Time will explain.

Jane Austen, Persuasion

Neural Language Models

RNN

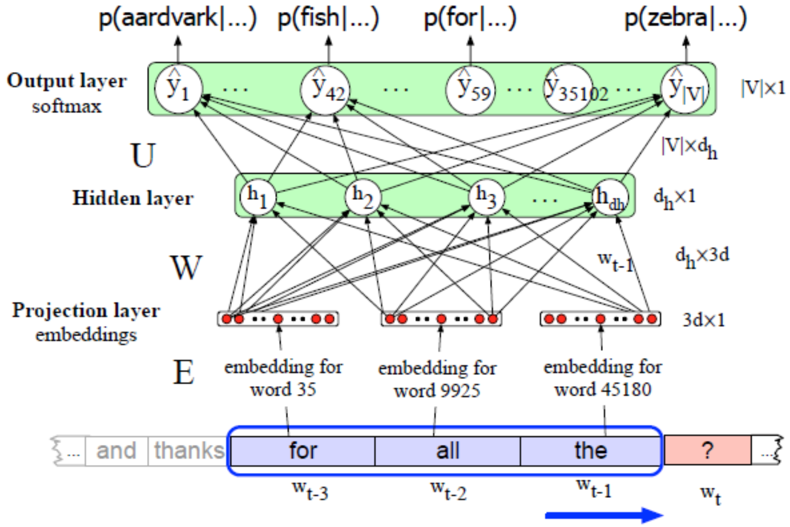
Dr. Uzair Ahmad

Program

- Previously
  - N-Gram Language Modeling
  - FFNN
- Neural Sequence Models
  - Design Criteria
  - Recurrent Neural Network
  - Capabilities
  - Limitations

Sequence Models

- N-Gram Language Models
  - “This morning I had Pizza for __________”
    - P(? | context words)
    - Limited History : Long-term dependencies
  - “Stop, do not let go”. Vs ”Do not stop, let go”.
    - Bag-of-words Representation
      - Counts do not preserve order
- Feedforward Neural Networks
  - “a b c d e” Vs “d e a b c”
    - Weights are tied to word positions
    - Cannot be shared



Jurafsky, SLP, Ch 9, P174

Sequence Models

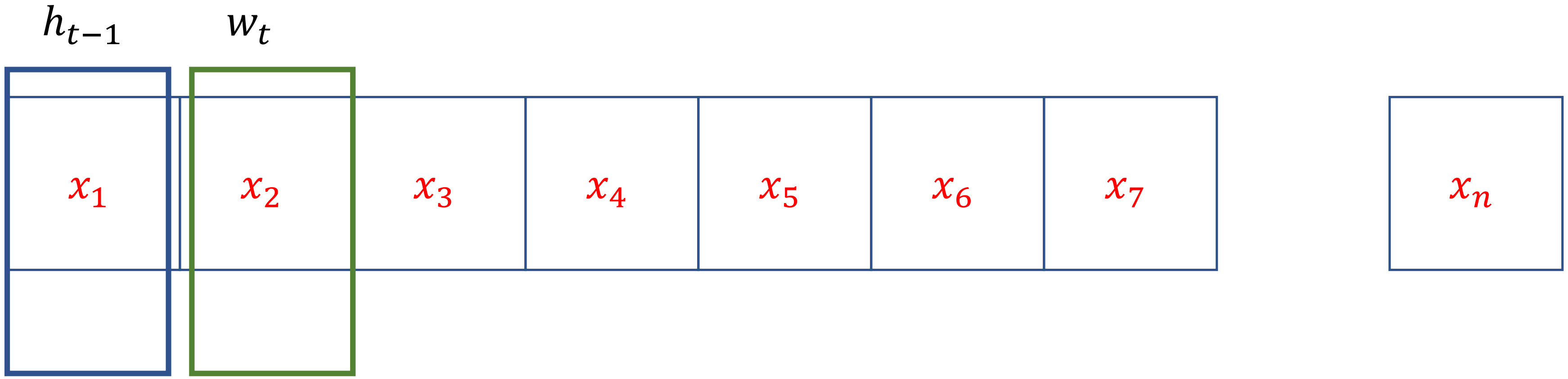
- Design criteria
  - Variable length sequences
  - Track long-term dependencies
  - Maintain information about “order of appearance”
  - Share parameters

