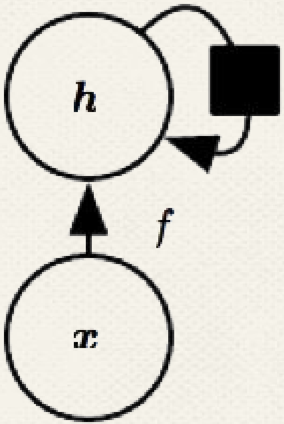
Chapter 10
Sequence Modelling: Recurrent and Recursive Nets

Introduction

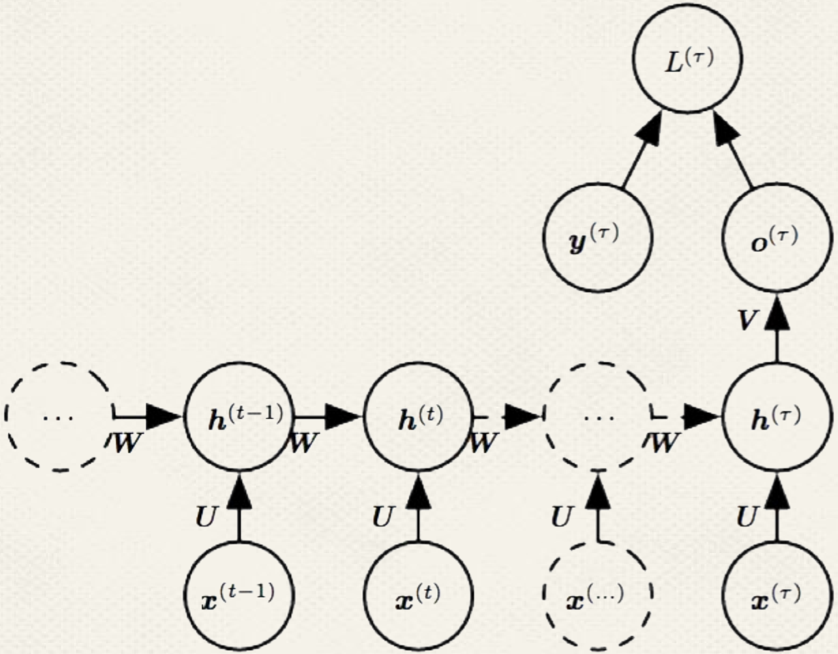


- Sequential data
- Scalable
- Infinitely deep

Introduction

- Parameter sharing through many layers
- Shared weights across time steps
- Cycles affect future values of variables
- Typical output; probability distribution

Structure



Prediction based on sequence of inputs

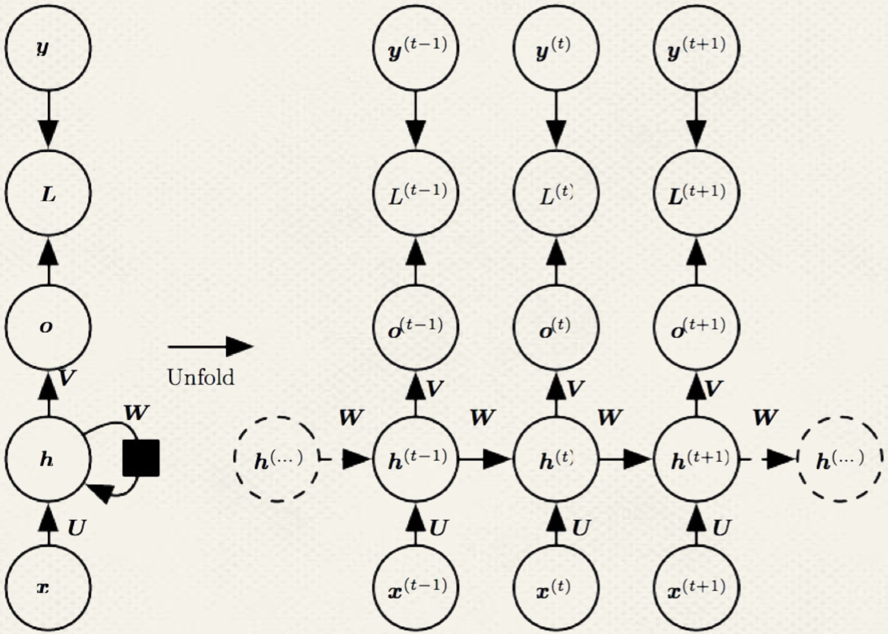
Unfolding

1. $s^{(t)} = f(s^{(t-1)};\theta)$
2. $t = 3$
3. $s^{(3)} = f(s^{(2)};\theta)$
4. $s^{(3)} = f(f(s^{(1)};\theta);\theta)$

Advantages

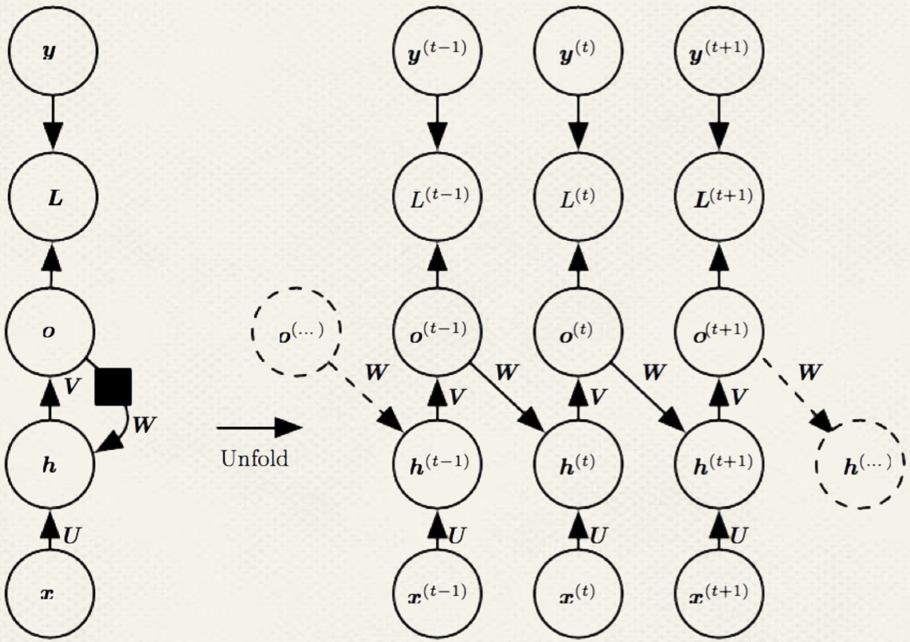
- No fixed input length
- Same transition function and parameters

