Recurrent Neural Networks (Continued)

Yu Meng
University of Virginia
yumeng5@virginia.edu

Sep 27, 2024

UNIVERSITY of VIRGINIA

Join at slido.com #2212 235

Overview of Course Contents

- Week 1: Logistics & Overview
- Week 2: N-gram Language Models
- Week 3: Word Senses, Semantics & Classic Word Representations
- Week 4: Word Embeddings
- Week 5: Sequence Modeling and Neural Language Models
- Week 6-7: Language Modeling with Transformers (Pretraining + Fine-tuning)
- Week 8: Large Language Models (LLMs) & In-context Learning
- Week 9-10: Knowledge in LLMs and Retrieval-Augmented Generation (RAG)
- Week 11: LLM Alignment
- Week 12: Language Agents
- Week 13: Recap + Future of NLP
- Week 15 (after Thanksgiving): Project Presentations

(Recap) Feedforward Network (FFN)

- Feedforward network (FFN) = multilayer network where the outputs from units in each layer are passed to units in the next higher layer
- FFNs are also called multi-layer perceptrons (MLPs)
- Model parameters in each layer in FFNs: a weight matrix $\boldsymbol{W}$ and a bias vector $\boldsymbol{b}$
  - Each layer has multiple hidden units
  - Recall: a single hidden unit has as a weight vector and a bias parameters
  - Weight matrix: combining the weight vector for each unit
  - Bias vector: combining the bias for each unit

(Recap) Example: 2-layer FFN

- Input: $\boldsymbol{x} = [x_1, x_2, \ldots, x_{n_0}]$
- Model parameters (weights & bias): $\boldsymbol{W} \in \mathbb{R}^{n_1 \times n_0}$, $\boldsymbol{U} \in \mathbb{R}^{n_2 \times n_1}$ & $\boldsymbol{b} \in \mathbb{R}^{n_1}$
- Forward computation:

First layer: $\boldsymbol{h} = \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$

Non-linear function (element-wise)

Second layer: $\boldsymbol{z} = \boldsymbol{U}\boldsymbol{h}$

Output: $\boldsymbol{y} = \mathrm{softmax}(\boldsymbol{z})$

Convert to probability distribution

$$= \left[\frac{\exp(z_1)}{\sum_{j=1}^{n_2} \exp(z_j)}, \ldots, \frac{\exp(z_{n_2})}{\sum_{j=1}^{n_2} \exp(z_j)} \right]$$

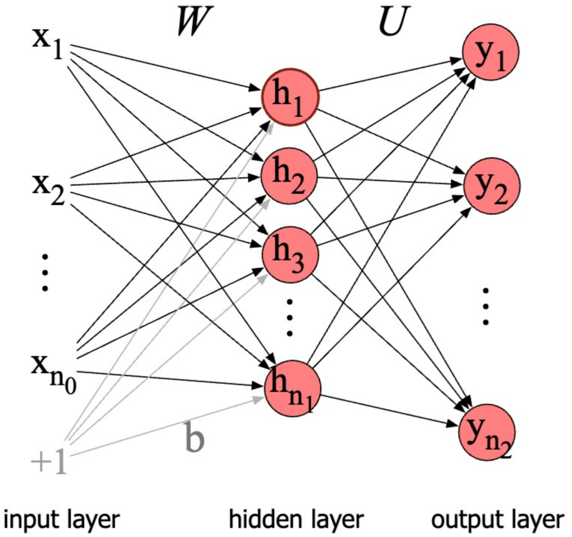


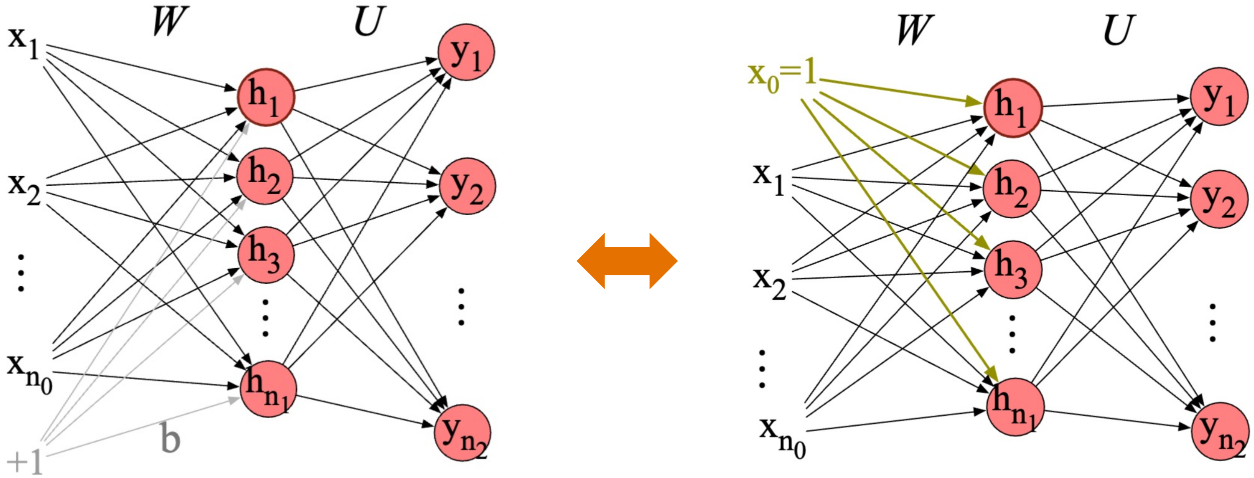
Figure source: https://web.stanford.edu/~jurafsky/slp3/7.pdf

UNIVERSITY of VIRGINIA

Join at
slido.com
#2212 235

(Recap) Replacing the Bias Term

- In neural network computations, we often use a slightly simplified notation that represents exactly the same function without an explicit bias node
- We assume the input will always have a dummy node $x_0 = 1$



$$\boldsymbol{h} = \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$$

$\boldsymbol{x} = [x_1, x_2, \ldots, x_{n_0}]$ $\boldsymbol{W} \in \mathbb{R}^{n_1 \times n_0}$

$$\boldsymbol{h} = \sigma(\boldsymbol{W}'\boldsymbol{x}')$$

$\boldsymbol{x}' = [1, x_1, x_2, \ldots, x_{n_0}]$ $\boldsymbol{W}' \in \mathbb{R}^{n_1 \times (n_0+1)}$

UNIVERSITY of VIRGINIA

Join at
slido.com
#2212 235

(Recap) Training Objective

- We'll need a **loss function** that models the distance between the model output and the gold/desired output
- The common loss function for classification tasks is **cross-entropy** (CE) loss

K-way classification (K classes): $$\mathcal{L}_{\mathrm{CE}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

Model output probability   Ground-truth probability

Usually a one-hot vector (one dimension is 1; others are 0): $\boldsymbol{y} = [0, \dots, 1, \dots, 0]$

$$\mathcal{L}_{\mathrm{CE}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\log \hat{y}_c = -\log \frac{\exp(z_c)}{\sum_{j=1}^{K} \exp(z_j)}$$

Also called "negative log likelihood (NLL) loss"

c is the ground-truth class

(Recap) Model Training

- Most optimization methods for DNNs are based on gradient descent
- First, randomly initialize model parameters
- In each optimization step, run two passes
  - **Forward pass**: evaluate the loss function given the input and current model parameters
  - **Backward pass:** update the parameters following the opposite direction of the gradient

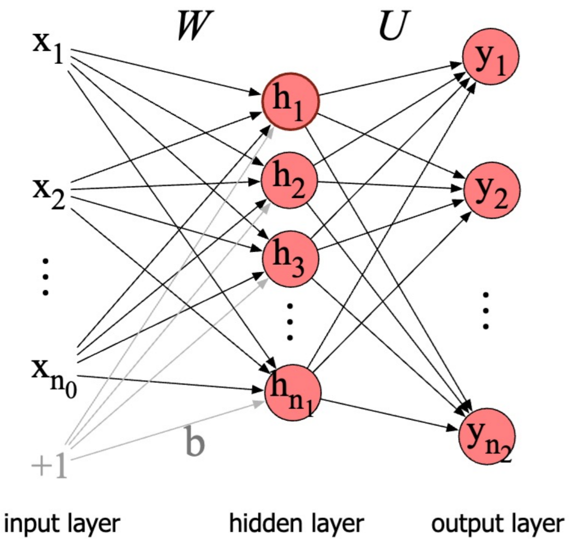
$$\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} - \eta \nabla_{\boldsymbol{w}} \mathcal{L}(\hat{\boldsymbol{y}}, \boldsymbol{y})$$

