Sequence Models

Machine Learning: Jordan Boyd-Graber
University of Maryland

LSTM VARIANTS
GRU simplifies slightly

No explicit memory

Only one gate
GRU simplifies slightly

\[ z_t = \sigma (W_z \cdot [h_{t-1}, x_t]) \]
\[ r_t = \sigma (W_r \cdot [h_{t-1}, x_t]) \]
\[ \tilde{h}_t = \tanh (W \cdot [r_t \times h_{t-1}, x_t]) \]
\[ h_t = (1 - z_t) \times h_{t-1} + z_t \times \tilde{h}_t \]

Slightly fewer parameters
What’s most important part of LSTM

Greff et al. explore

- No Input Gate (NIG)
- No Forget Gate (NFG)
- No Output Gate (NOG)
- No Input Activation Function (NIAF)
- No Output Activation Function (NOAF)
- No Peepholes (NP)
- Coupled Input and Forget Gate (CIFG) : GRU, $f_t = 1 - i_t$
- Full Gate Recurrence (FGR): Original LSTM paper
What’s most important part of LSTM

Greff et al. explore

- No Input Gate (NIG)
- No Forget Gate (NFG)
- No Output Gate (NOG)
- No Input Activation Function (NIAF)
- No Output Activation Function (NOAF)
- No Peepholes (NP)
- Coupled Input and Forget Gate (CIFG): GRU, $f_t = 1 - i_t$
- Full Gate Recurrence (FGR): Original LSTM paper
What’s most important part of LSTM

Greff et al. explore

- No Input Gate (NIG)
- No Forget Gate (NFG)
- No Output Gate (NOG)
- No Input Activation Function (NIAF)
- No Output Activation Function (NOAF)
- No Peepholes (NP)
- Coupled Input and Forget Gate (CIFG) : GRU, $f_t = 1 - i_t$
- Full Gate Recurrence (FGR): Original LSTM paper
What’s most important part of LSTM

Greff et al. explore

- No Input Gate (NIG)
- No Forget Gate (NFG)
- No Output Gate (NOG)
- No Input Activation Function (NIAF)
- No Output Activation Function (NOAF)
- No Peepholes (NP)
- Coupled Input and Forget Gate (CIFG) : GRU, $f_t = 1 - i_t$
- Full Gate Recurrence (FGR): Original LSTM paper
What’s most important part of LSTM

Greff et al. explore
- No Input Gate (NIG)
- No Forget Gate (NFG)
- No Output Gate (NOG)
- No Input Activation Function (NIAF)
- No Output Activation Function (NOAF)
- No Peepholes (NP)
- Coupled Input and Forget Gate (CIFG): GRU, $f_t = 1 - i_t$
- Full Gate Recurrence (FGR): Original LSTM paper
What’s most important part of LSTM

Greff et al. explore

- No Input Gate (NIG)
- No Forget Gate (NFG)
- No Output Gate (NOG)
- No Input Activation Function (NIAF)
- No Output Activation Function (NOAF)
- No Peepholes (NP)
- Coupled Input and Forget Gate (CIFG) : GRU, \( f_t = 1 - i_t \)
- Full Gate Recurrence (FGR): Original LSTM paper
What’s most important part of LSTM

Greff et al. explore

- No Input Gate (NIG)
- No Forget Gate (NFG)
- No Output Gate (NOG)
- No Input Activation Function (NIAF)
- No Output Activation Function (NOAF)
- No Peepholes (NP)
- Coupled Input and Forget Gate (CIFG): GRU, $f_t = 1 - i_t$
- Full Gate Recurrence (FGR): Original LSTM paper
What’s most important part of LSTM

Greff et al. explore

- No Input Gate (NIG)
- No Forget Gate (NFG)
- No Output Gate (NOG)
- No Input Activation Function (NIAF)
- No Output Activation Function (NOAF)
- No Peepholes (NP)
- Coupled Input and Forget Gate (CIFG) : GRU, $f_t = 1 - i_t$
- Full Gate Recurrence (FGR): Original LSTM paper
What’s most important part of LSTM

Greff et al. explore

- No Input Gate (NIG)
- No Forget Gate (NFG)
- No Output Gate (NOG)
- No Input Activation Function (NIAF)
- No Output Activation Function (NOAF)
- No Peepholes (NP)
- Coupled Input and Forget Gate (CIFG) : GRU, \( f_t = 1 - i_t \)
- Full Gate Recurrence (FGR): Original LSTM paper
What's most important part of LSTM
Bi-directional LSTMs

Simple extension, often slightly improve performance (but don’t always make sense for task)
Comparing architectures

- GRUs seem competitive
- LSTM seems to be good tradeoff
- Bi-directional often offers slight improvement