Design patterns



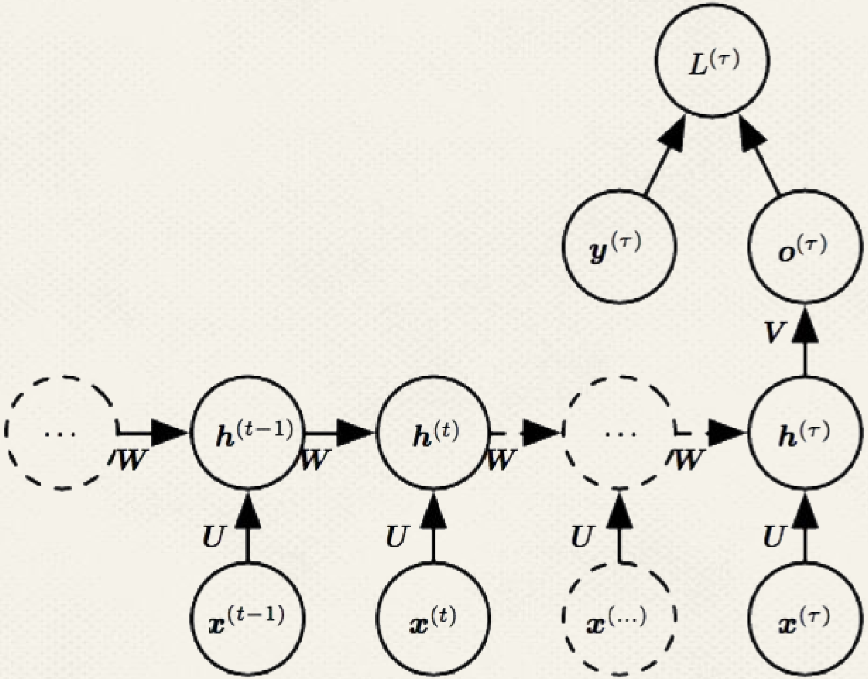
- Recurrent connections between hidden nodes
- Output at each time step
- Powerful

Design patterns



- Recurrent connections from output to hidden
- Output at each time step
- Less powerful, but easier to train

Design patterns



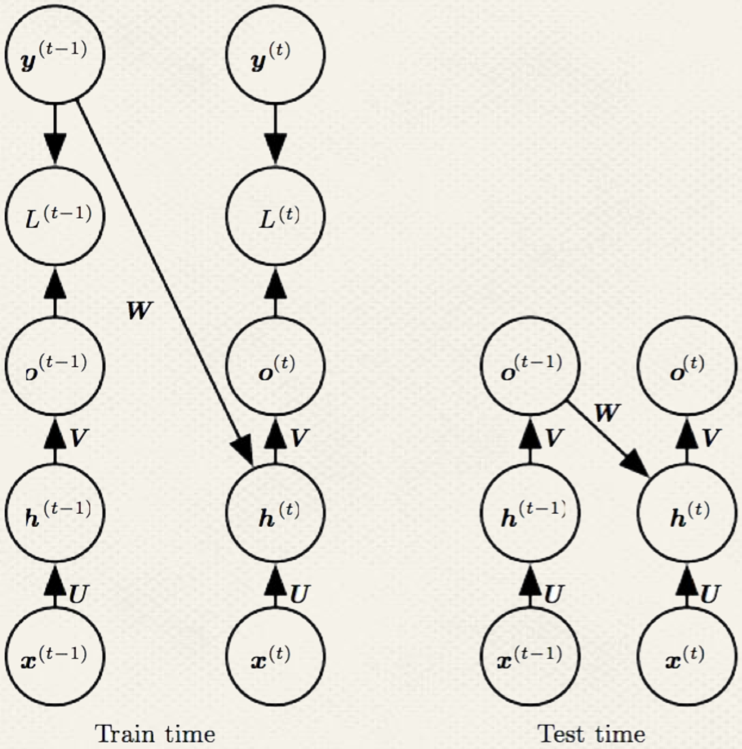
- Recurrent connections between hidden noden
- Single output
- Reads entire sequence

Training

Teacher forcing

(But in a good way.)

- Ideal value > actual output
- Isolates each time step
- Parallel training
- Can cause undesirable effects during testing



Sampling

- Need to stop generation
- Two main methods
  - Reserved character
  - Train another output to decide

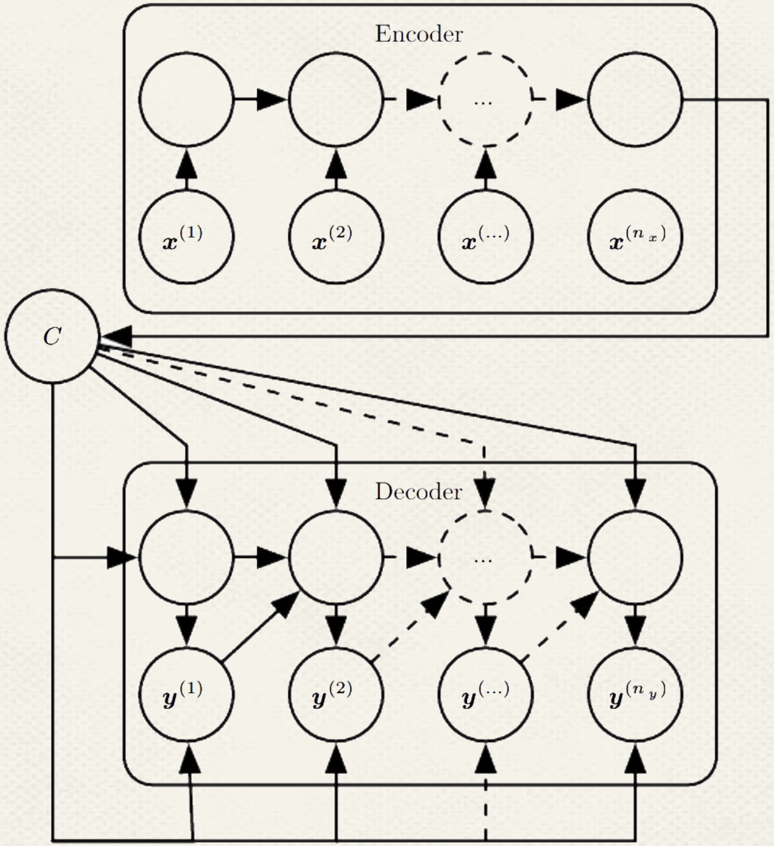
"I think in thy time
Thou hideless time, thought a "