- Gradient computed via the chain rule $\nabla_{\boldsymbol{w}} \mathcal{L}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{w}}$

Gradient computation taken care of by deep learning libraries
(e.g., PyTorch)

(Recap) Simple Neural Language Model

Instantiate FFN as a neural language model



2-layer FFN

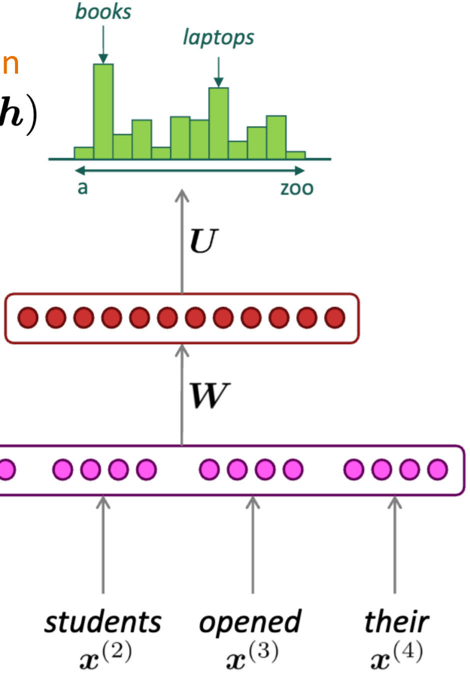
Output distribution

$$\boldsymbol{y} = \text{softmax}(\boldsymbol{U}\boldsymbol{h})$$

Hidden layer

$$\boldsymbol{h} = \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$$

Word embeddings



2-layer neural language model

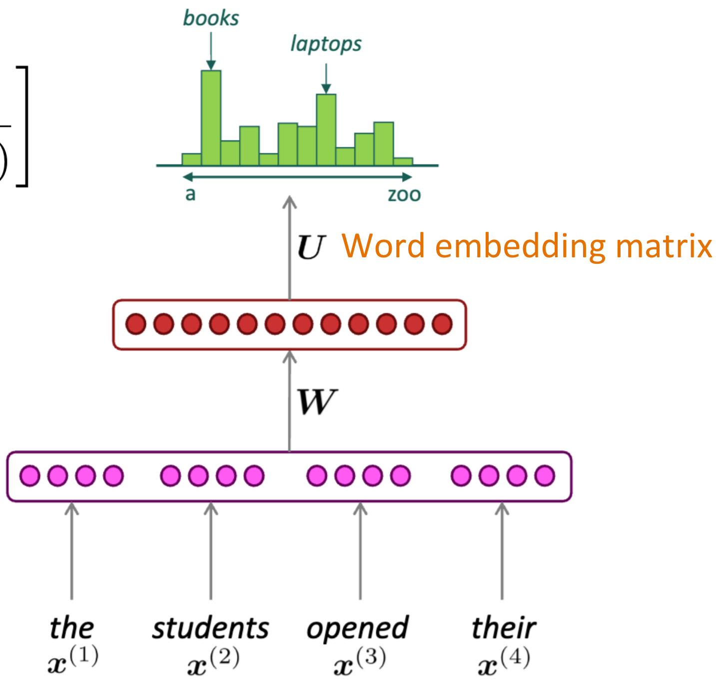
Figure source: https://web.stanford.edu/class/cs224n/slides/cs224n-spr2024-lecture05-rnnlm.pdf

Join at
slido.com
#2212 235

(Recap) Benefits of Neural Language Models

Output distribution

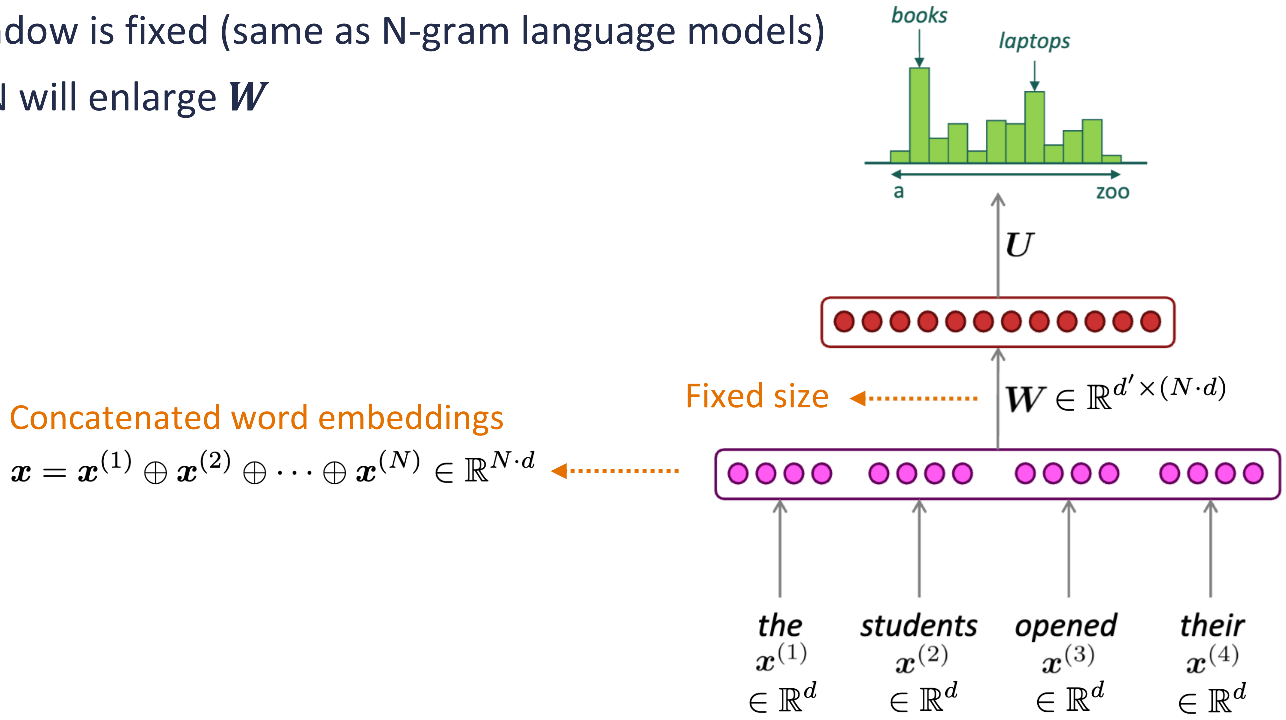
$$\boldsymbol{y} = \mathrm{softmax}(\boldsymbol{U}\boldsymbol{h}) = \left[\underbrace{\frac{\exp(\boldsymbol{u}_1 \cdot \boldsymbol{h})}{\sum_{j=1}^{|\mathcal{V}|} \exp(z_j)}, \ldots, \frac{\exp(\boldsymbol{u}_{|\mathcal{V}|} \cdot \boldsymbol{h})}{\sum_{j=1}^{\mathcal{V}} \exp(z_j)}}_{|\mathcal{V}|\text{-dimensions}} \right]$$



- Address sparsity issue:
  - Strictly positive probability on every token in the vocabulary
  - Semantically similar words tend to have similar probabilities

(Recap) Limitations of Neural Language Models

- Context window is fixed (same as N-gram language models)
- Increasing N will enlarge $\boldsymbol{W}$



Agenda

Join at
slido.com
#2212 235

- Feedforward Network (FFN)
- Simple Neural Language Model
- Recurrent Neural Network (RNN)
- RNN Limitations
- Advanced RNNs

