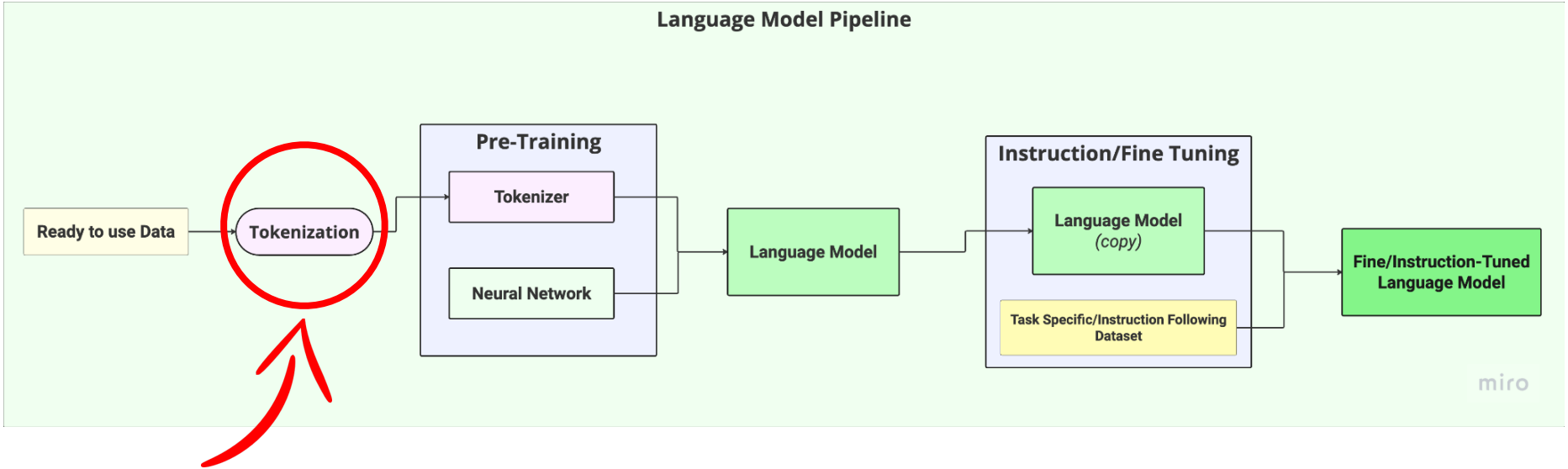
NLP4Web
Practice Session 8

Neural Language Models: RNN to LSTM

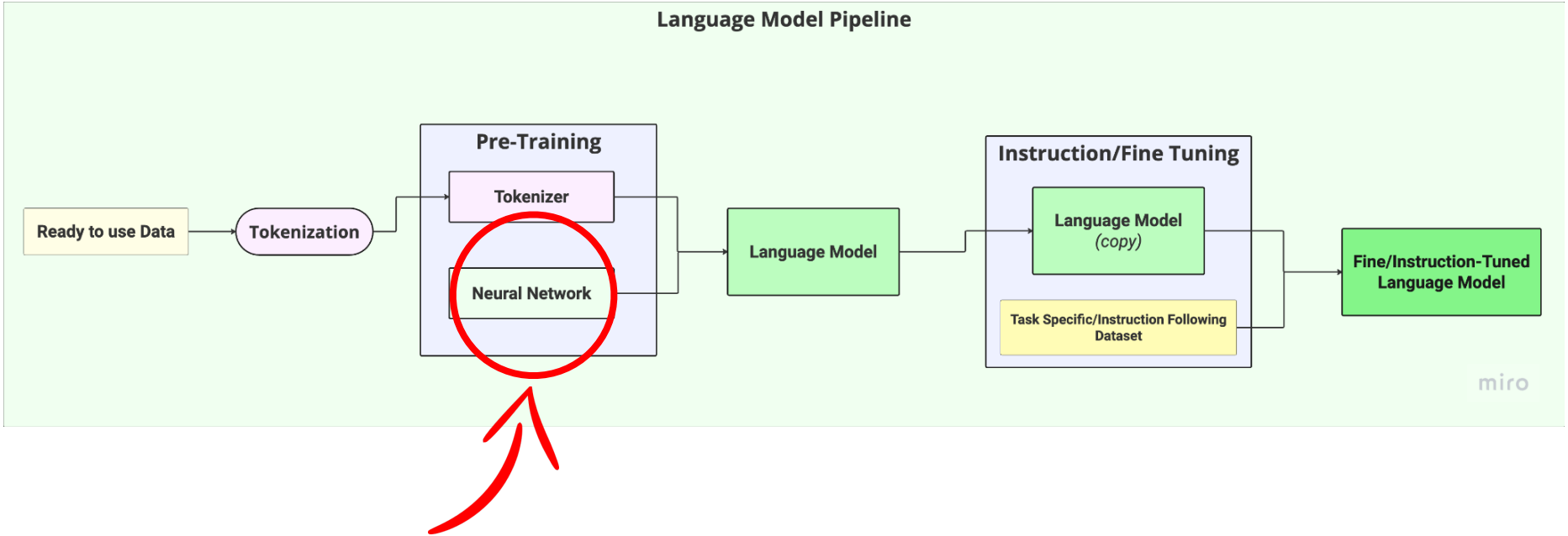
Hovhannes Tamoyan