- Badly sampled RNN on Shakespeare

Bidirectional RNN

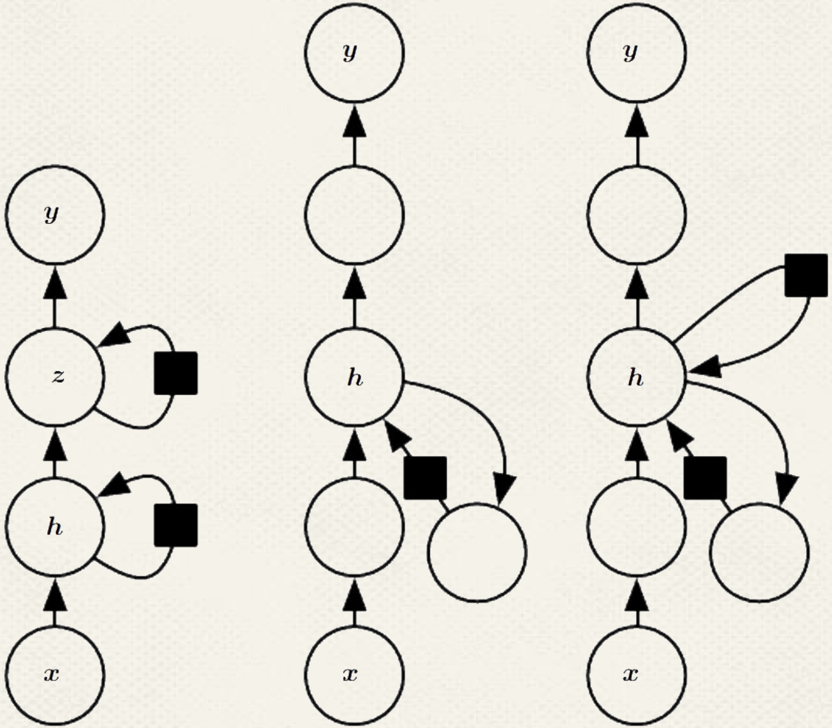
- Forward and backwards in time
- Degrading sensitivity from time t
- Handwriting recognition
- Speech recognition
- Extensions to other dimensions

Encoder-Decoder



- Variable input length
- Encoder→Context→Decoder
- Translation

Deep Recurrent Networks



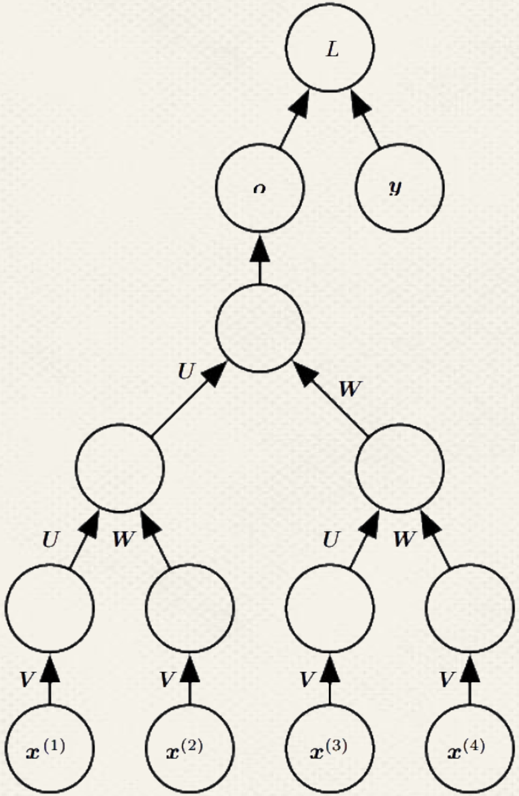
We have to go deeper

- Hierarchically
- Deep recursion (MLP)
- Skip connections

\+ Greater potential

\- Harder to learn

Recursive Neural Networks



- ○ Process data structures for neural networks
- ○ Natural language processing
- ○ Computer vision

+ Reduced depth
- Hard to structure the tree

The Challenge of Long-Term Dependencies

Chapter 10.7

The Challenge of Long-Term Dependencies

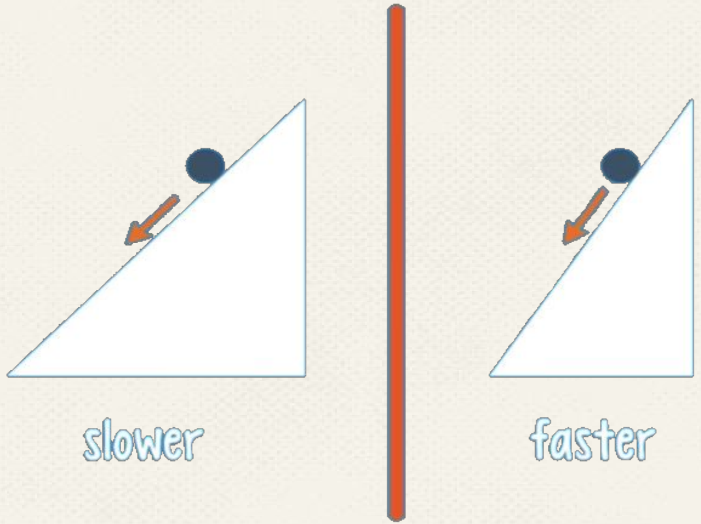
Vanishing gradient problem

Exploding gradient problem

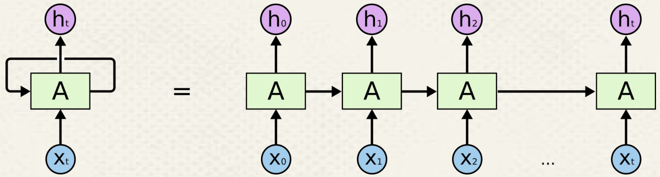
The Challenge of Long-Term Dependencies

Gradient

- = Rate at which the cost changes with respect to weights and biases
- Cost is lowered by making small adjustments to w and b



The Challenge of Long-Term Dependencies



- RNN involve composition of the same function multiple times → extremely nonlinear
- $\boldsymbol{h}^{(t)} = \boldsymbol{W}^\top \boldsymbol{h}^{(t-1)} \quad \longrightarrow \quad \boldsymbol{h}^{(t)} = \boldsymbol{Q}^\top \boldsymbol{\Lambda}^t \boldsymbol{Q} \boldsymbol{h}^{(0)}$