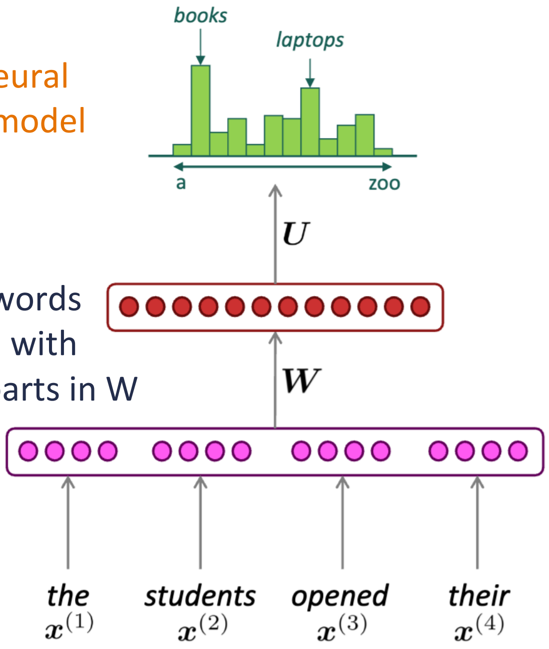
UNIVERSITY of VIRGINIA

Join at
slido.com
#2212 235

Recurrent Neural Network (RNN) Overview

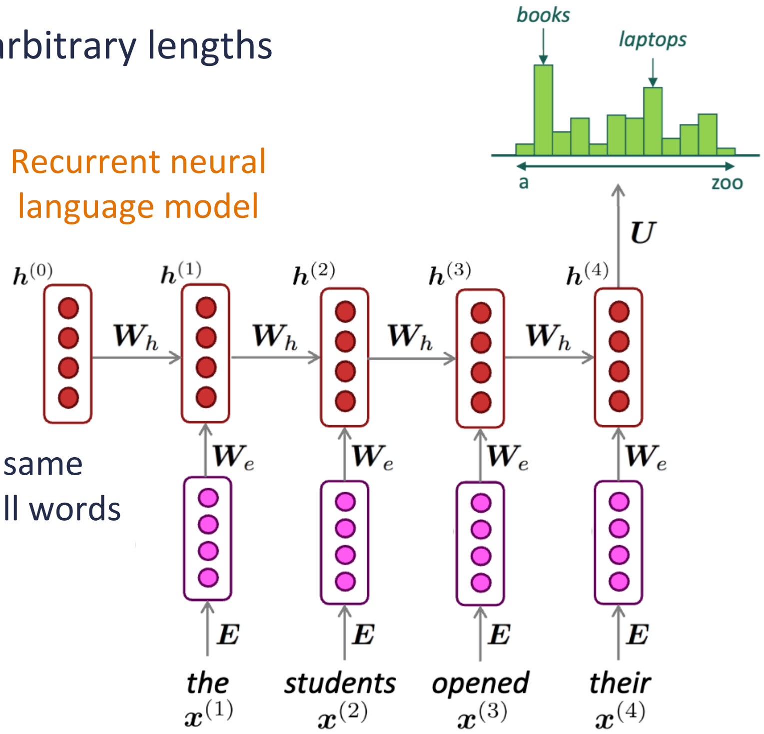
A neural language model that can process inputs of arbitrary lengths

Simple neural language model



Different words multiplied with different subparts in W

Recurrent neural language model



Reuse the same weights for all words

Figure source: https://web.stanford.edu/class/cs224n/slides/cs224n-spr2024-lecture05-rnnlm.pdf

UNIVERSITY of VIRGINIA

Join at
slido.com
#2212 235

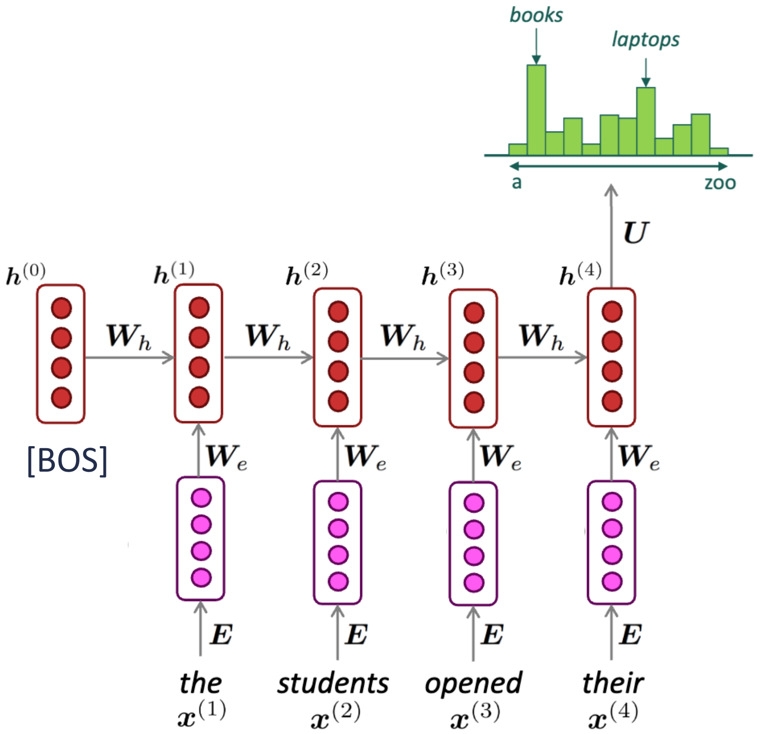
RNN Computation

- Hidden states in RNNs are computed based on
  - The hidden state at the previous step (memory)
  - The word embedding at the current step
- Parameters:
  - $\boldsymbol{W}_h$: weight matrix for the recurrent connection
  - $\boldsymbol{W}_e$: weight matrix for the input connection

$$\boldsymbol{h}^{(t)} = \sigma\left(\boldsymbol{W}_h \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_e \boldsymbol{x}^{(t)}\right)$$

Hidden states at the previous word (time step)

Word embedding of the current word (time step)



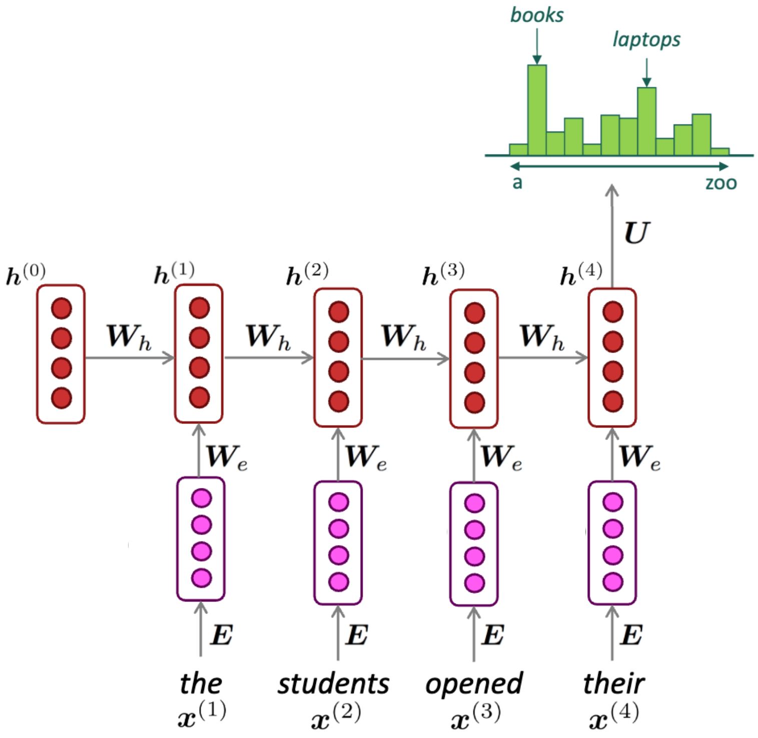
RNN Computation

- Input: $\boldsymbol{x} = [\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(N)}]$
- Initialize $\boldsymbol{h}^{(0)}$
- For each time step (word) in the input:
  - Compute hidden states:

$$\boldsymbol{h}^{(t)} = \sigma\left(\boldsymbol{W}_h \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_e \boldsymbol{x}^{(t)}\right)$$

  - Compute output:

$$\boldsymbol{y}^{(t)} = \text{softmax}\left(\boldsymbol{U}\boldsymbol{h}^{(t)}\right)$$



