L101: Machine Learning for Language Processing

Lecture 7

Guy Emerson

Today’s Lecture

- Neural networks
- Sequence labelling
- Language modelling

Features

input $\longrightarrow$ features $\longrightarrow$ prediction

Features

input ⟶ features ⟶ prediction

engineered

Features

input ⟶ features ⟶ prediction

engineered trained

Features

input ⟶ features ⟶ prediction

trained trained

Features

input ⟶ features ⟶ prediction

trained trained

- Engineering at a more abstract level

Feedforward Networks

$$x \mapsto f_1(x) \mapsto f_2(f_1(x))$$

Feedforward Networks

$$x \mapsto f_1(x) \mapsto f_2(f_1(x))$$

- Linear: $f(x) = Ax$

Feedforward Networks

$$x \mapsto f_1(x) \mapsto f_2(f_1(x))$$

- Linear: $f(x) = Ax$
- but can simplify matrix multiplication $AB = C$

Feedforward Networks

$$x \mapsto f_1(x) \mapsto f_2(f_1(x))$$

- Nonlinear: $f(x) = g(Ax)$

Feedforward Networks

$$x \mapsto f_1(x) \mapsto f_2(f_1(x))$$

- Nonlinear: $f(x) = g(Ax)$
  (g applied componentwise)

Feedforward Networks

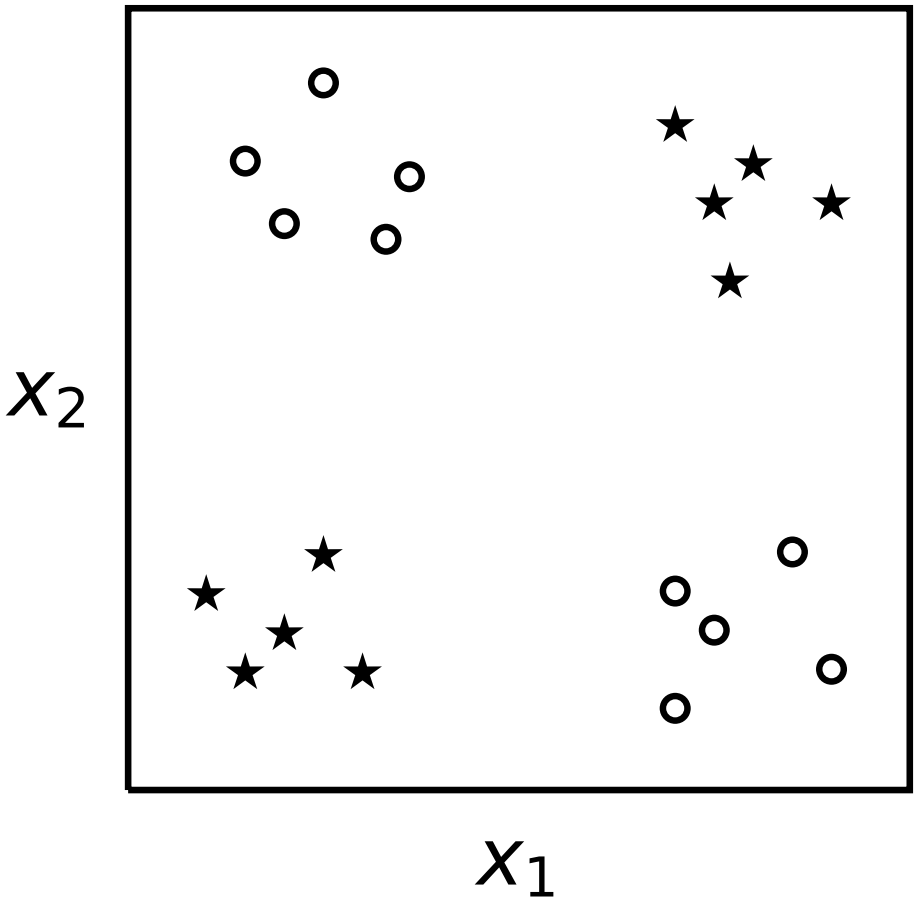
$$x \mapsto f_1(x) \mapsto f_2(f_1(x))$$

- Nonlinear: $f(x) = g(Ax)$
  (g applied componentwise)
- Can approximate any function