The Challenge of Long-Term Dependencies

Vanishing gradient problem

The gradient get smaller as we move backward through the hidden layers

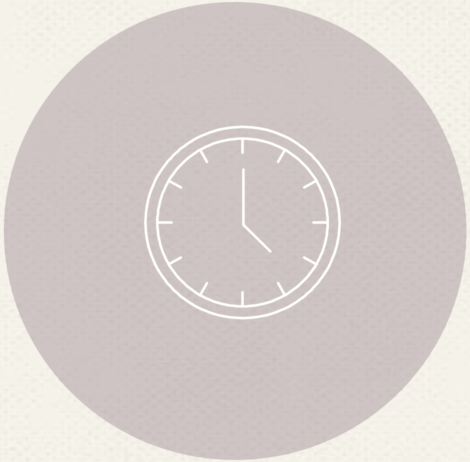
- Neurons in earlier layers learn much slower than neurons in later layers

Exploding gradient problem

The gradient gets much larger in earlier layers, will lead to big changes of weights

- Will cause the network to forget almost everything it has learned
- Rare

The Challenge of Long-Term Dependencies



Too long to train

Inaccurate

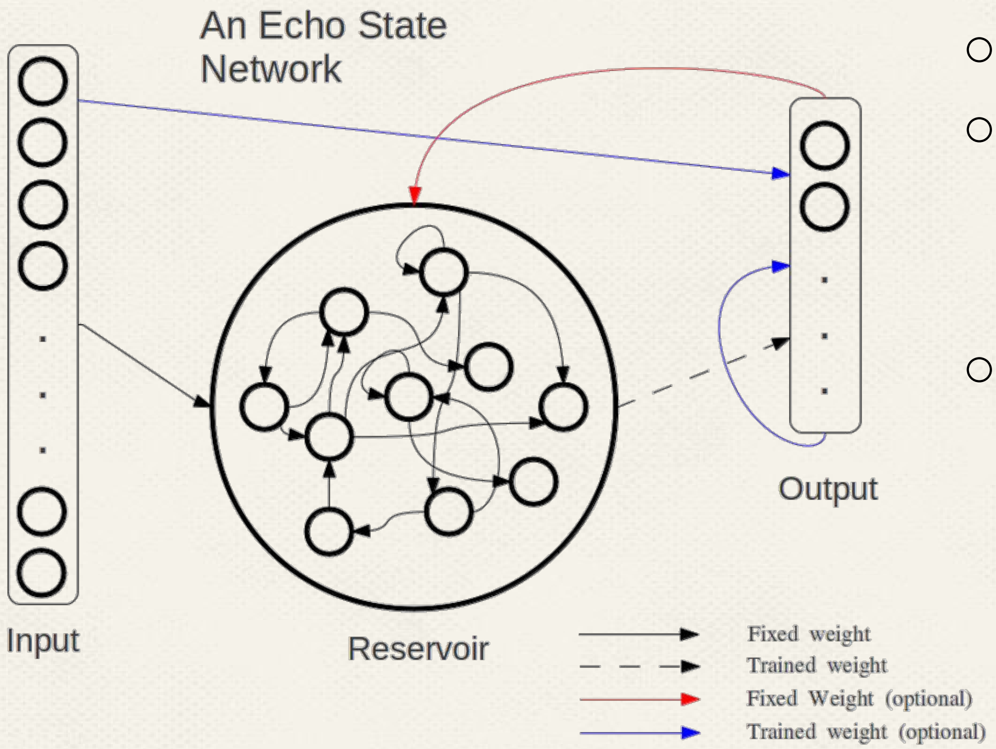
Echo State Network

Chapter 10.8

Echo State Network

- **Difficult parameters to learn:**
  - Recurrent weight mapping from $h^{(t-1)}$ to $h^{(t)}$
  - Input weights mapping from $x^{(t)}$ to $h^{(t)}$
- **Solution**: Only learn the output weights

Echo State Network



- <u>Only learn the output weights</u>
- **Reservoir computing**: hidden units form a reservoir of temporal features
- **Strategy**: Fix weights such that information is carried forward through time but <u>does not explode</u>

Leaky Units and Other Strategies for Multiple Time Scales

Chapter 10.9

Leaky units and other strategies for multiple time scales

- Design a model that operates at multiple time scales
  - Fine grained time scales
  - Coarse time scales

Leaky units and other strategies for multiple time scales

- **Adding Skip Connections through Time**
  - Time-delay of d
  - Gradients now diminish exponentially as a function of T/d rather than T
  - Ensuring that a unit always can learn to be <u>influenced by a value from *d* time steps earlier</u>

Leaky units and other strategies for multiple time scales

- **Leaky units**
  - Each hidden state u(t) is now a "summary of history"
    - History up to state u(t-1)
    - Present time v(t)

$$\mu^{(t)} \leftarrow \alpha\mu^{(t-1)} + (1-\alpha)v^{(t)}$$

Leaky units and other strategies for multiple time scales

- **Leaky units**
  - Strategies for setting the time constants used by leaky units:
    - Constant
    - Free parameters and learn them