UNIVERSITY of VIRGINIA

Join at
slido.com
#2212 235

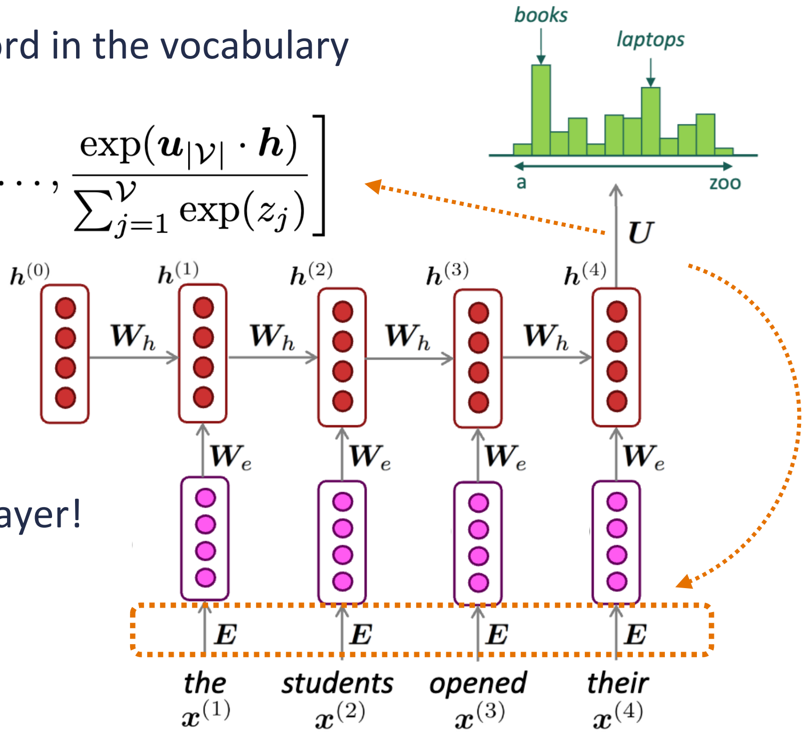
RNN Weight Tying

- Role of matrix $\boldsymbol{U}$: score the likelihood of each word in the vocabulary

$$\boldsymbol{y} = \mathrm{softmax}(\boldsymbol{U}\boldsymbol{h}) = \left[\frac{\exp(\boldsymbol{u}_1 \cdot \boldsymbol{h})}{\sum_{j=1}^{|\mathcal{V}|} \exp(z_j)}, \ldots, \frac{\exp(\boldsymbol{u}_{|\mathcal{V}|} \cdot \boldsymbol{h})}{\sum_{j=1}^{\mathcal{V}} \exp(z_j)} \right]$$

$$\boldsymbol{U} \in \mathbb{R}^{|\mathcal{V}| \times d}$$

Same dimensionality as the
word embedding matrix!



- Use the same input embeddings in the softmax layer!
- Weight tying benefits:
  - Improve learning efficiency & effectiveness
  - Reduce the number of parameters in the model

UNIVERSITY of VIRGINIA

Join at
slido.com
#2212 235

RNN for Language Modeling

- Recall that language modeling predicts the next word given previous words

$$p(\boldsymbol{x}) = p\left(x^{(1)}\right) p\left(x^{(2)} | x^{(1)}\right) \cdots p\left(x^{(n)} | x^{(1)}, \ldots, x^{(n-1)}\right) = \prod_{t=1}^{n} p\left(x^{(t)} | x^{(1)}, \ldots, x^{(t-1)}\right)$$

- How to use RNNs to represent $p\left(x^{(t)} | x^{(1)}, \ldots, x^{(t-1)}\right)$?

Output probability at (t-1) step: $\boldsymbol{y}^{(t-1)} = \mathrm{softmax}\left(\boldsymbol{U}\boldsymbol{h}^{(t-1)}\right) := f\left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(t-2)}, \boldsymbol{x}^{(t-1)}\right)$

$\boldsymbol{h}^{(t-1)}$ is a function of $\boldsymbol{h}^{(t-2)}$ and $\boldsymbol{x}^{(t-1)}$: $\boldsymbol{h}^{(t-1)} = \sigma\left(\boldsymbol{W}_h \boldsymbol{h}^{(t-2)} + \boldsymbol{W}_e \boldsymbol{x}^{(t-1)}\right) := g\left(\boldsymbol{h}^{(t-2)}, \boldsymbol{x}^{(t-1)}\right)$

$\boldsymbol{h}^{(t-2)}$ is a function of $\boldsymbol{h}^{(t-3)}$ and $\boldsymbol{x}^{(t-2)}$: $\boldsymbol{h}^{(t-2)} = \sigma\left(\boldsymbol{W}_h \boldsymbol{h}^{(t-3)} + \boldsymbol{W}_e \boldsymbol{x}^{(t-2)}\right) := g\left(\boldsymbol{h}^{(t-3)}, \boldsymbol{x}^{(t-2)}\right)$

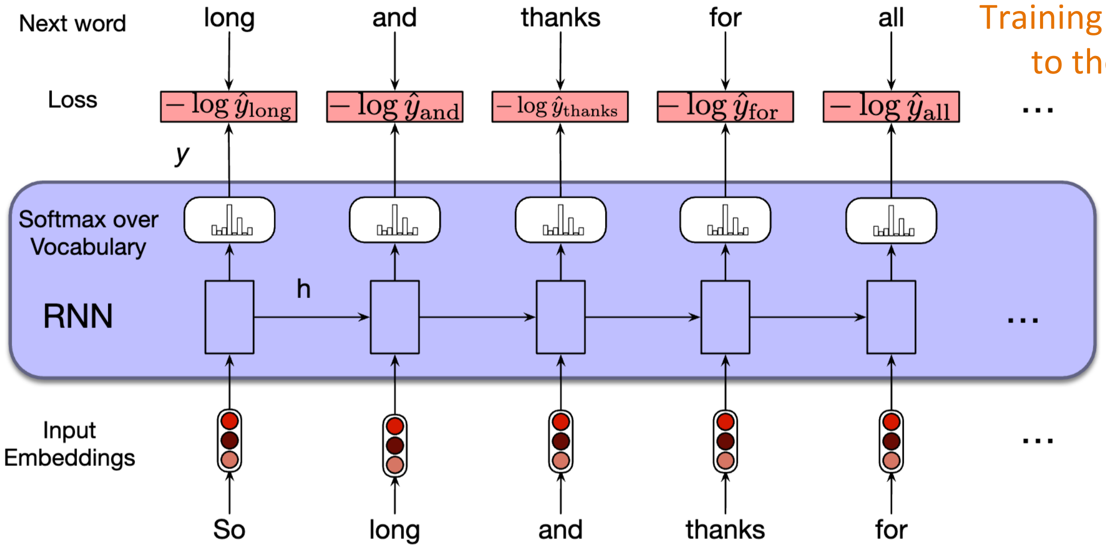
$\vdots \qquad \vdots$

$\boldsymbol{h}^{(1)}$ is a function of $\boldsymbol{h}^{(0)}$ and $\boldsymbol{x}^{(1)}$: $\boldsymbol{h}^{(1)} = \sigma\left(\boldsymbol{W}_h \boldsymbol{h}^{(0)} + \boldsymbol{W}_e \boldsymbol{x}^{(1)}\right) := g\left(\boldsymbol{h}^{(0)}, \boldsymbol{x}^{(1)}\right)$

RNN Language Model Training

Train the output probability at each time step to predict the next word

$$\mathcal{L}_{\mathrm{LM}}(\boldsymbol{x}) = \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}_{\mathrm{CE}}\left(\hat{\boldsymbol{y}}^{(t)}, \boldsymbol{y}^{(t)}\right) = \frac{1}{n}\sum_{t=1}^{n} -\log \hat{y}^{(t)}_{x^{(t)}} = \frac{1}{n}\sum_{t=1}^{n} -\log \frac{\exp\left(x^{(t)}\right)}{\sum_{w' \in \mathcal{V}} \exp(w')}$$



Training target is the input shifted to the left by one time step

Figure source: https://web.stanford.edu/~jurafsky/slp3/8.pdf

RNN for Text Generation

- Input [BOS] (beginning-of-sequence) token to the model
- Sample a word from the softmax distribution at the first time step
- Use the word embedding of that first word as the input at the next time step
- Repeat until the [EOS] (end-of-sequence) token is generated

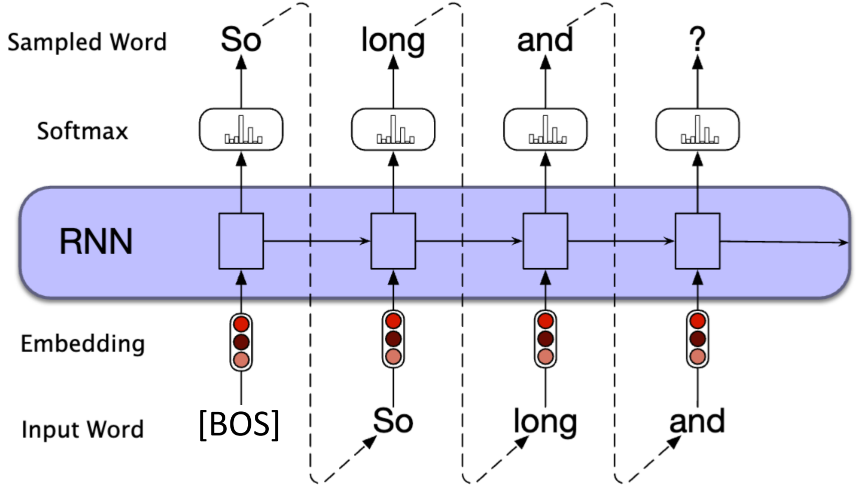


Figure source: https://web.stanford.edu/~jurafsky/slp3/8.pdf

Agenda

- Feedforward Network (FFN)
- Simple Neural Language Model
- Recurrent Neural Network (RNN)
- RNN Limitations
- Advanced RNNs