tamohannes.com

To not get lost in space over time, let’s
Use a **mind map**

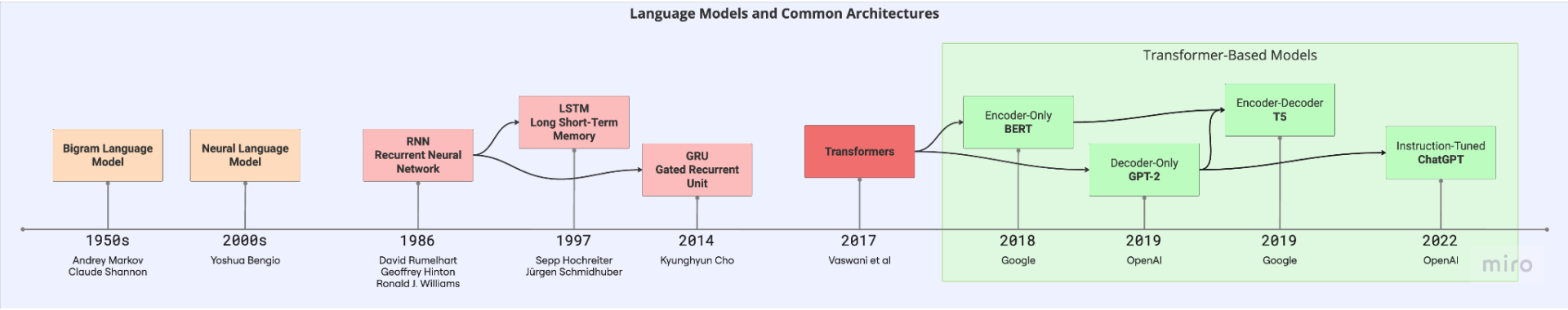
Last time we covered: **Tokenization**