Gated RNNs

Gated RNNs

- Most efficient sequence models
  - Gated RNNs
    - LSTM
    - GRU

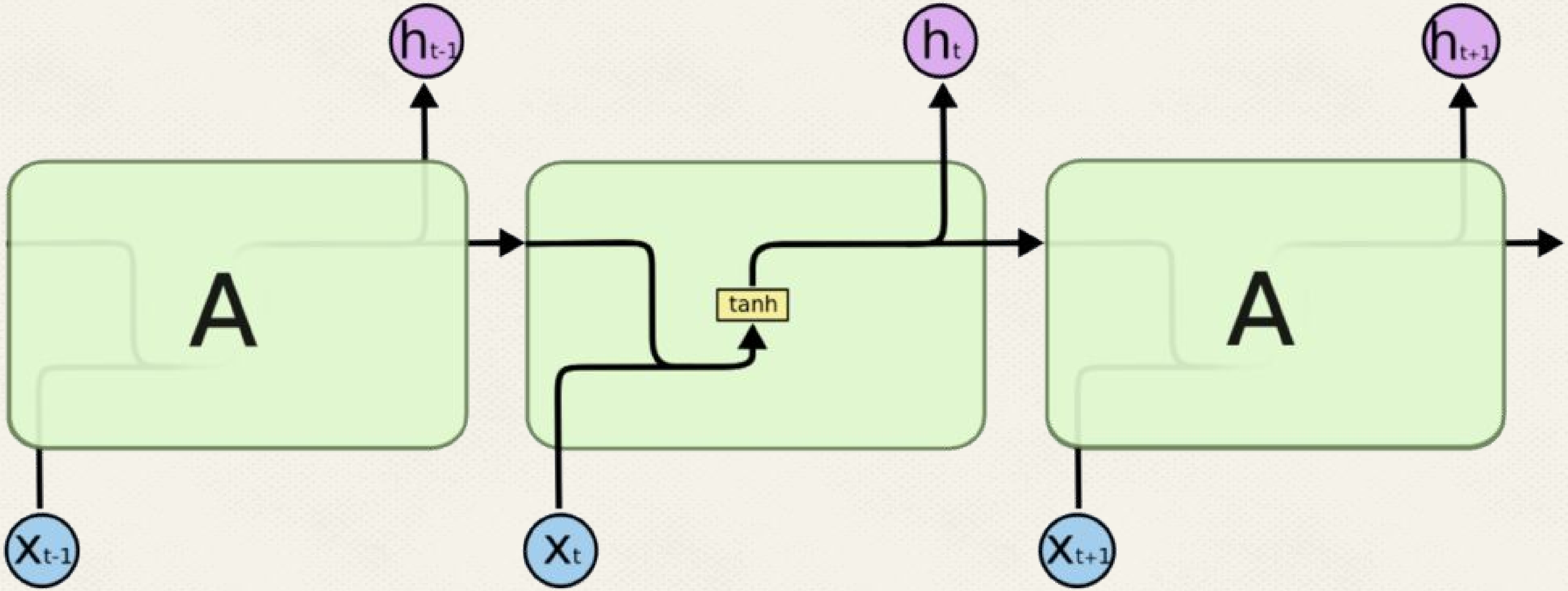
Gated RNNs

- Same idea as leaky units
  - Connection weights that ~~*were either manually chosen constants or were parameters*~~
  - Connection weights that *may change at each time*
- Forget old state - by learning

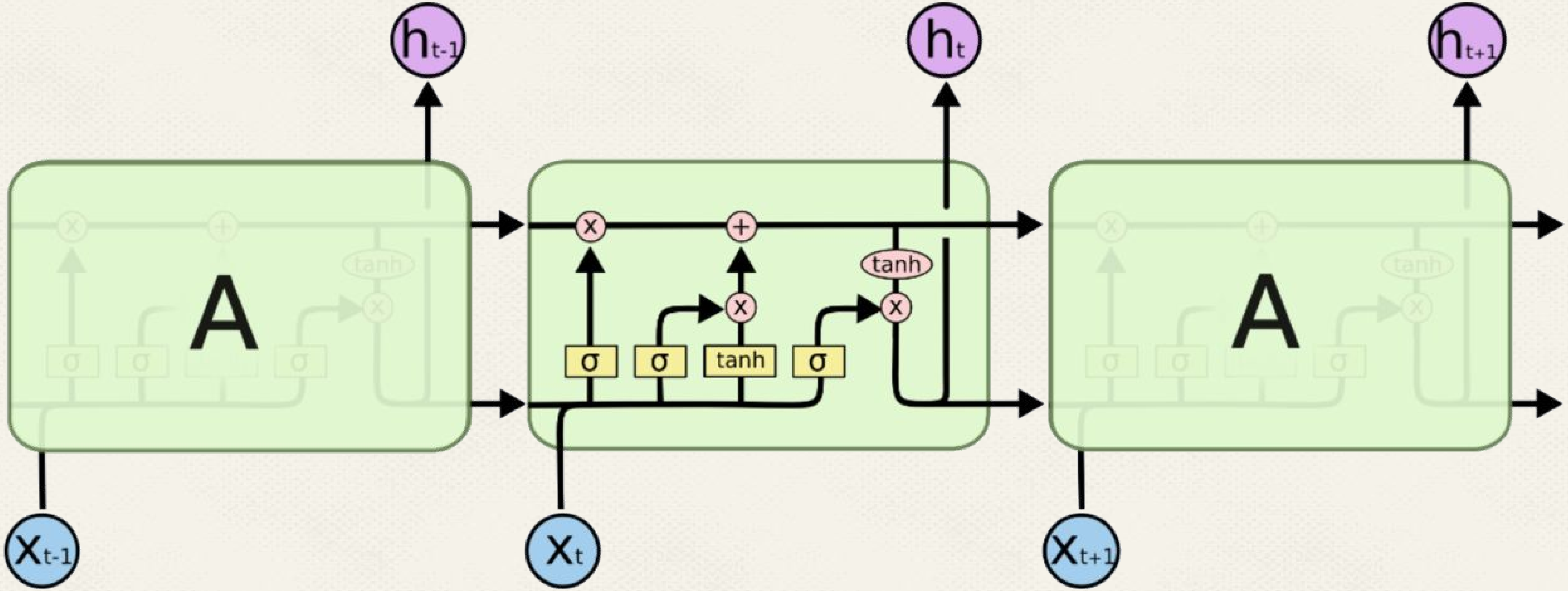
LSTM

- Handwritten recognition
- Speech recognition
- Machine translation
- Image captioning
- Parsing