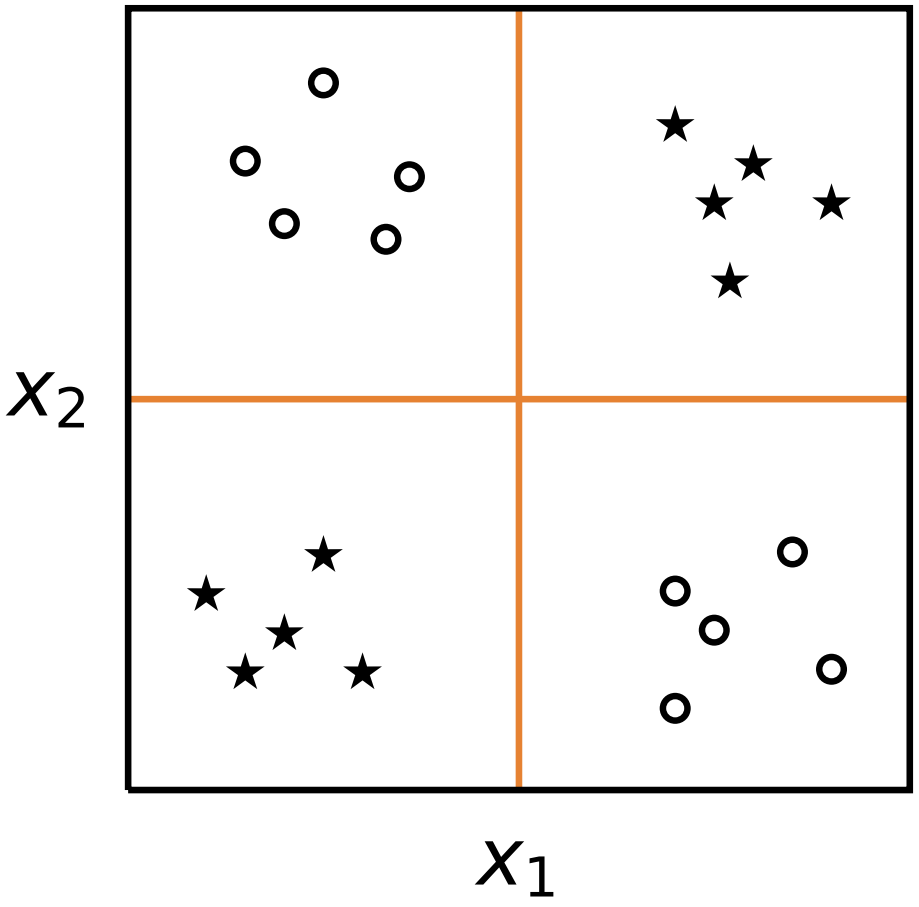
Nonlinear Activation Functions

- $\frac{1}{1+e^{-x}}$ “sigmoid”
- $\frac{1-e^{-2x}}{1+e^{-2x}}$ “tanh”
- $\max\{x, 0\}$ “rectified linear”
- $\log(1 + e^{x})$ “softplus”

Nonlinear Decision Boundaries



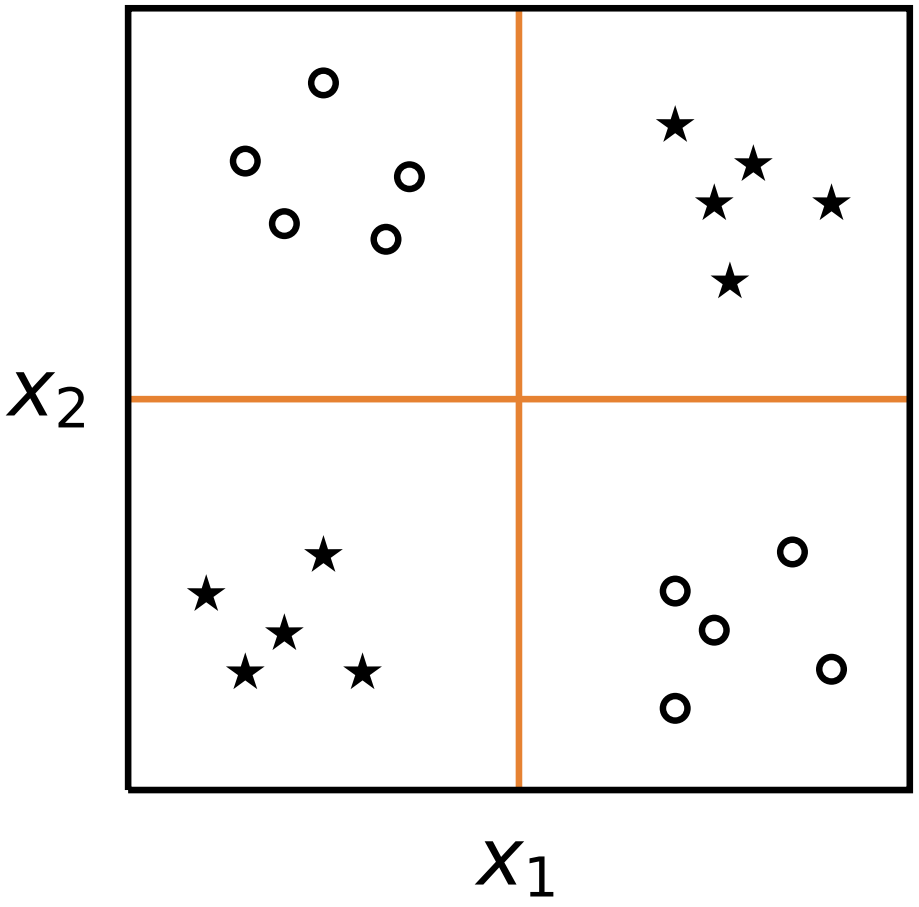
Nonlinear Decision Boundaries



Quadratic kernel:

$$x_1 x_2 - x_1 - x_2 + 1 = 0$$

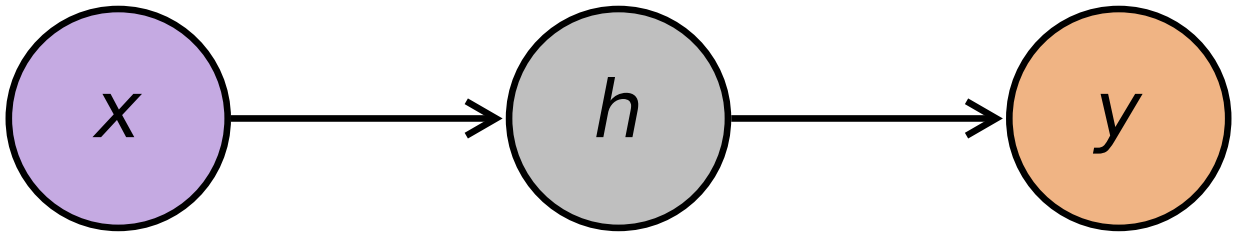
Nonlinear Decision Boundaries



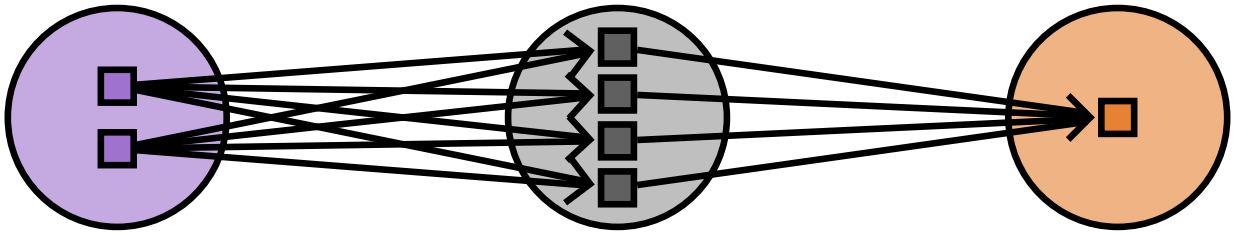
Rectified linear units:

$$
\begin{aligned}
& r(\ \ x_1 + x_2 - 2) \\
+\, & r(-x_1 - x_2 + 2) \\
-\, & r(\ \ x_1 - x_2) \\
-\, & r(-x_1 + x_2) \\
= \, & 0
\end{aligned}
$$