Sequence Models

- Alternate to direct estimation of $p(x_{t+1}|x_1, x_2, x_3 \cdots x_t)$
  - Word prediction as discriminative task
  - $\mathrm{p}(x_m|h_m)$

- Reparameterization of $\mathrm{p}(w \mid \mu)$
  - $\mathrm{p}(x_m|h_m) = \frac{e^{(\beta_x \,.\, v_\mu)}}{\sum_{x' \in v} e^{(\beta_{x'} \,.\, v_\mu)}}$

Sequence Models: RNN

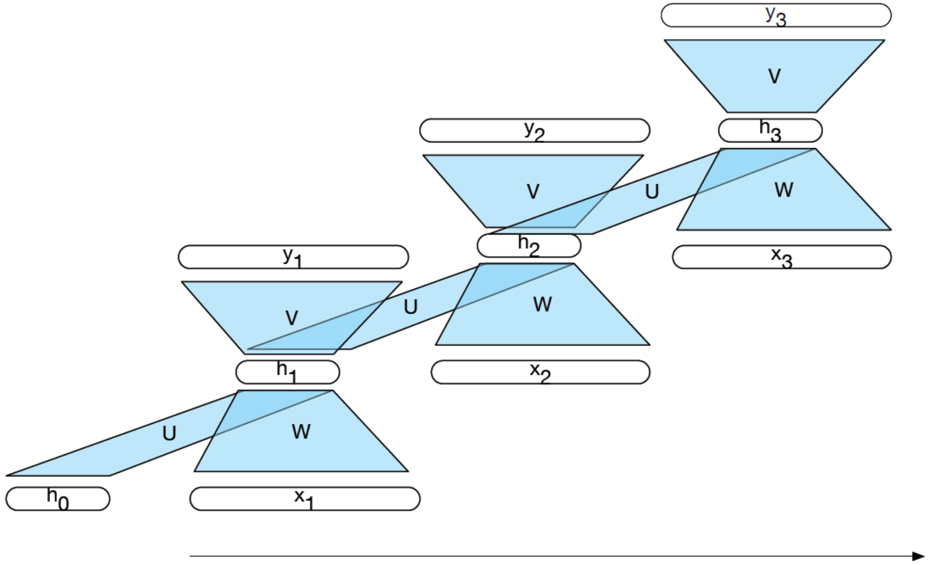


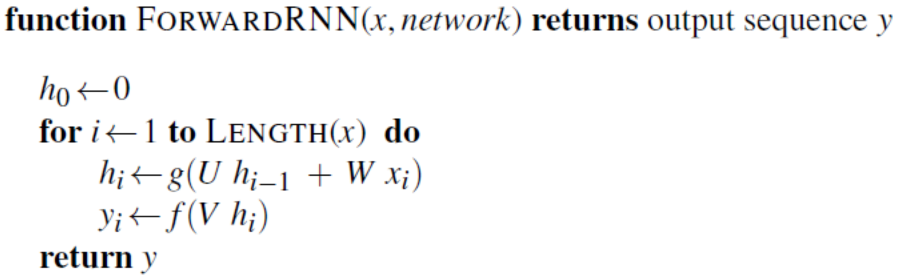
$$w_t \triangleq \phi x_t$$

$$h_t \ = RNN(w_t, h_{t-1})$$

$$p(x_{t+1}|x_1, x_2, x_3 \cdots x_t) = \frac{e^{(\beta_{x_{t+1}} \, . \, h_t)}}{\sum_{w' \in v} e^{(\beta_x \, . \, h_t)}}$$