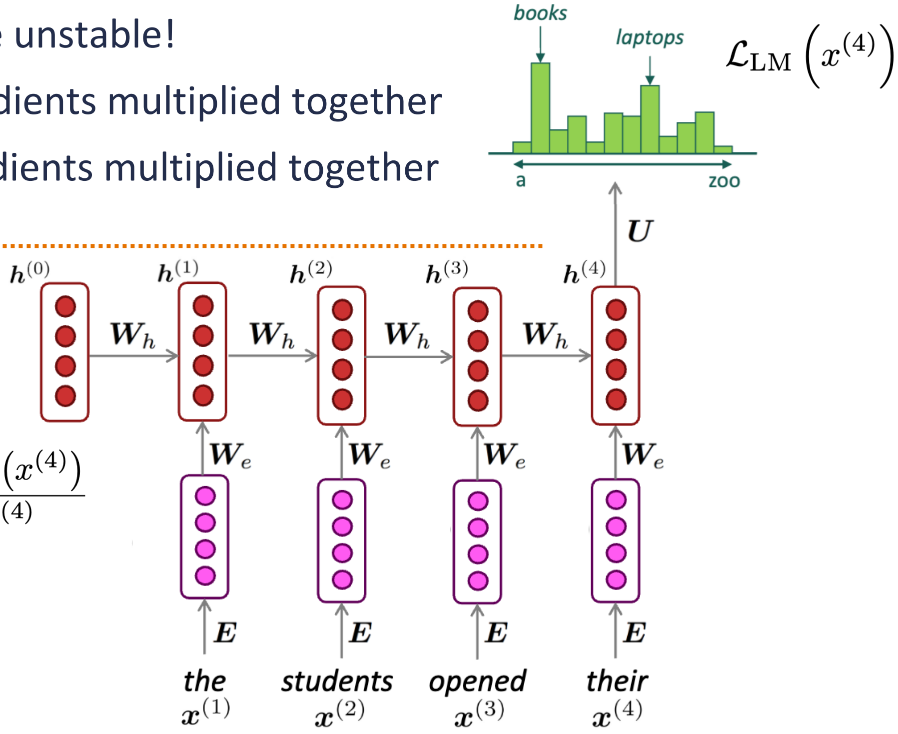
Vanishing & Exploding Gradient

- Gradient signal from far away can be unstable!
- Vanishing gradient = many small gradients multiplied together
- Exploding gradient = many large gradients multiplied together

Gradient backpropagation

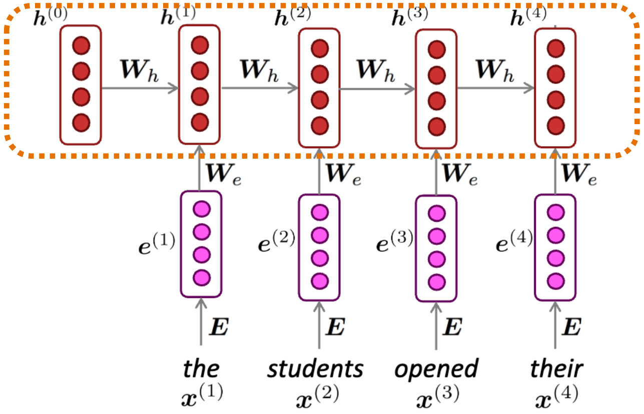
Lots of gradient multiplications!

$$\frac{\partial \mathcal{L}_{\mathrm{LM}}\left(x^{(4)}\right)}{\partial \boldsymbol{h}^{(0)}} = \boxed{\frac{\partial \boldsymbol{h}^{(1)}}{\partial \boldsymbol{h}^{(0)}} \frac{\partial \boldsymbol{h}^{(2)}}{\partial \boldsymbol{h}^{(1)}} \frac{\partial \boldsymbol{h}^{(3)}}{\partial \boldsymbol{h}^{(2)}} \frac{\partial \boldsymbol{h}^{(4)}}{\partial \boldsymbol{h}^{(3)}}} \frac{\partial \mathcal{L}_{\mathrm{LM}}\left(x^{(4)}\right)}{\partial \boldsymbol{h}^{(4)}}$$



Difficulty in Capturing Long-Term Dependencies

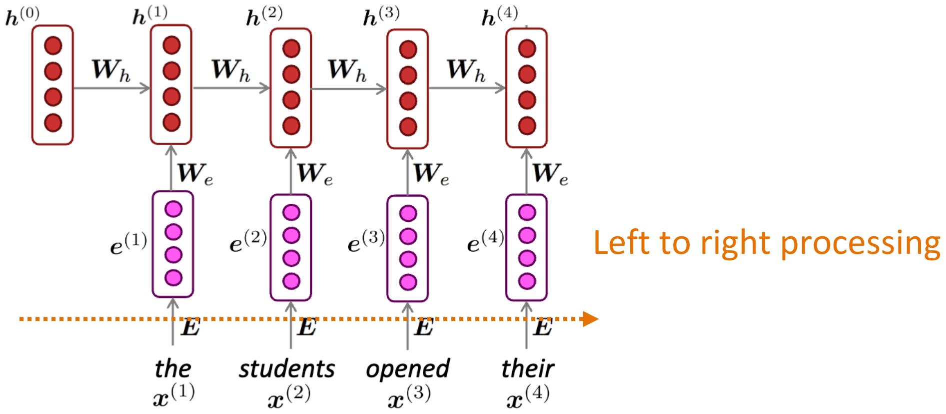
- RNNs are theoretically capable of remembering information over arbitrary lengths of input, but they struggle in practice with long-term dependencies
- RNNs use a fixed-size hidden state to encode an entire sequence of variable length; the hidden state is required to compress a lot of information
- RNNs might give more weight to the most recent inputs and may ignore or "forget" important information at the beginning of the sentence while processing the end



Fixed size hidden states!

Lack of Bidirectionality

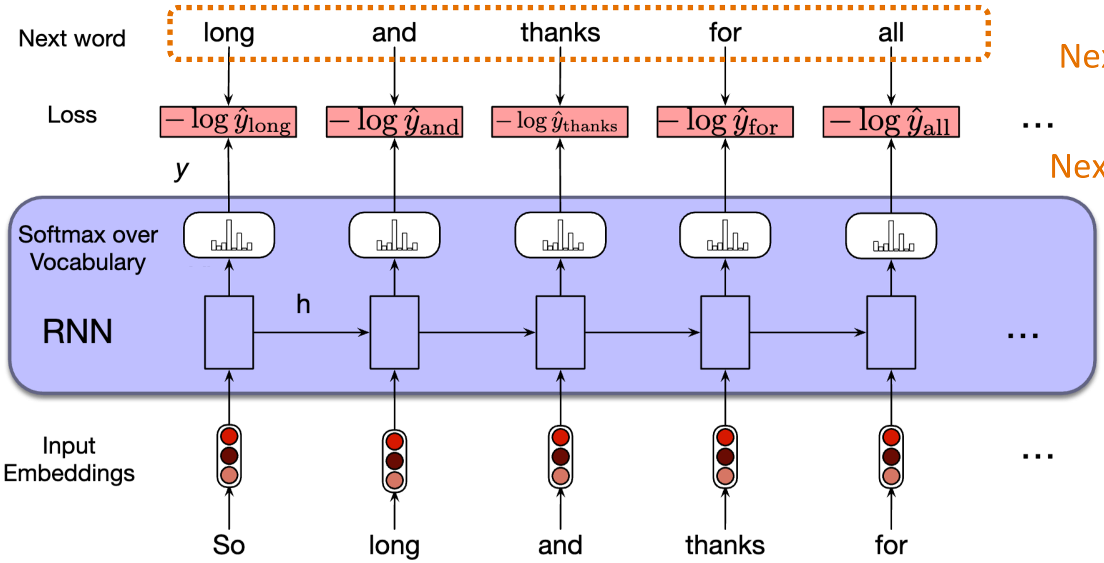
- RNNs process the input sequence step by step from the beginning to the end (left to right for English)
- At each time step, the hidden state only has access to the information from the past without being able to leverage future contexts
- Example: "The bank is on the river" the word "bank" can be correctly disambiguated only if the model has access to the word "river" later in the sentence