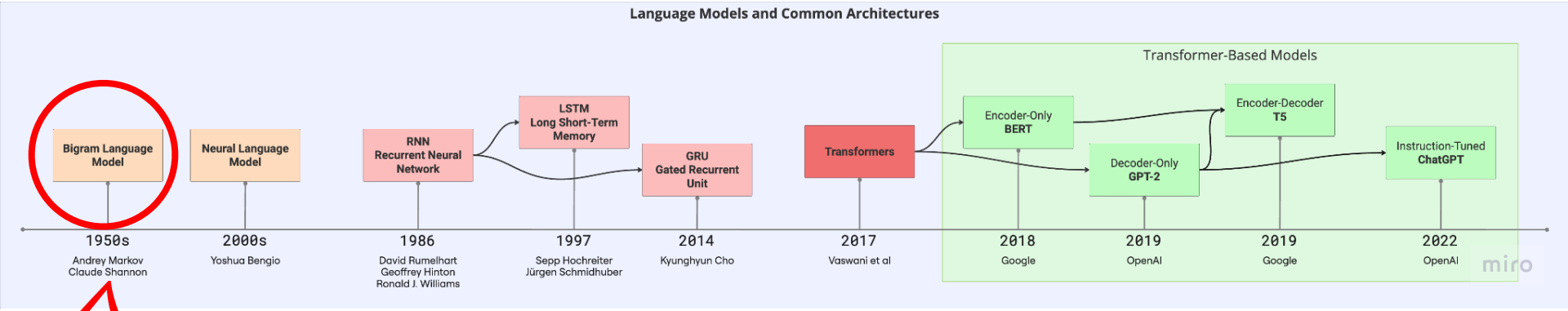
Today's subject: **Neural Networks**



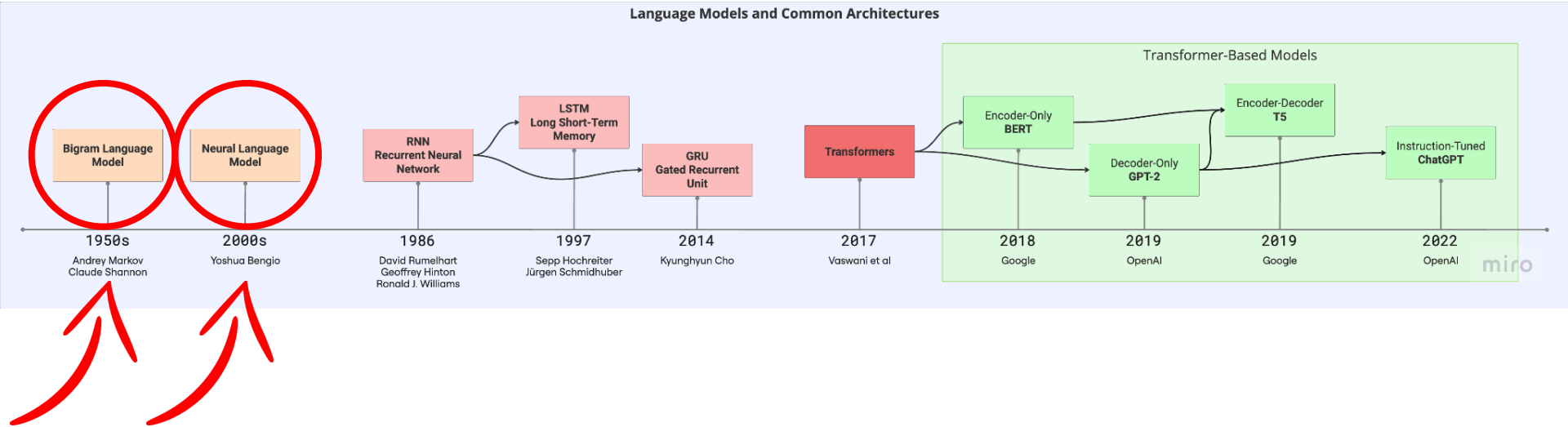
Language Models and Commonly used Architectures