Sequence Models: RNN



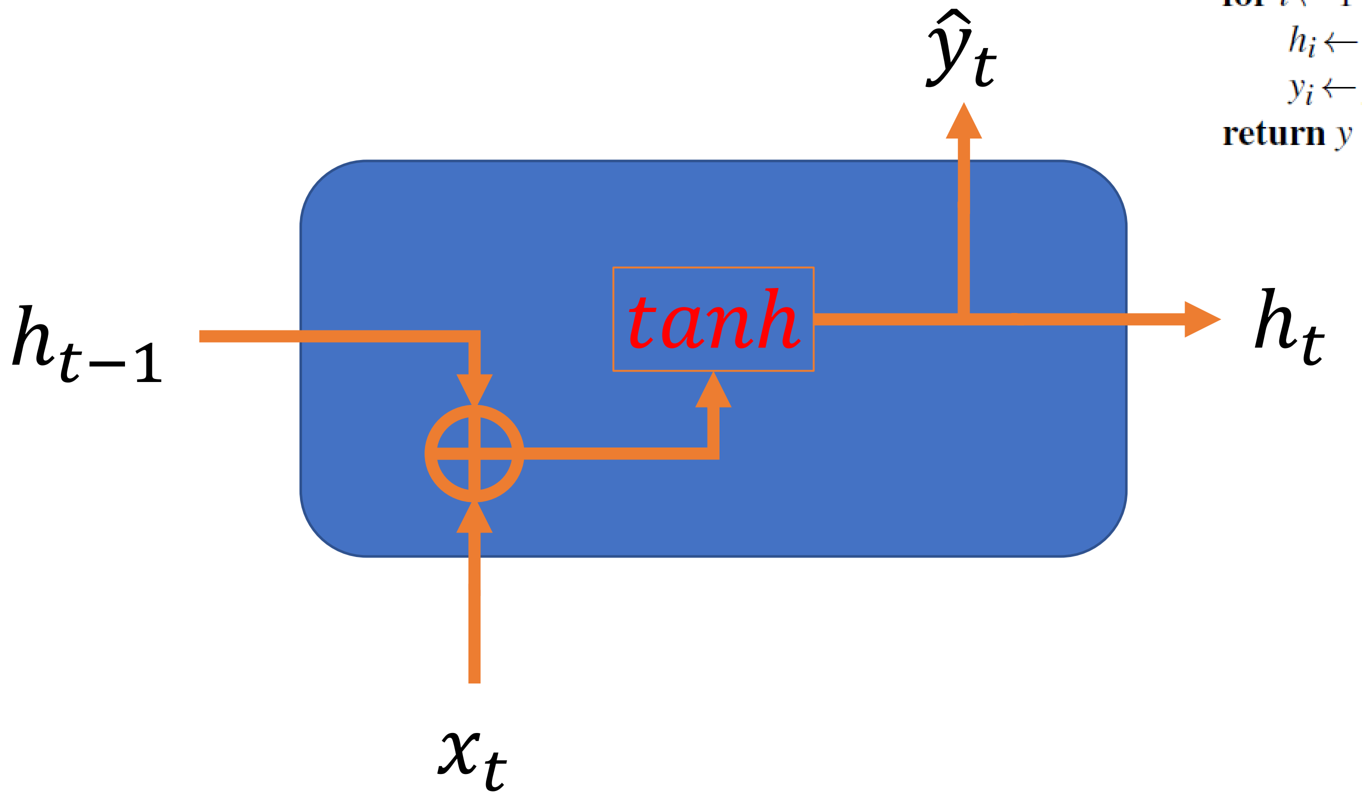


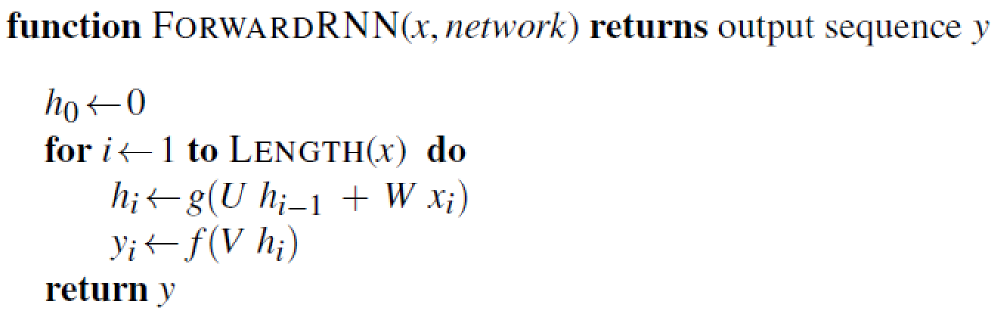
```
function FORWARDRNN($x$, $network$) returns output sequence $y$

  $h_0 \leftarrow 0$
  for $i \leftarrow 1$ to LENGTH($x$) do
      $h_i \leftarrow g(U\ h_{i-1}\ +\ W\ x_i)$
      $y_i \leftarrow f(V\ h_i)$
  return $y$
```