Join at
slido.com
#2212 235

Exposure Bias

- **Teacher forcing/exposure bias**: during RNN training, the model always receives the **correct** next word from the training data as input for the next step
- When the model has to predict sequences on its own, it may perform poorly if it hasn't learned how to correct its own mistakes



During training:
Next word = actual next word

During generation:
Next word = model's prediction

Join at
slido.com
#2212 235

Agenda

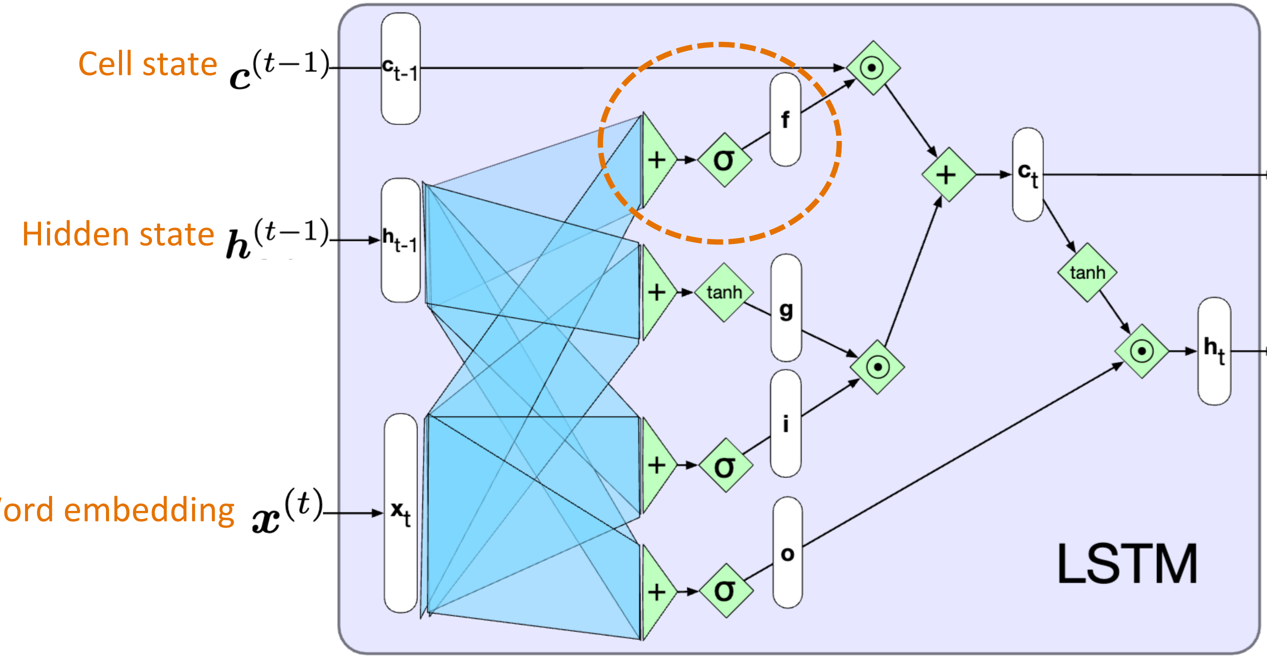
- Feedforward Network (FFN)
- Simple Neural Language Model
- Recurrent Neural Network (RNN)
- RNN Limitations
- Advanced RNNs

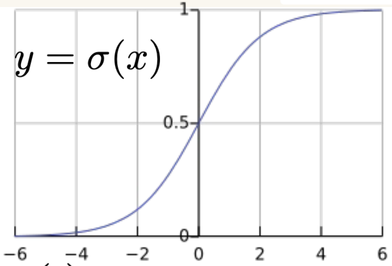
Long Short-Term Memory (LSTM)

- Challenge in RNNs: information encoded in hidden states tends to be local; distant information gets lost
- LSTM design intuition:
  - Remove information no longer needed from the context
  - Add information likely to be needed for future time steps
- Inputs at each time step:
  - Word embedding of the current word
  - Hidden state from the previous time step
  - **Memory/cell state**
- Three gates:
  - Forget gate
  - Add/input gate
  - Output gate

LSTM Computation (Forget Gate)

$$\boldsymbol{f}^{(t)} = \sigma\left(\boldsymbol{U}_f \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_f \boldsymbol{x}^{(t)}\right)$$





Decides what information to discard from the cell state

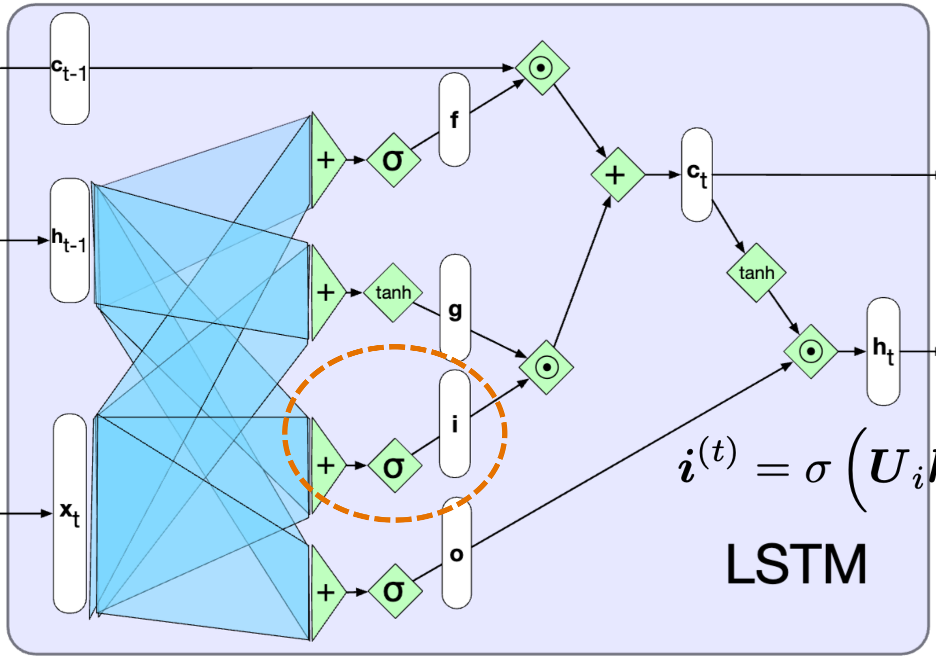
Figure source: https://web.stanford.edu/~jurafsky/slp3/8.pdf

LSTM Computation (Add/Input Gate)

Cell state $\boldsymbol{c}^{(t-1)}$

Hidden state $\boldsymbol{h}^{(t-1)}$

Word embedding $\boldsymbol{x}^{(t)}$



$\boldsymbol{c}^{(t)}$

$\boldsymbol{h}^{(t)}$

Decides what new information to store to the cell state

$$\boldsymbol{i}^{(t)} = \sigma\left(\boldsymbol{U}_i \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_i \boldsymbol{x}^{(t)}\right)$$