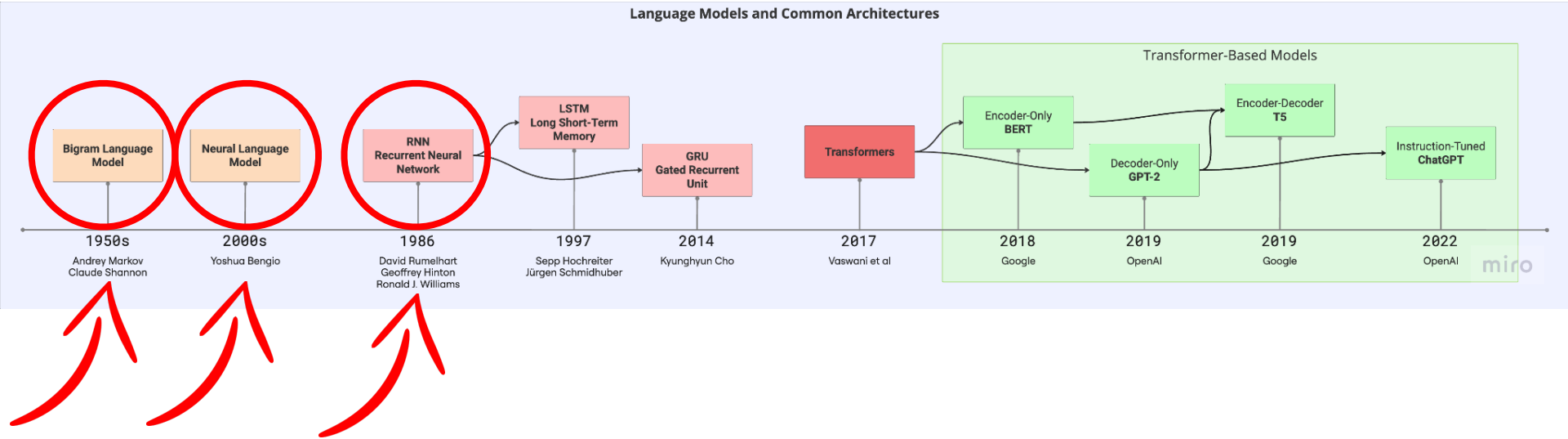
Bigram Language Model



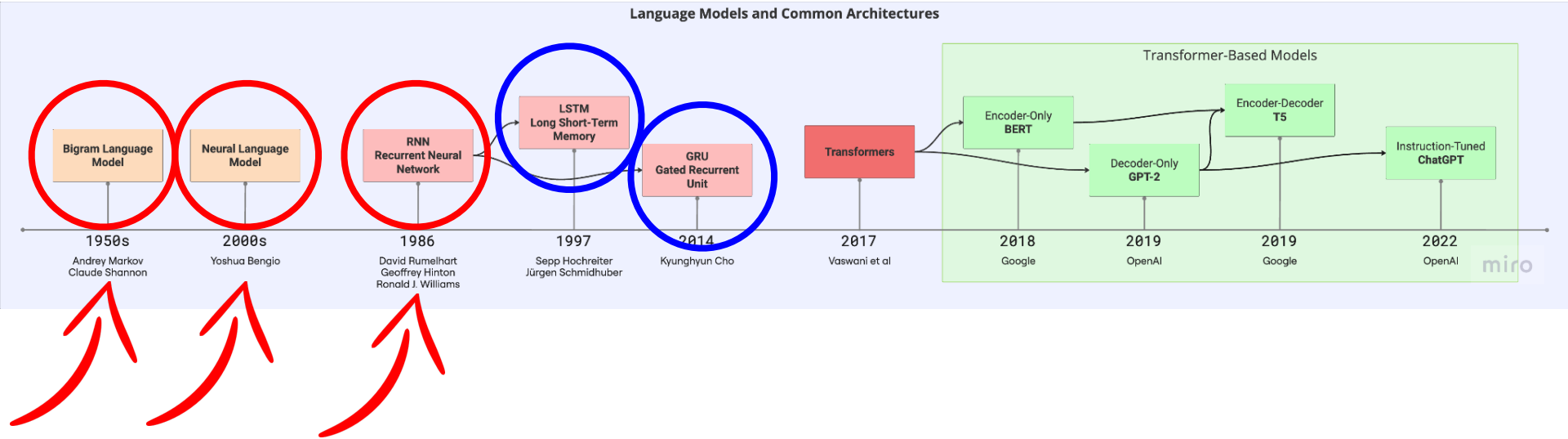
Neural Language Model



Recurrent Neural Network (RNN)



LSTMs and GRUs are for HW6



Recap of Language Modeling

The intermediate objective is to predict what word comes next.
e.g. “The students opened their ____.”

More formally: given a sequence of words $x_1, x_2, \ldots, x_t$ compute the probability distribution of the next word x_{t+1} by learning a predictor parameterized as θ.

$$P(x_{t+1} \mid x_t \ldots x_1; \theta)$$

Where x_{t+1} can be any word in the vocabulary V

Recap of Neural Language Model (NLM)