Jurafky, SLP, P179

Recurrent Neural Networks

$$h_t = f_W(h_{t-1}, x_t)$$





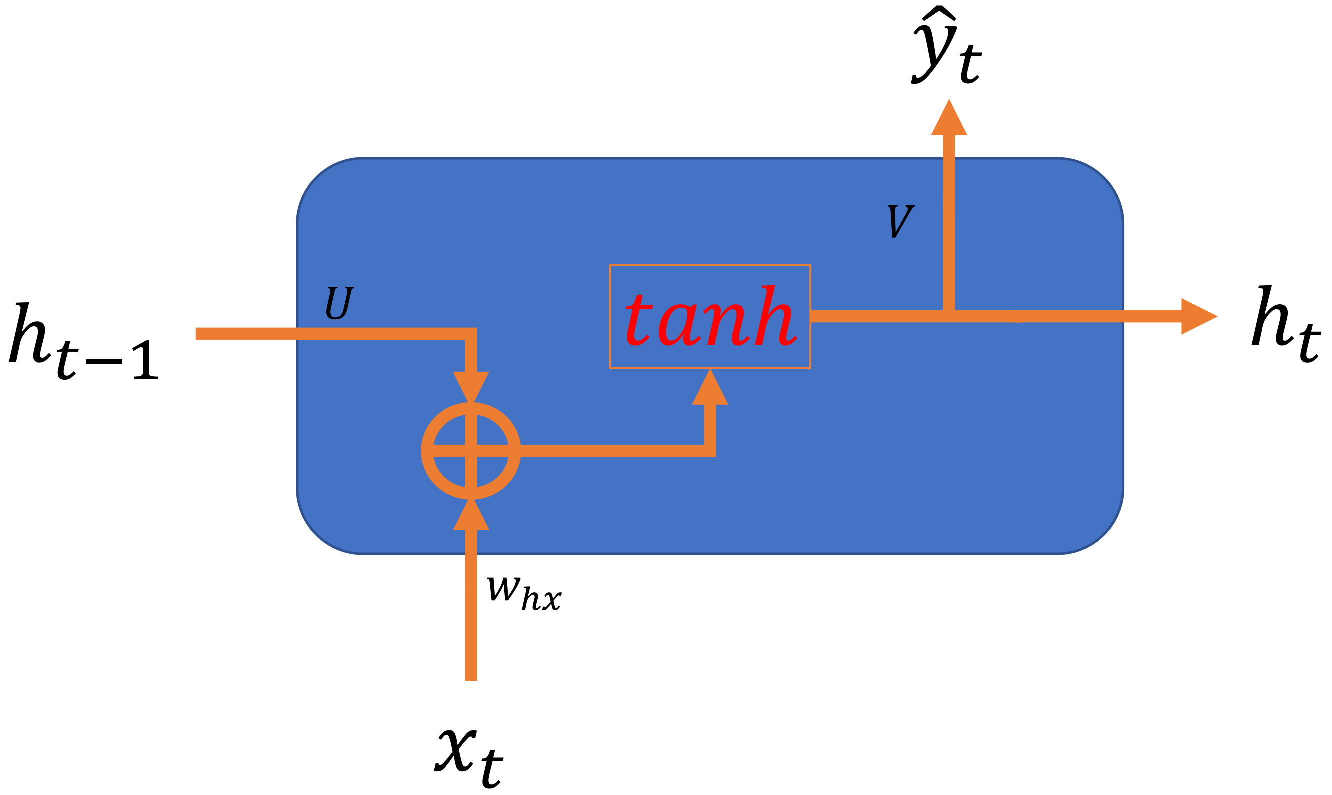
```
function FORWARDRNN($x$, $network$) returns output sequence $y$

  $h_0 \leftarrow 0$
  for $i \leftarrow 1$ to LENGTH($x$) do
      $h_i \leftarrow g(U\ h_{i-1} + W\ x_i)$
      $y_i \leftarrow f(V\ h_i)$
  return $y$
```