Figure source: https://web.stanford.edu/~jurafsky/slp3/8.pdf

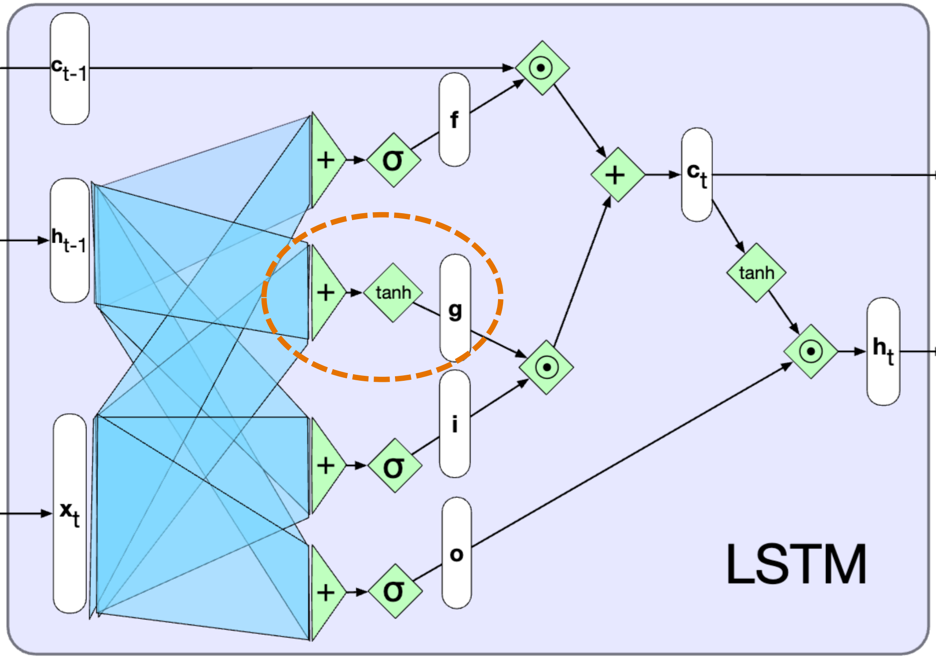
LSTM Computation (Candidate Cell State)

$$\boldsymbol{g}^{(t)} = \tanh\left(\boldsymbol{U}_g \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_g \boldsymbol{x}^{(t)}\right)$$

Cell state $\boldsymbol{c}^{(t-1)}$

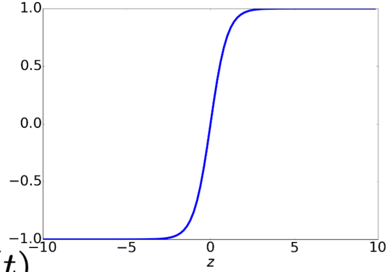
Hidden state $\boldsymbol{h}^{(t-1)}$

Word embedding $\boldsymbol{x}^{(t)}$



$\boldsymbol{c}^{(t)}$

$\boldsymbol{h}^{(t)}$



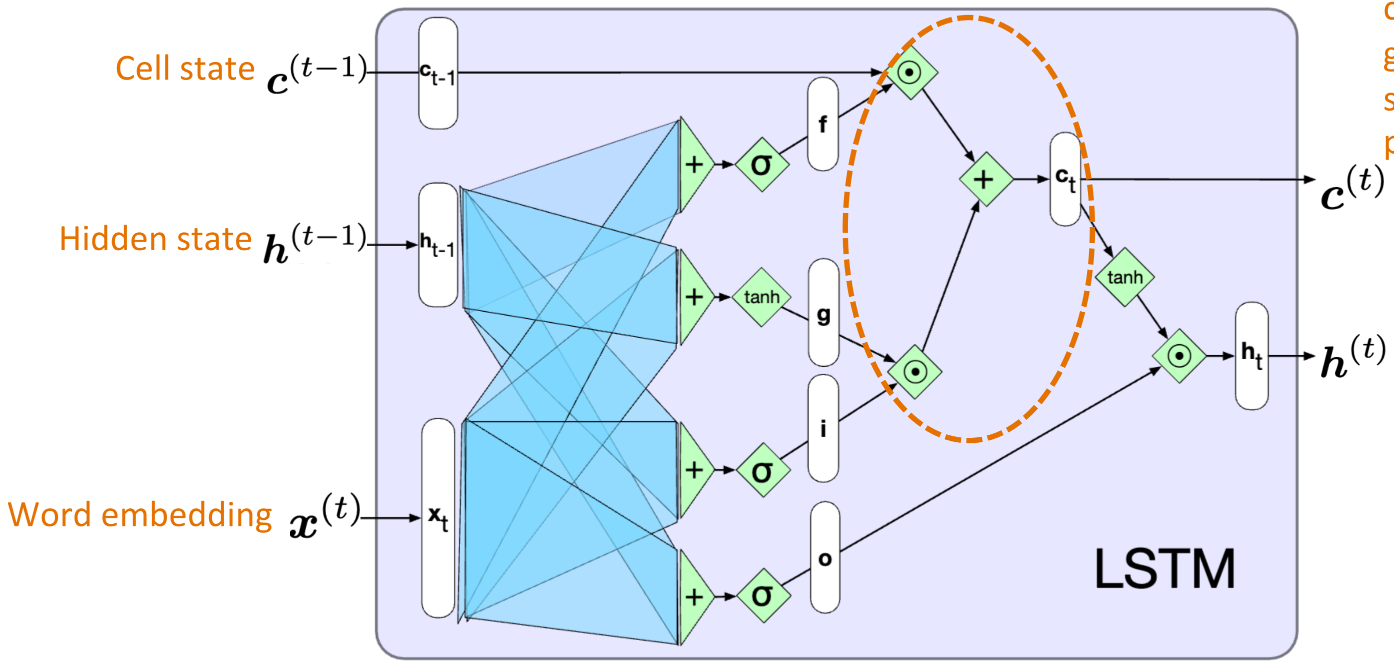
$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

Compute information needed from the previous hidden state and current inputs

Figure source: https://web.stanford.edu/~jurafsky/slp3/8.pdf

LSTM Computation (Cell State Update)

$$\boldsymbol{c}^{(t)} = \boldsymbol{i}^{(t)} \odot \boldsymbol{g}^{(t)} + \boldsymbol{f}^{(t)} \odot \boldsymbol{c}^{(t-1)}$$



Cell state updated by combining the input gate, candidate cell state, forget gate & previous cell state

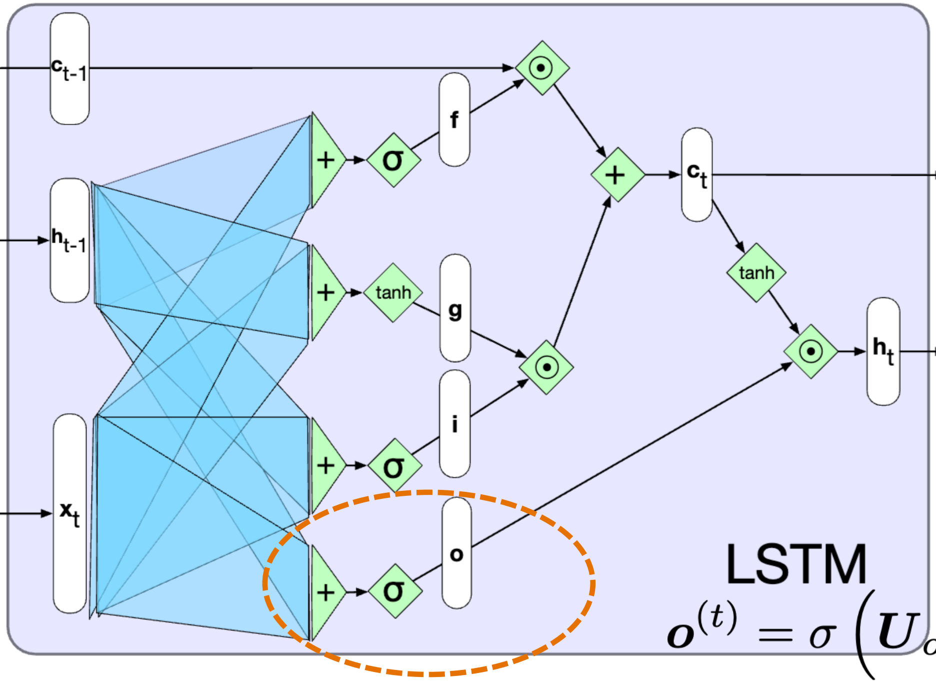
Figure source: https://web.stanford.edu/~jurafsky/slp3/8.pdf

LSTM Computation (Output Gate)

Cell state $\boldsymbol{c}^{(t-1)}$

Hidden state $\boldsymbol{h}^{(t-1)}$

Word embedding $\boldsymbol{x}^{(t)}$



$\boldsymbol{c}^{(t)}$

$\boldsymbol{h}^{(t)}$

Decides what parts of the cell state will be output

$$\boldsymbol{o}^{(t)} = \sigma\left(\boldsymbol{U}_o \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_o \boldsymbol{x}^{(t)}\right)$$