output distribution

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2) \in \mathbb{R}^{|V|}$$

hidden layer

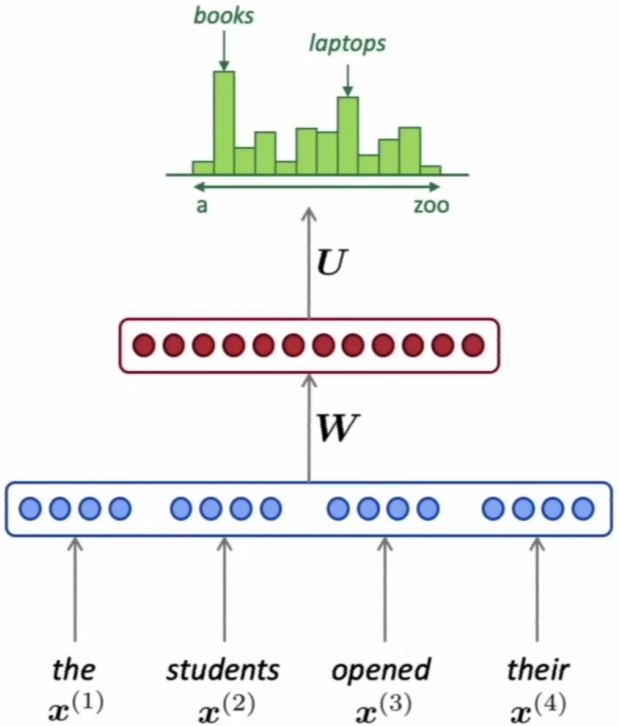
$$\boldsymbol{h} = f(\boldsymbol{W}\boldsymbol{e} + \boldsymbol{b}_1)$$

concatenated word embeddings

$$\boldsymbol{e} = [\boldsymbol{e}^{(1)}; \boldsymbol{e}^{(2)}; \boldsymbol{e}^{(3)}; \boldsymbol{e}^{(4)}]$$

words

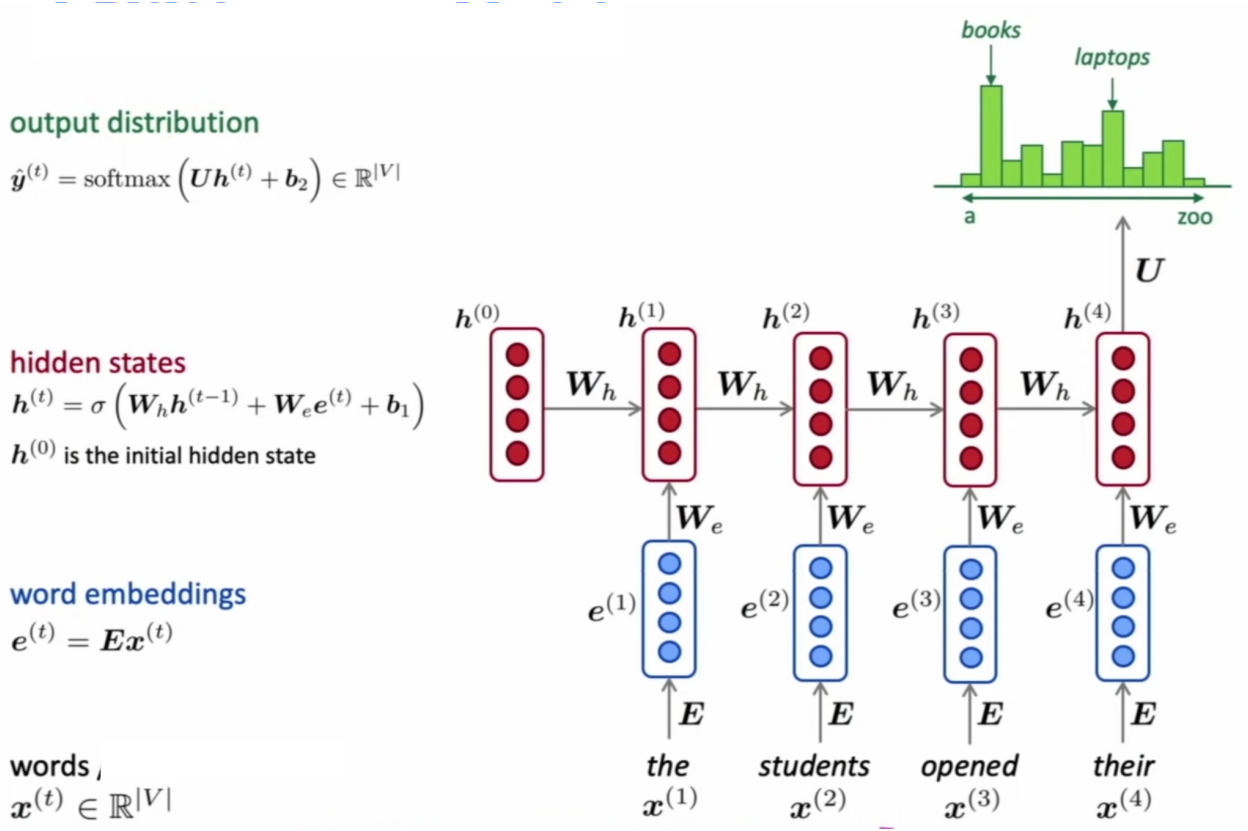
$$\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}, \boldsymbol{x}^{(4)}$$



NLM pros and cons

- Improvements over n-gram LM:
  - No sparsity problem,
  - Don’t need to store all observed n-gram,
- Remaining problems:
  - Fixed window is too small,
  - Enlarging window enlarges W
  - Window can never be large enough
  - No symmetry in how the inputs are processed. X s are multiplied by completely different portion of W.