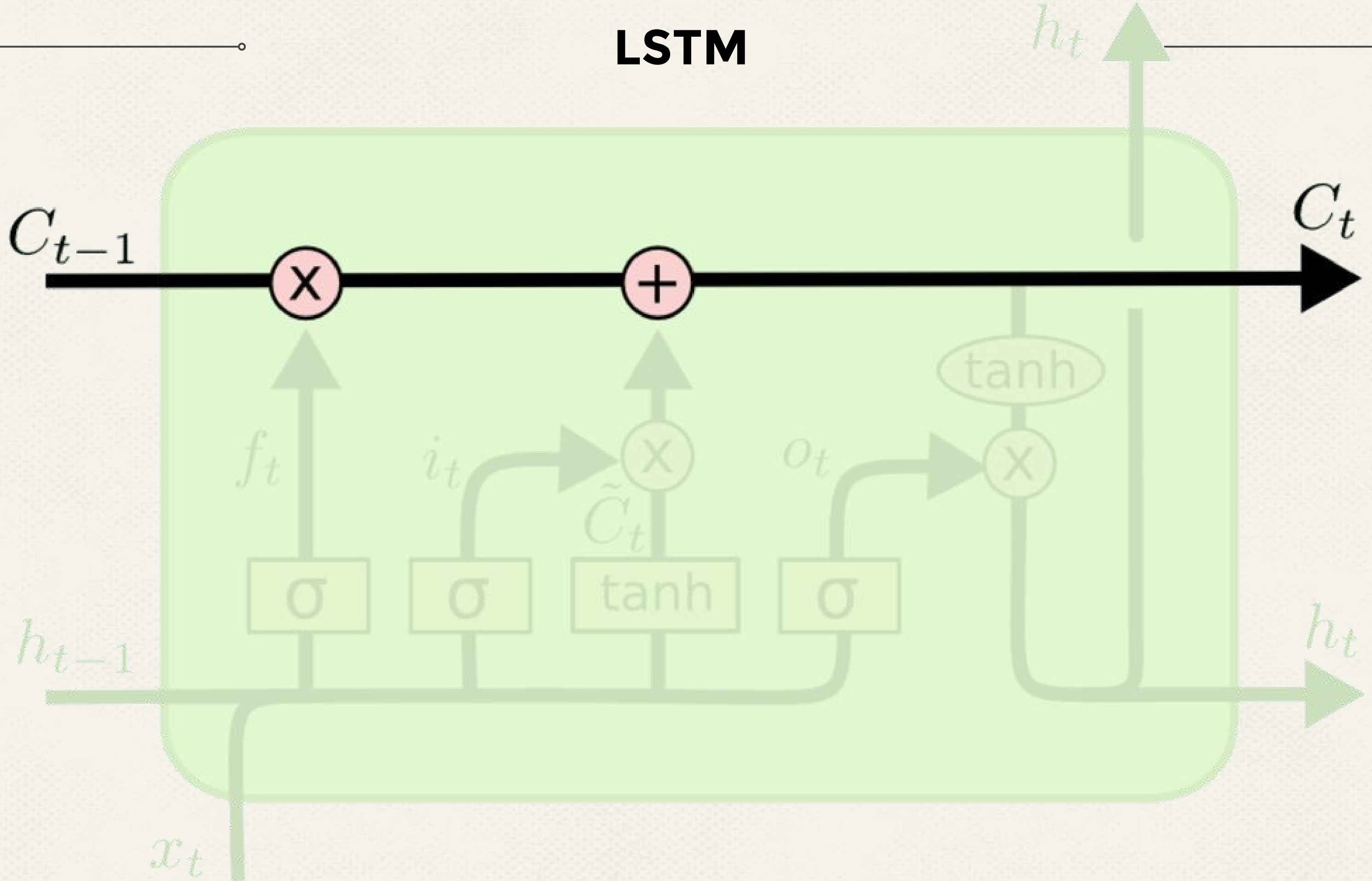
LSTM



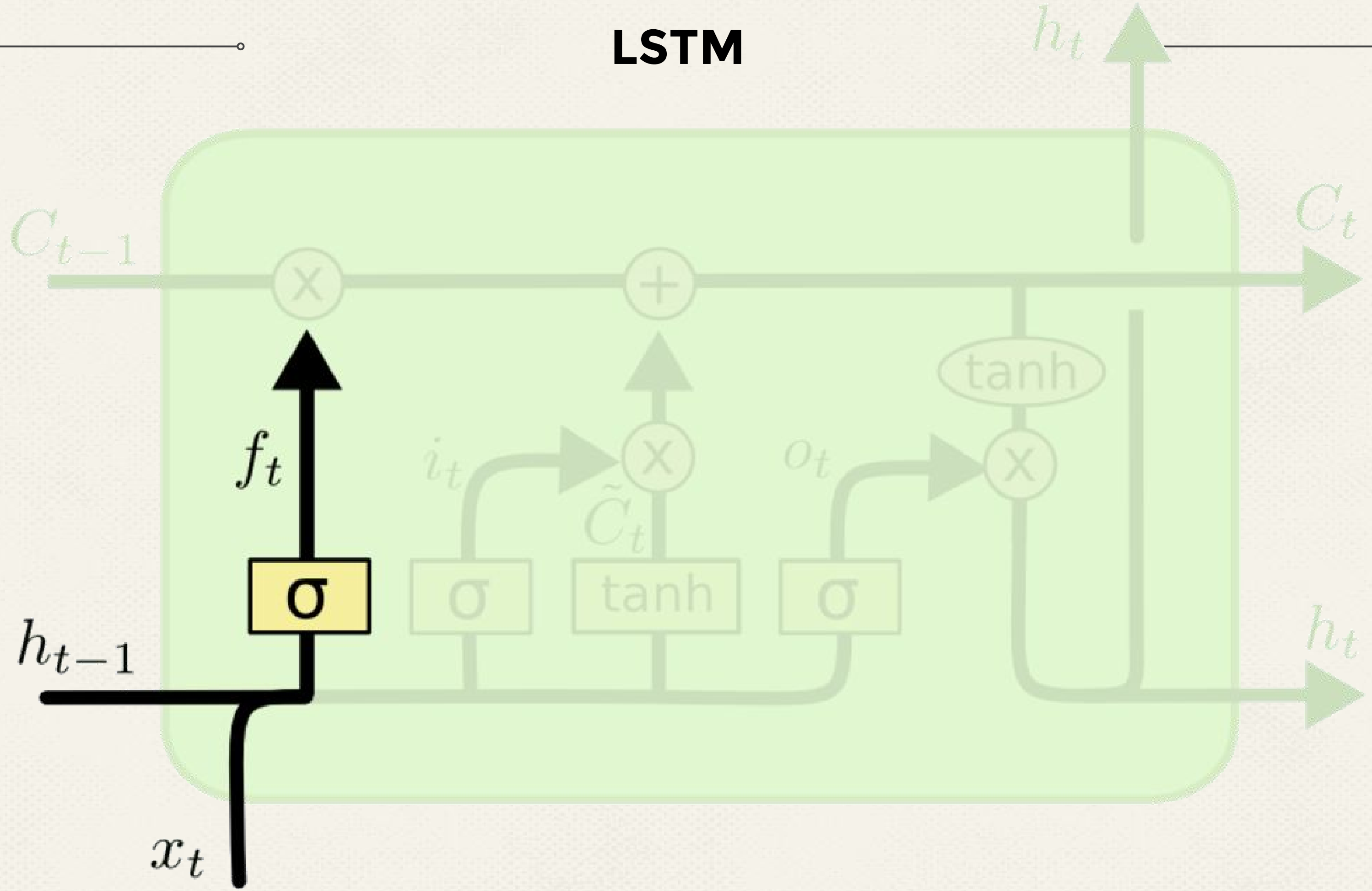
LSTM



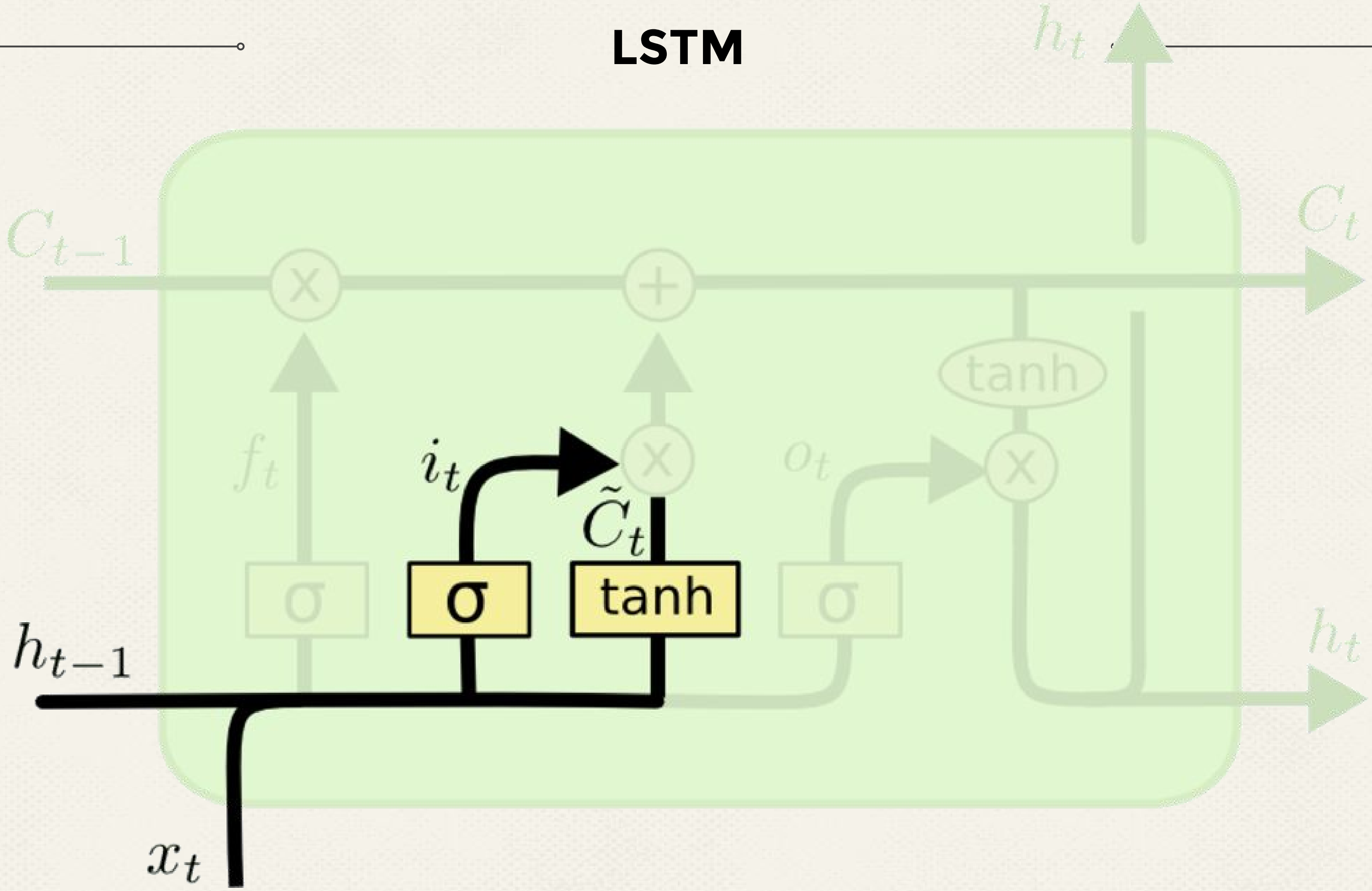
LSTM



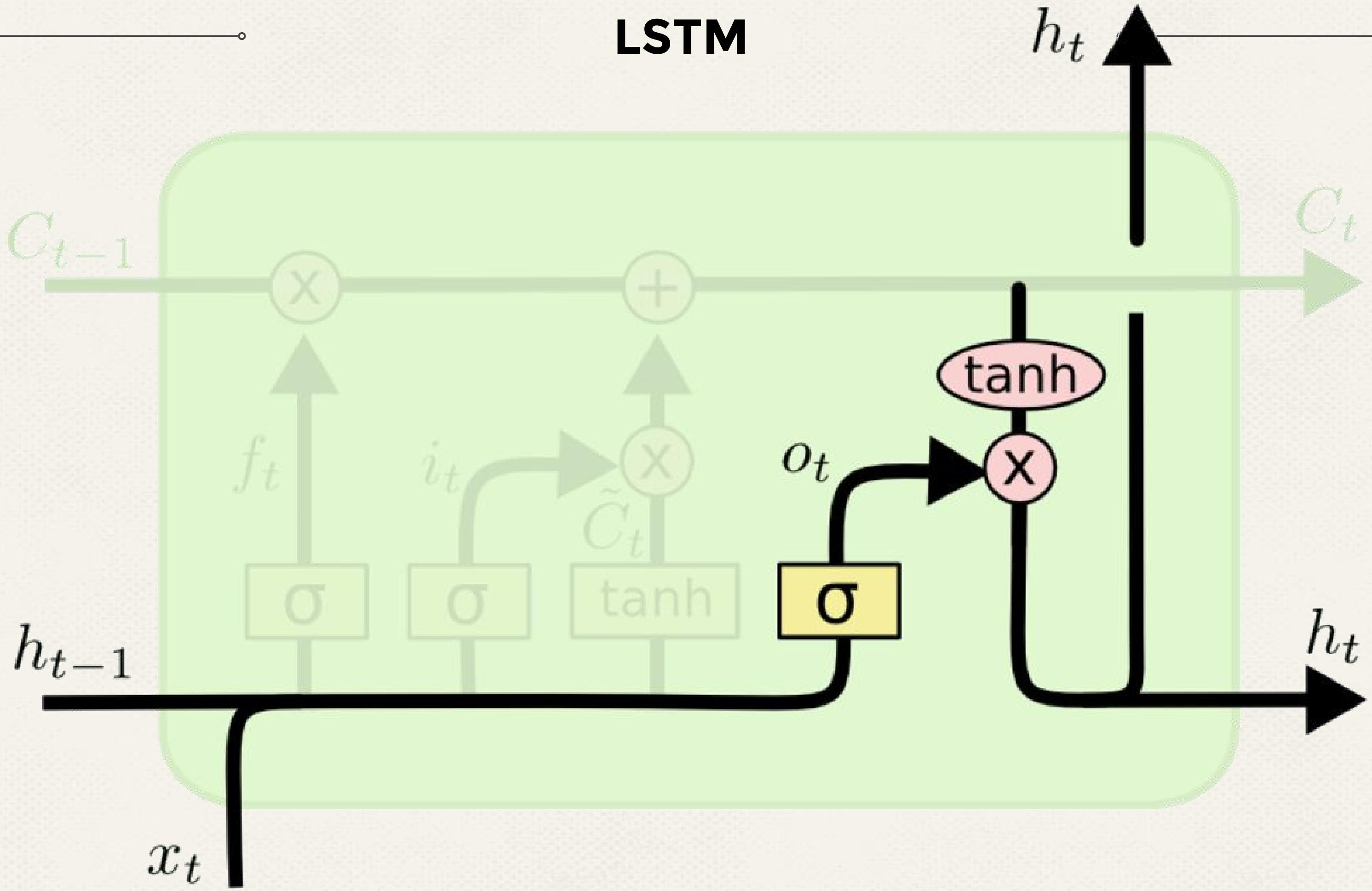
LSTM



LSTM

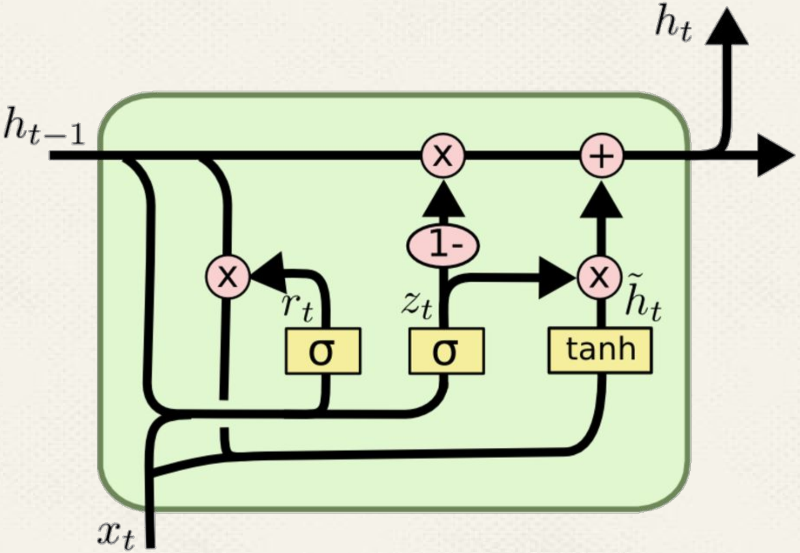


LSTM



GRU

- Less gates
  - Update gate
  - Reset gate
- Simpler and easier to train
- Performs slightly worse than LSTM

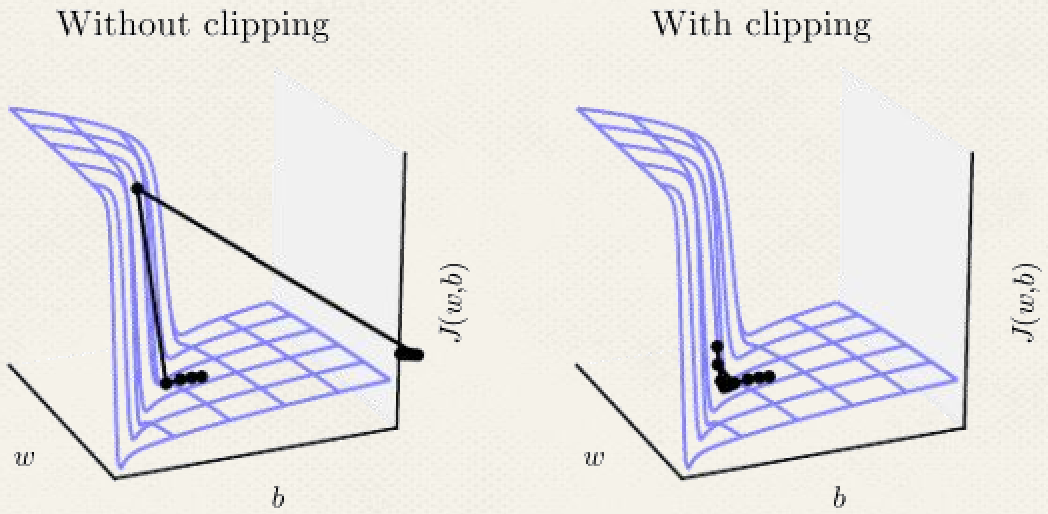


Optimization

- Exploding gradients
- Vanishing gradients

Exploding gradients

- Clipping gradients

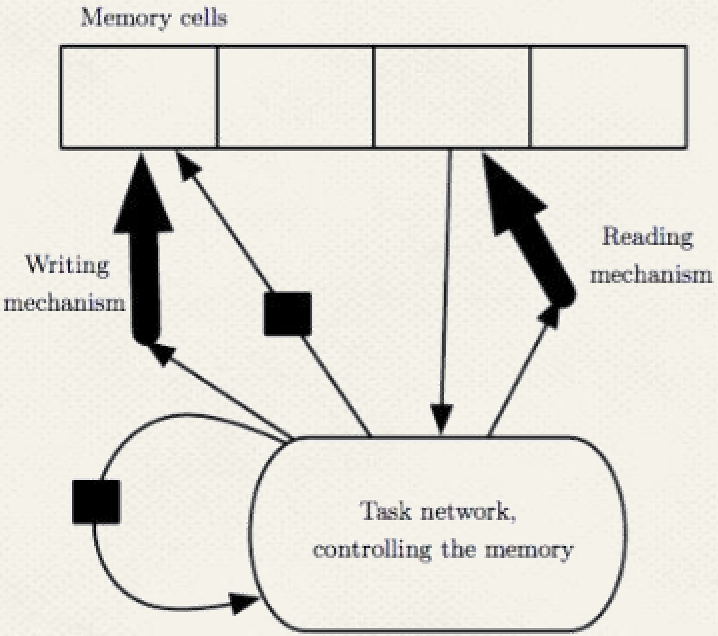


Vanishing gradients

- LSTMs and other self-loops
- Regularization