Figure source: https://web.stanford.edu/~jurafsky/slp3/8.pdf

UNIVERSITY of VIRGINIA

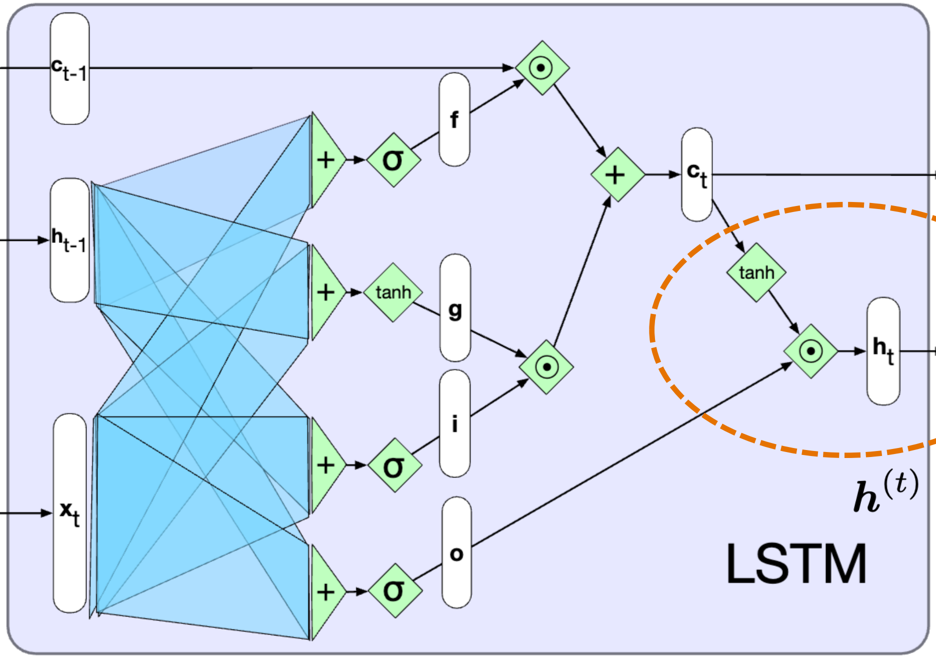
Join at slido.com #2212 235

LSTM Computation (Hidden State Update)

Cell state $\boldsymbol{c}^{(t-1)}$

Hidden state $\boldsymbol{h}^{(t-1)}$

Word embedding $\boldsymbol{x}^{(t)}$



$\boldsymbol{c}^{(t)}$

$\boldsymbol{h}^{(t)}$

$$\boldsymbol{h}^{(t)} = \boldsymbol{o}^{(t)} \odot \tanh\left(\boldsymbol{c}^{(t)}\right)$$

Hidden state updated using the output gate & the updated cell state

Figure source: https://web.stanford.edu/~jurafsky/slp3/8.pdf

Bidirectional RNNs

- Separate models are trained in the forward and backward directions
- Hidden states from both RNNs are concatenated as the final representations

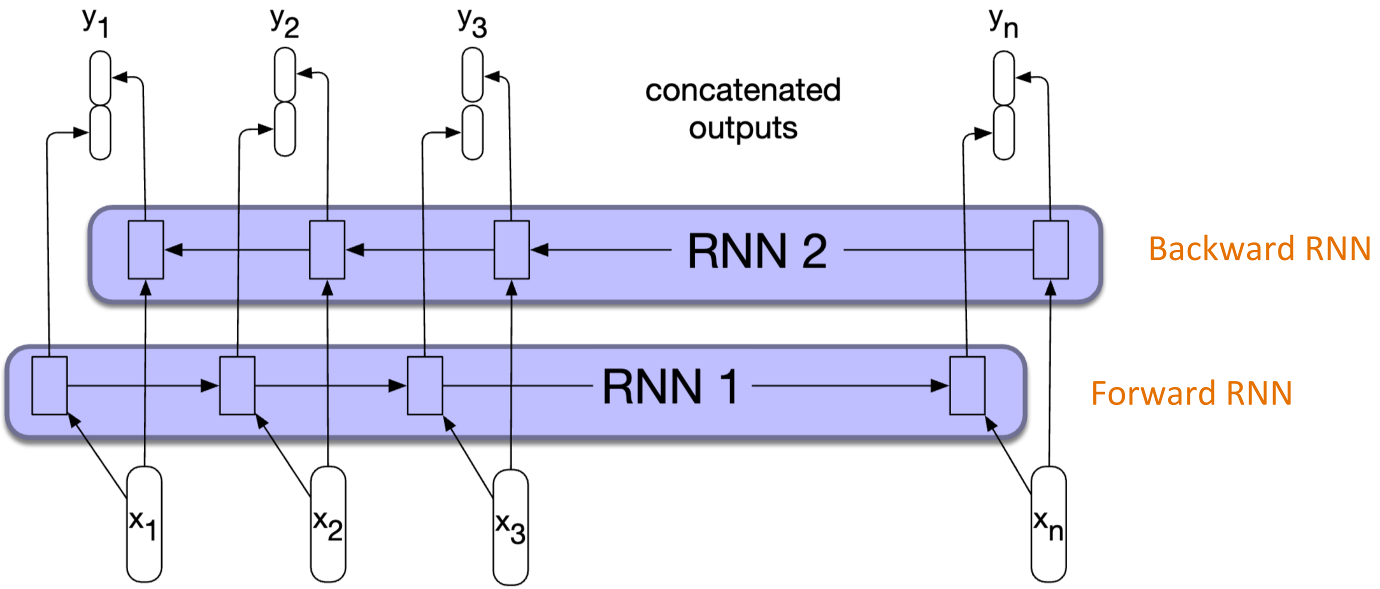


Figure source: https://web.stanford.edu/~jurafsky/slp3/8.pdf

Deep RNNs

- We can stack multiple RNN layers to build deep RNNs
- The output of a lower level serves as the input to higher levels
- The output of the last layer is used as the final output

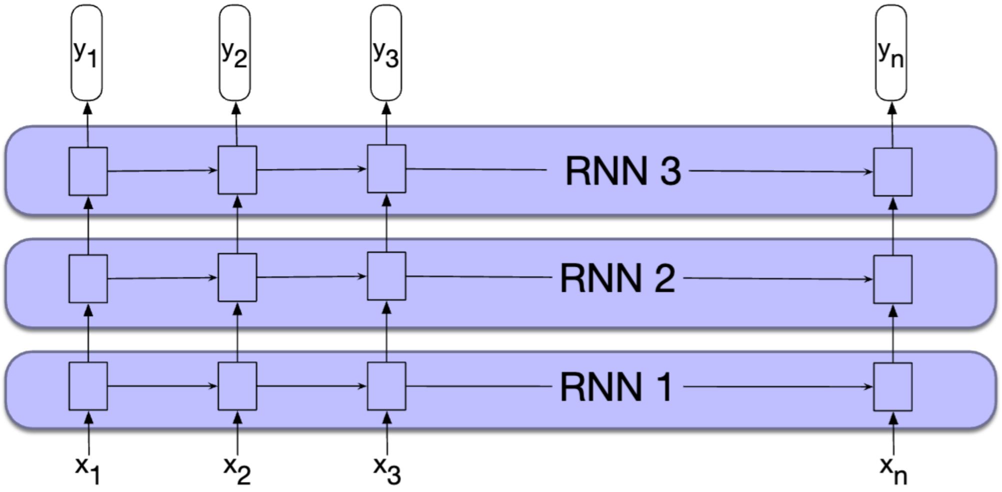


Figure source: https://web.stanford.edu/~jurafsky/slp3/8.pdf

Summary: Sequence Modeling

- Sequence modeling goals:
  - Learn context-dependent representations
  - Capture long-range dependencies
  - Handle complex relationships among large text units
- Use deep learning architectures to understand, process, and generate text sequences
- Why DNNs?
  - The multi-layer structure in DNNs mirrors the hierarchical structures in language
  - DNNs learn multiple levels of semantics across layers: low-level patterns (e.g., relations between words) in lower layers & high-level patterns (e.g., sentence meanings) in higher layers

Join at
slido.com
#2212 235

Summary: Neural Language Models

- Address the sparsity issue in N-gram language models by computing the output distribution based on distributed representations (with semantic information)
- Simple neural language models based on feedforward networks suffer from the fixed context window issue
  - Can only model a fixed number of words (similar to N-gram assumption)
  - Increasing the context window requires adding more model parameters

Summary: Recurrent Neural Networks

- General idea: Use the same set of model weights to process all input words
- RNNs as language models
  - Theoretically able to process infinitely long sequences
  - Practically can only keep track of recent contexts
- Training issues: vanishing & exploding gradients
- LSTM is a prominent RNN variant to keep track of both long-term and short-term memories via multiple gates

Thank You!

Yu Meng
University of Virginia
yumeng5@virginia.edu