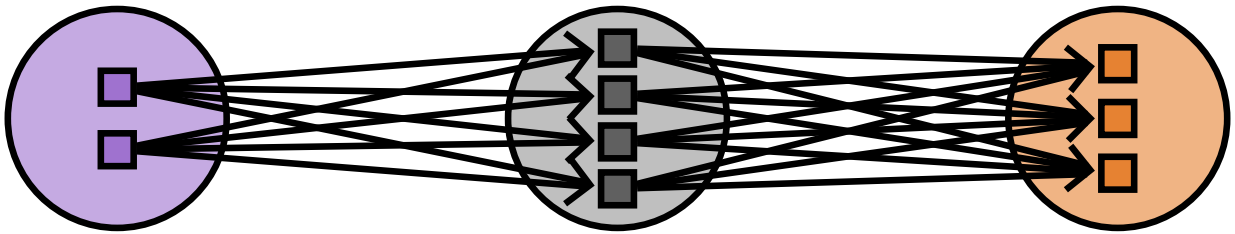
Feedforward Networks



Feedforward Networks

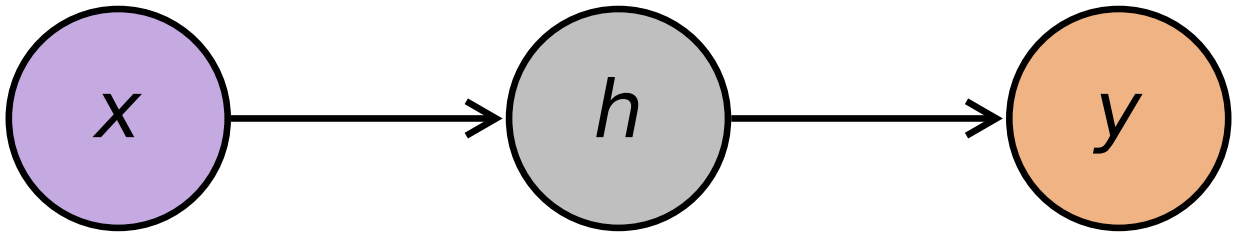


Feedforward Networks

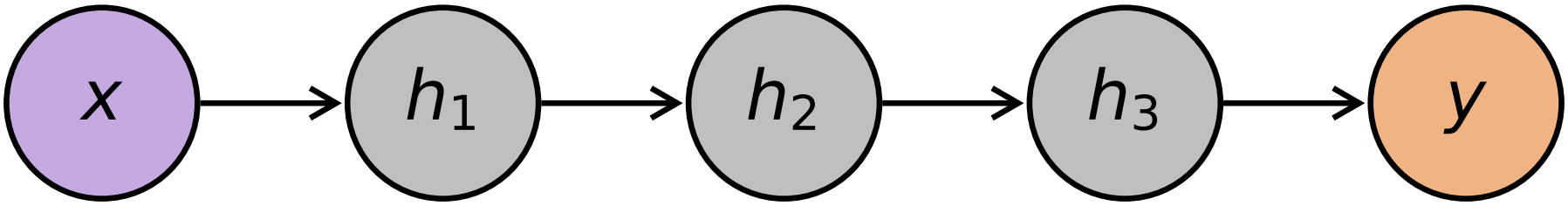


Multiple classes: “softmax”
(like logistic regression)