  - Not as good as LSTM

Explicit memory

- Memory networks
- Neural Turing machine



DEMO

https://github.com/sherjilozair/char-rnn-tensorflow

250 epochs
Works of Shakespeare

POLIXENES:
Thou art your father: let's have as help?

QUEEN ELIZABETH:
Here a poison grass, you come; and so?
stamp, and not so villains up'st thoe you suffer's kindles
Hastingness in't in his friends, the loving
break not the rest as they away.

HENRY BOLINGBROKE:
What as he signing of your bed!

50 epochs of Don Quijote

(In spanish)

“Que podrá responder que gambiera mi pequeña iguarna donde tambores a mayor tener

pequeña modo que

hija y muy animal, se puso y le corona caber matando, les sabe iba la menteca las lenguas magns

Anselmo, y aderezada, para Resol fertiguoso ha de ser”

“Con éstos iba ensartando otros disparates, todos al modo de los que sus

libros le habían enseñado, imitando en cuanto podía su lenguaje. Con esto,

caminaba tan despacio, y el sol entraba tan apriesa y con tanto ardor”

25 epochs
Excerpt of The Holy Bible

“

Galatians 4:5 But thy words again, and thou break chosen: but now will I porters to me, breught in the sea, Gives, even unto his own eye is upon Isreating, and to another, and the ark of God from the ass down.