RNN

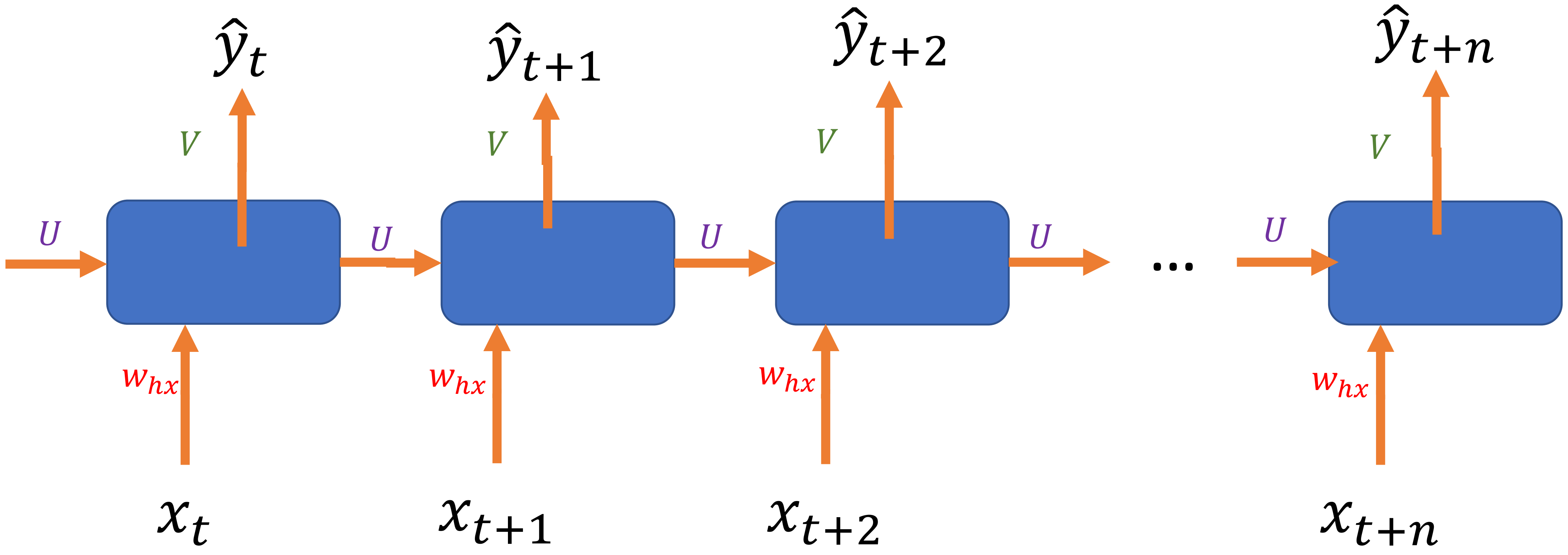
$x_t \triangleq \phi w_t$

$h_t = tanh(Uh_{t-1} + w_{hx}x_t)$

$\hat{y}_t = V h_t$

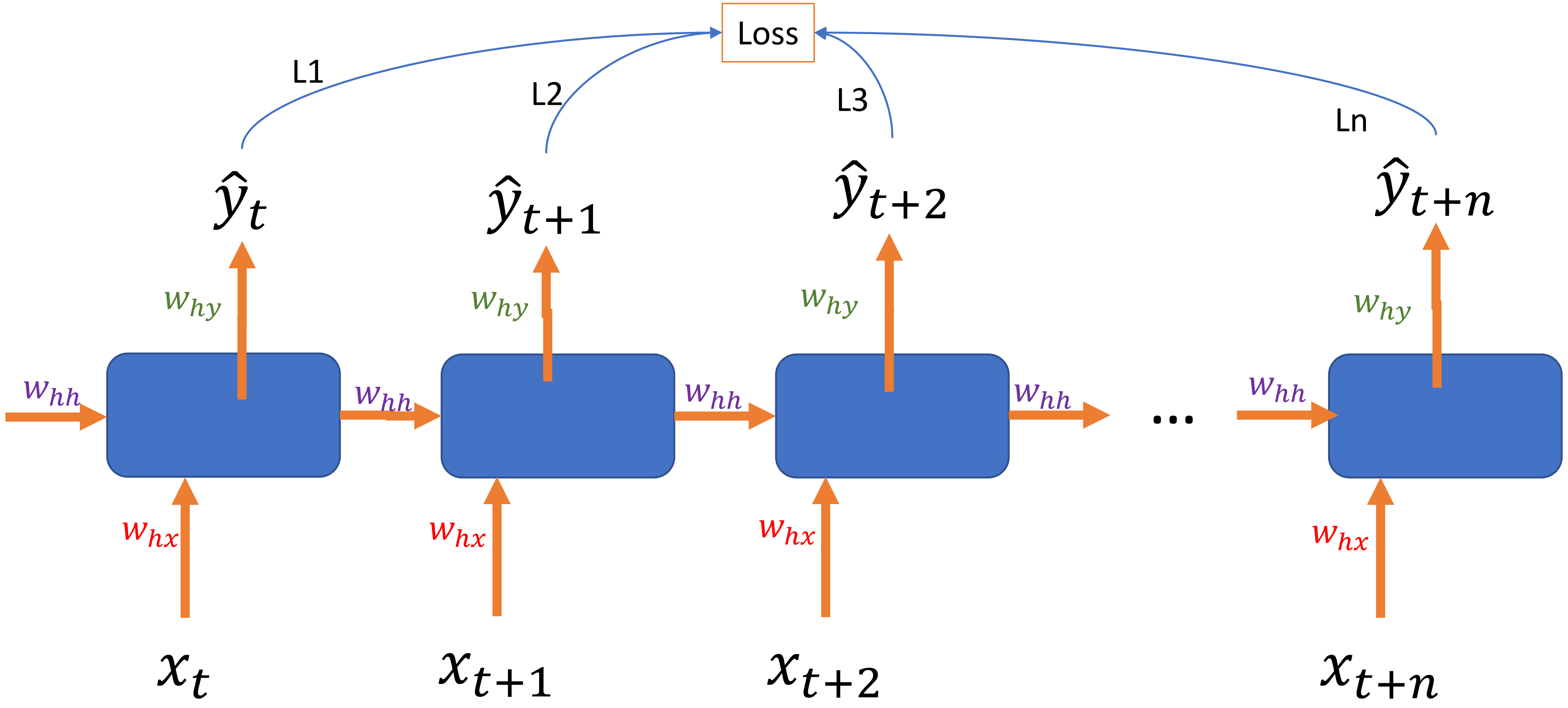


RNN

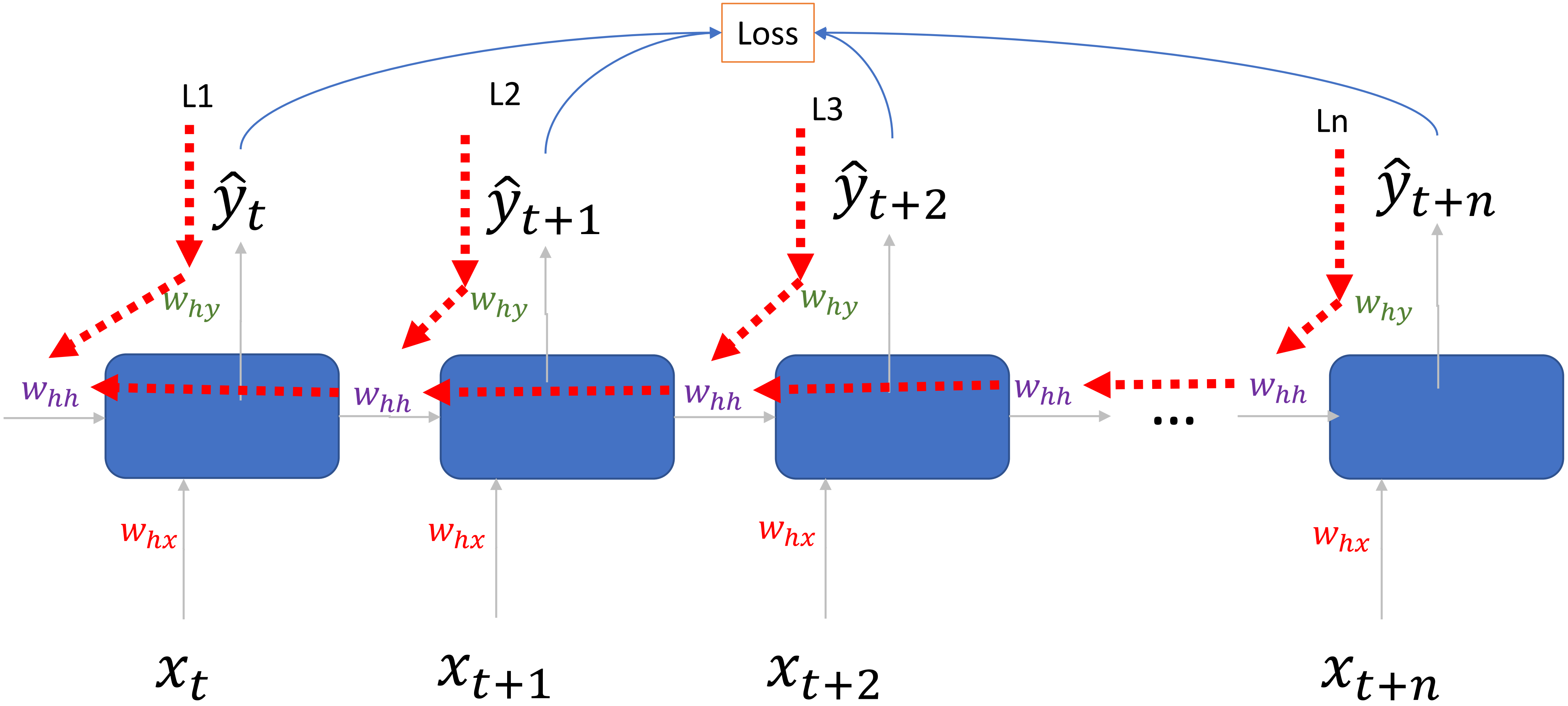


$\hat{y}_t$
$\hat{y}_{t+1}$
$\hat{y}_{t+2}$
$\hat{y}_{t+n}$
V
V
V
V
U
U
U
U
…
U
w_{hx}
w_{hx}
w_{hx}
w_{hx}
x_t
x_{t+1}
x_{t+2}