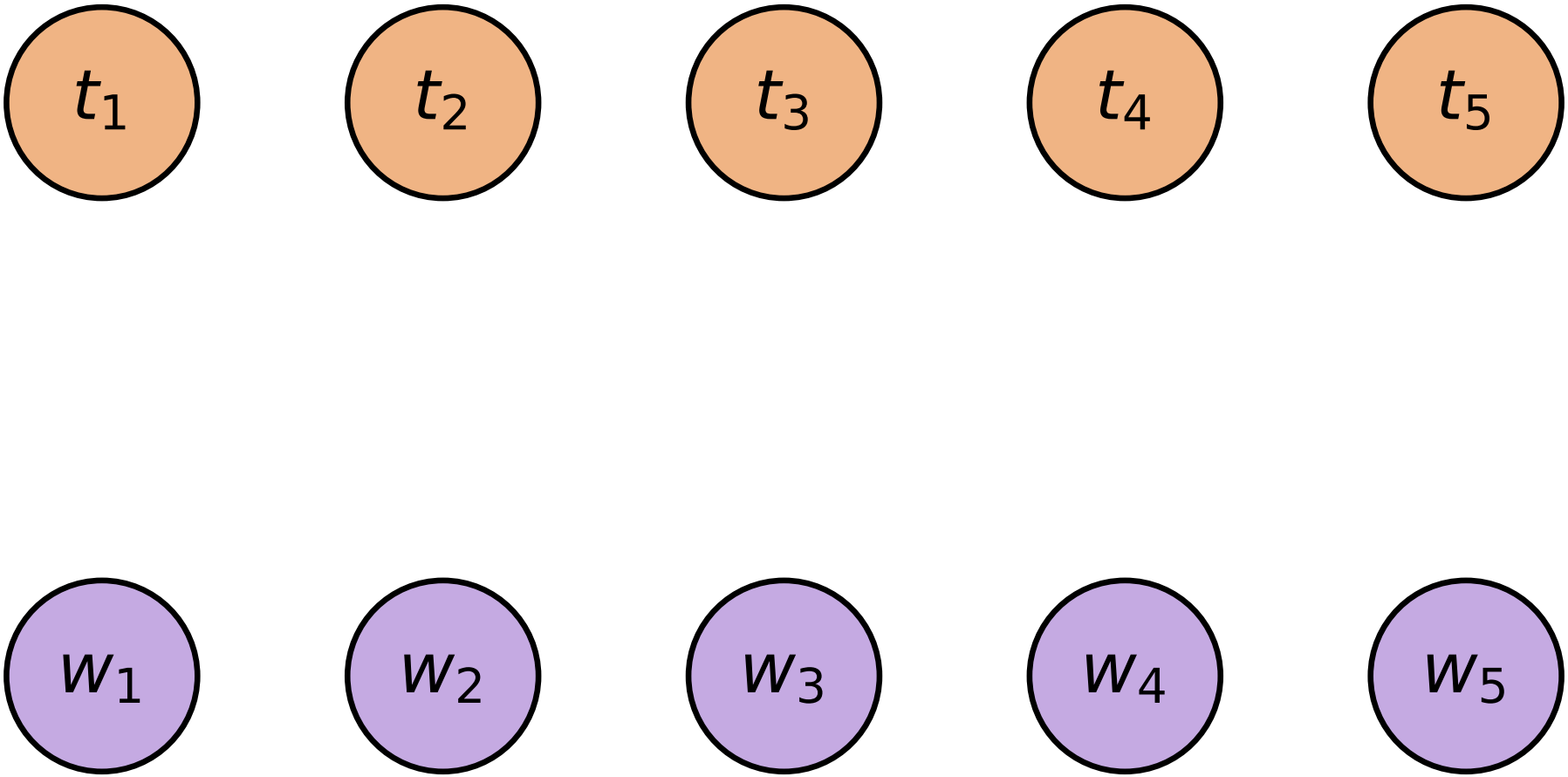
Feedforward Networks



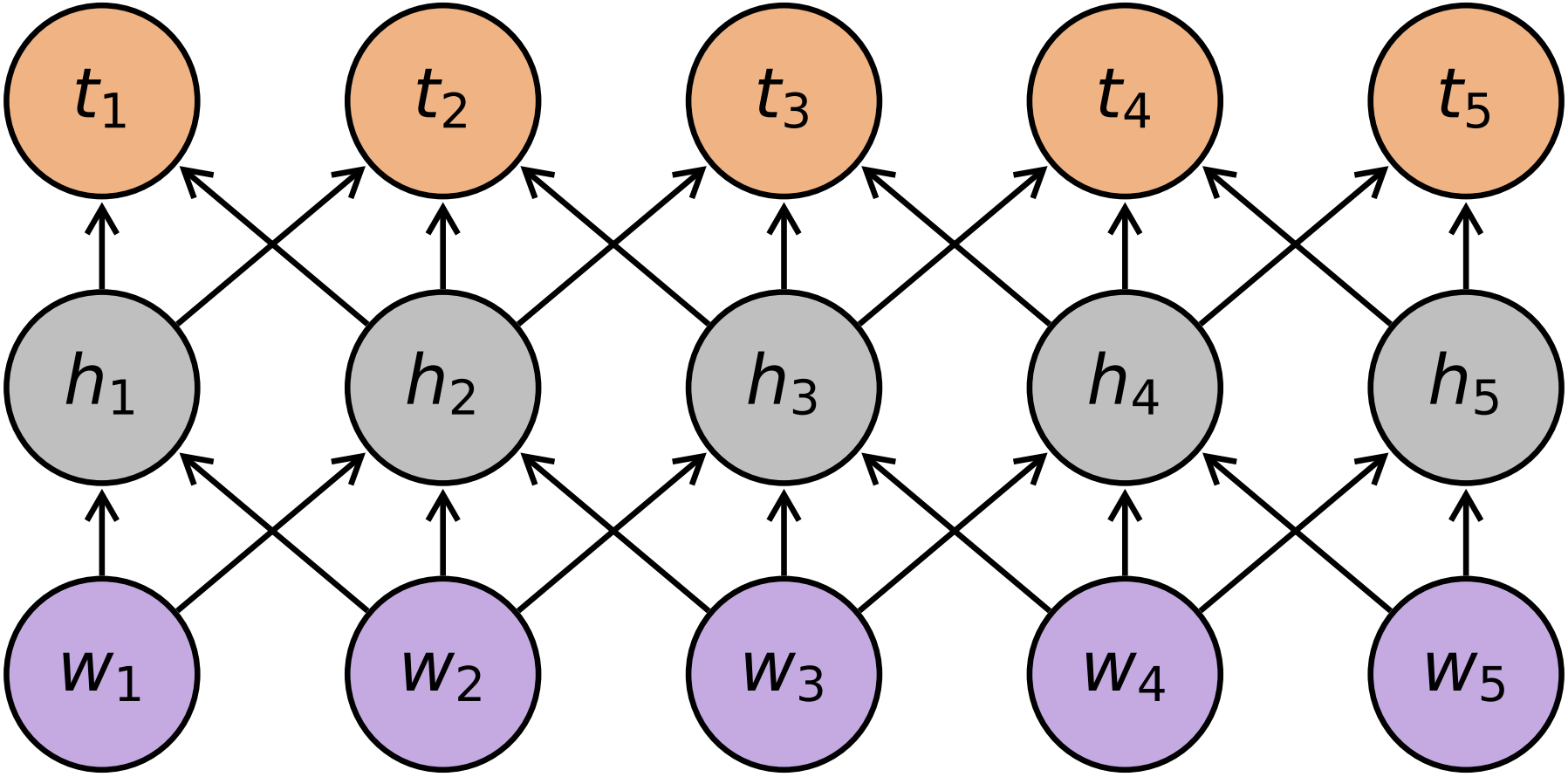
“Deep” Feedforward Networks



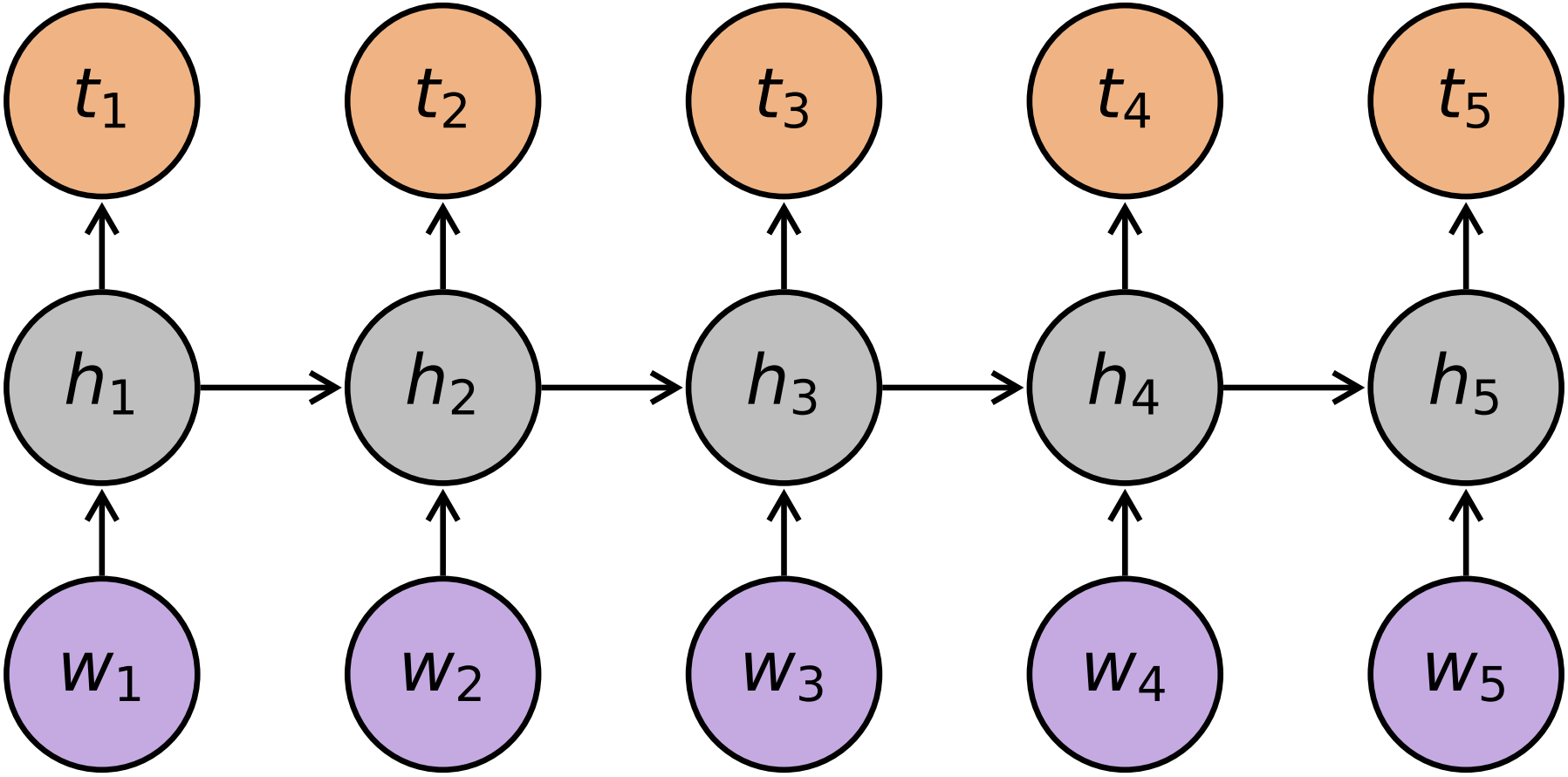
Sequence Labelling



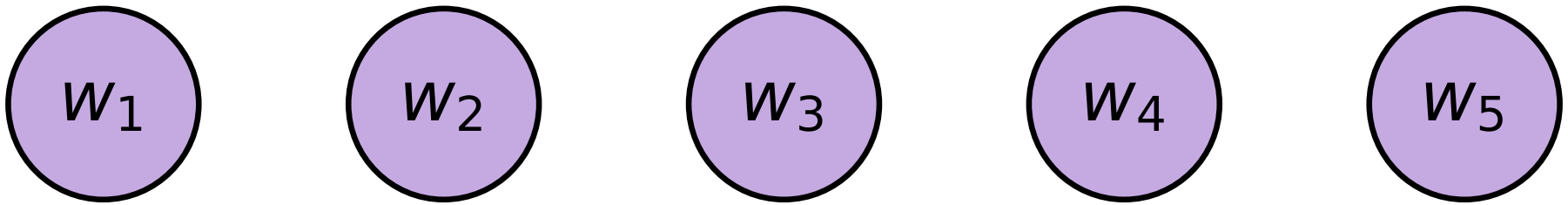
Convolutional Neural Net



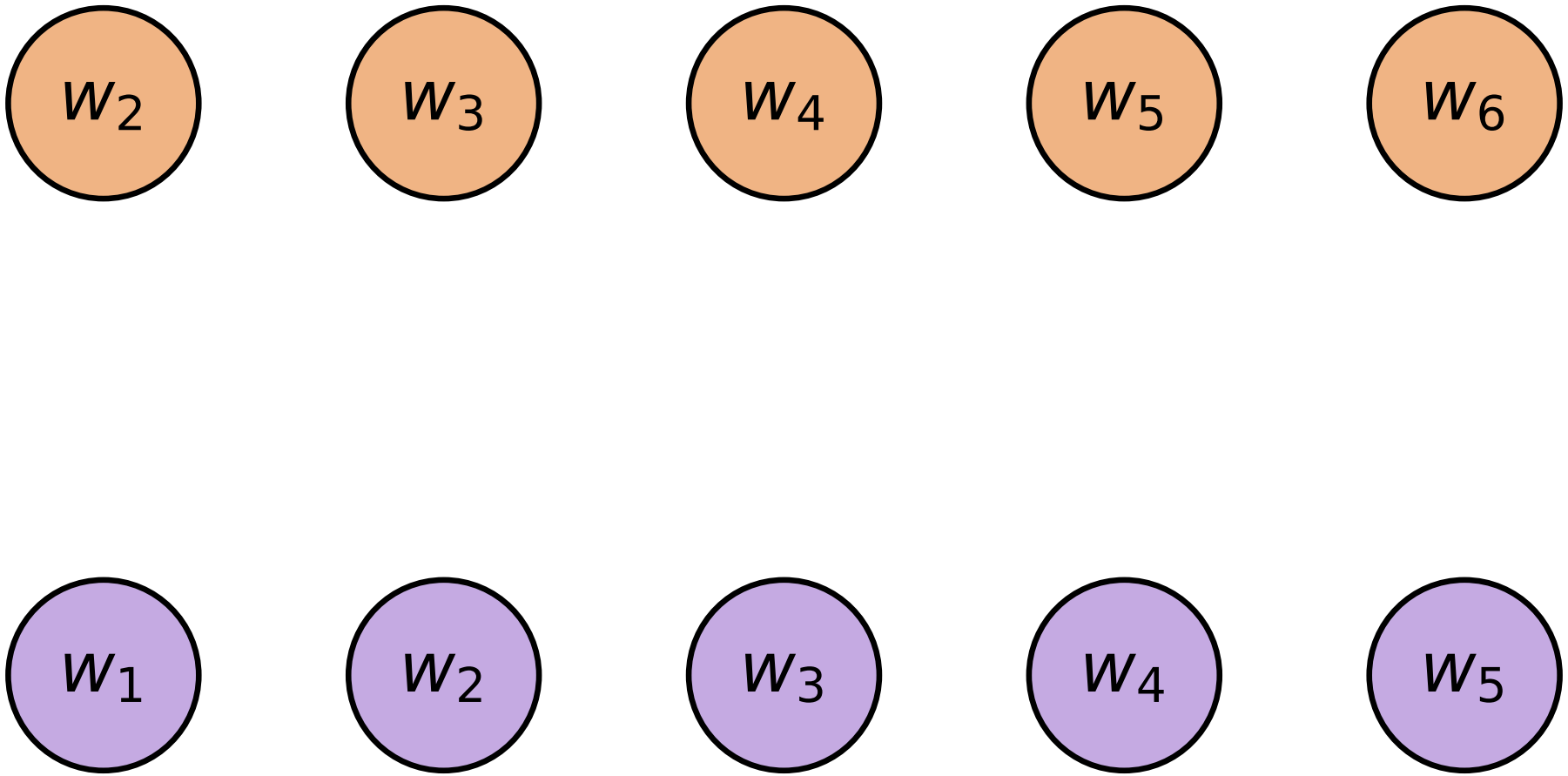
