t-1})$ inside $\tilde{h}_t$
- r_t is always b/w 0 and 1 due to activation fn
- used to determine which part of h_{t-1} to remember.
- As $r_t \to 0$, $(r_t \odot h_{t-1}) \to 0$ i.e "reset" h_{t-1} back to 0