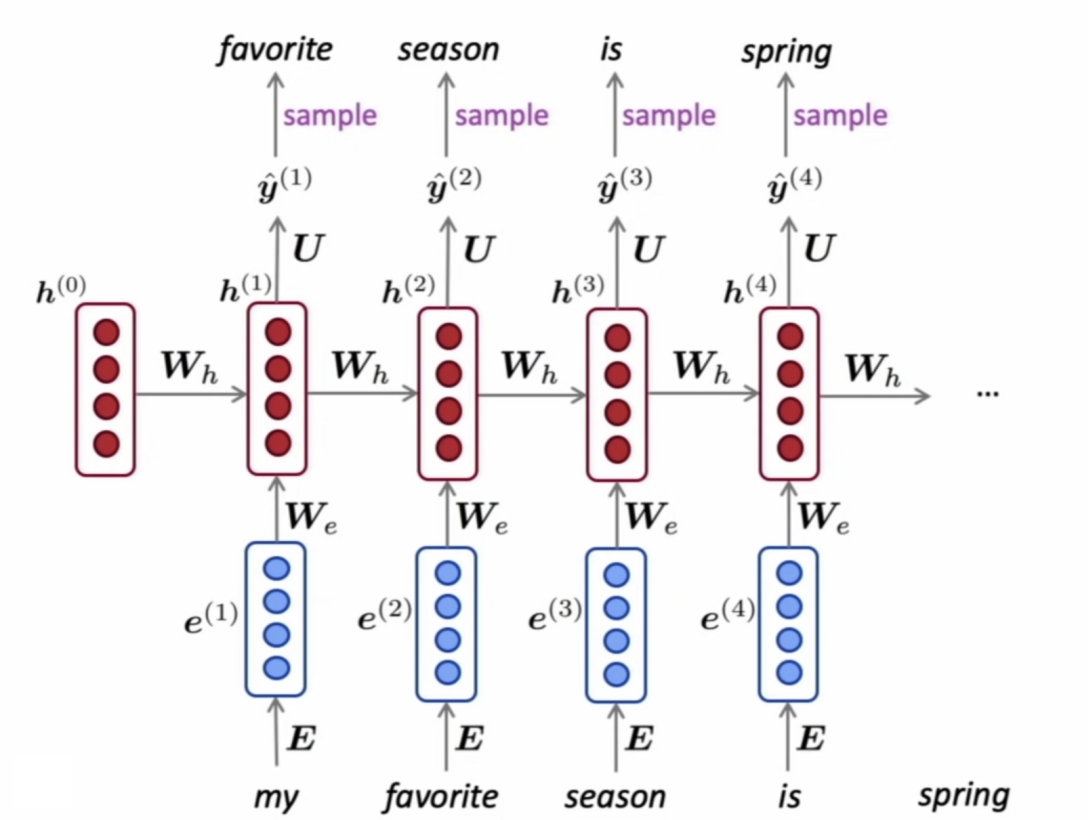
Recap of Recurrent Neural Network (RNN)



RNN pros and cons

- Improvements over NLM:
  - Model size doesn’t increase for longer inputs,
  - Same weights applied on every timestamp, so there is symmetry in how inputs are processed.
- Remaining problems:
  - We need to wait for each token to be processed; the process cannot be sped up.

Text generation with RNN Language Model



Some notations applicable throughout all sessions

- V - Vocabulary size (the number of unique tokens in the tokenizer's vocabulary)
- L – Number of layers in a deep model (commonly used in transformer-based models).
- T - Number of tokens in a sequence, alternatively sequence length
- E - Embedding dimension
- B - Batch size
- H - Hidden dimension depending on the context(layer)
- A – Number of attention heads in a multi-head attention mechanism.