x_{t+n}

Training

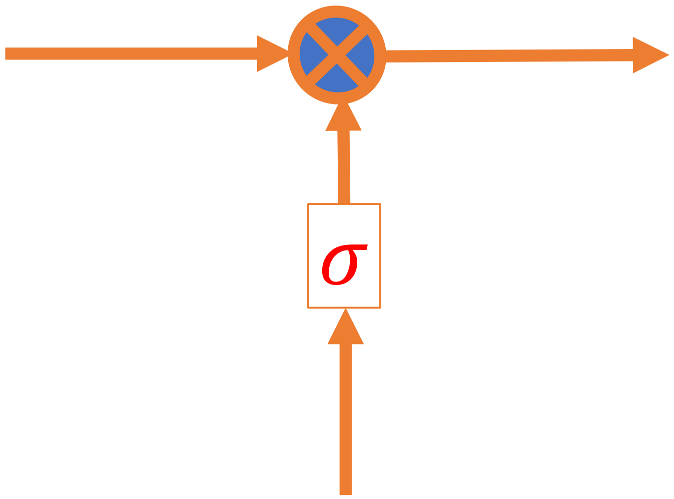

Loss
L1
L2
L3
Ln
$\hat{y}_t$
$\hat{y}_{t+1}$
$\hat{y}_{t+2}$
$\hat{y}_{t+n}$
w_{hy}
w_{hy}
w_{hy}
w_{hy}
w_{hh}
w_{hh}
w_{hh}
w_{hh}
…
w_{hh}
w_{hx}
w_{hx}
w_{hx}
w_{hx}
x_t
x_{t+1}
x_{t+2}
x_{t+n}

Recurrent Neural Networks



Recurrent Neural Networks

- Vanishing Gradients
  - Activation functions
    - Sigmoid
    - Tanh
    - Relu
  - Gated cells
    - GRU
    - LSTM



Evaluation of Language Models

- Extrinsic
- Intrinsic
  - Held-out data: $\ell(w) = \sum_{m=1}^{M} \log p(w_m | w_{m-1}, \dots, w_1)$
  - $Perplexity(w) = 2^{-\frac{\ell(w)}{M}}$

Summary

- Sequential Language Models

- RNN
  - Variable length computation graph