Recurrent Neural Net



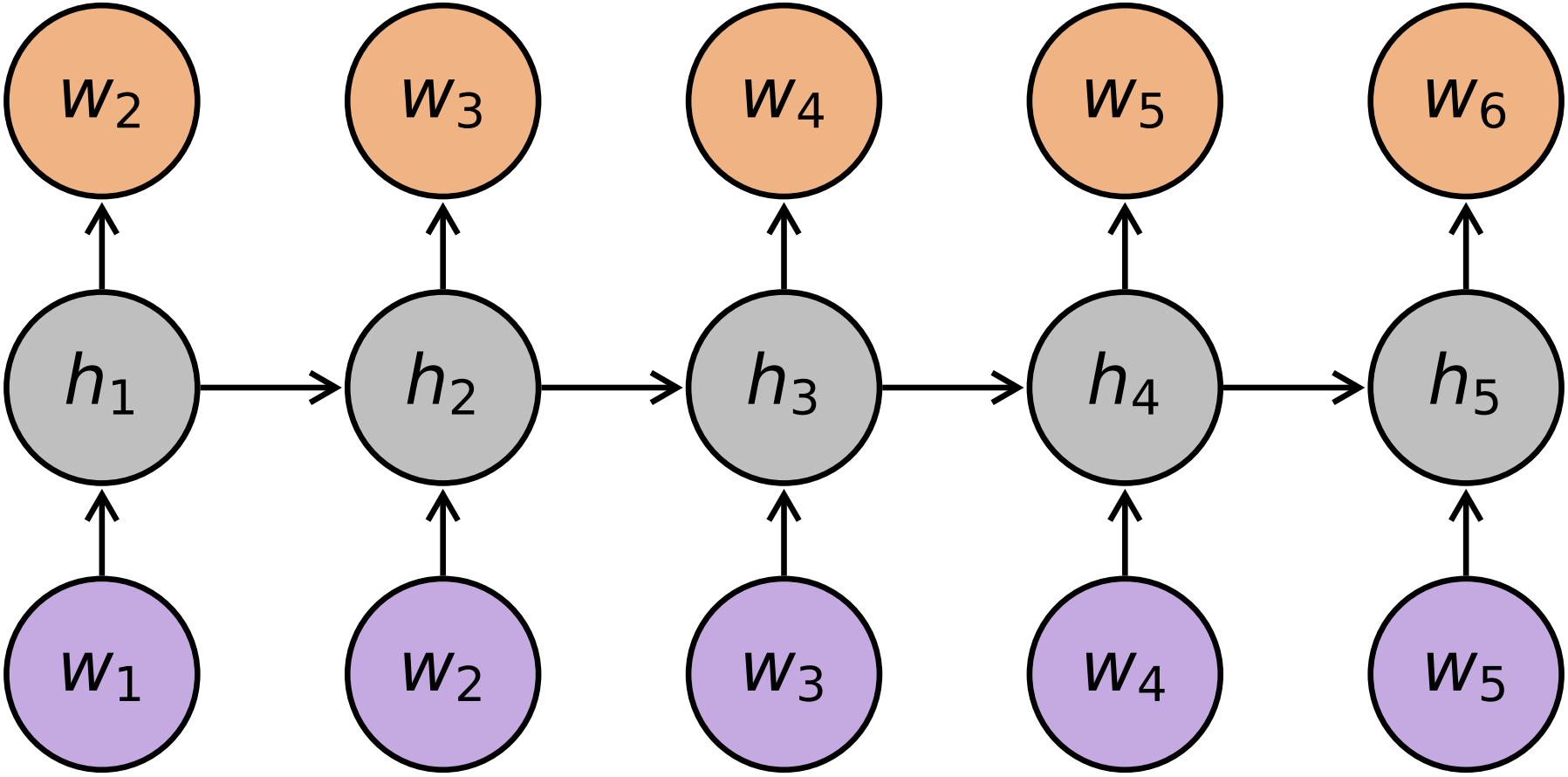
Language Modelling



Language Modelling



Language Modelling



Inference and Training

- Defined for fast inference
  - No beam search / dynamic programming

Inference and Training

- Defined for fast inference
  - No beam search / dynamic programming
- Train with gradient descent
  - Backpropagation: efficient chain rule

Short-Term Memory

- “Vanilla” RNNs, in ideal case:
  - Can remember long history

Short-Term Memory

- “Vanilla” RNNs, in ideal case:
  - Can remember long history
- “Vanilla” RNNs, in practice:
  - Very forgetful

Exploding/Vanishing Gradients

- Gradient descent for vanilla RNNs:
  - Backprop through recurrent connections

Exploding/Vanishing Gradients

- Gradient descent for vanilla RNNs:
  - Backprop through recurrent connections
  - Repeated multiplications

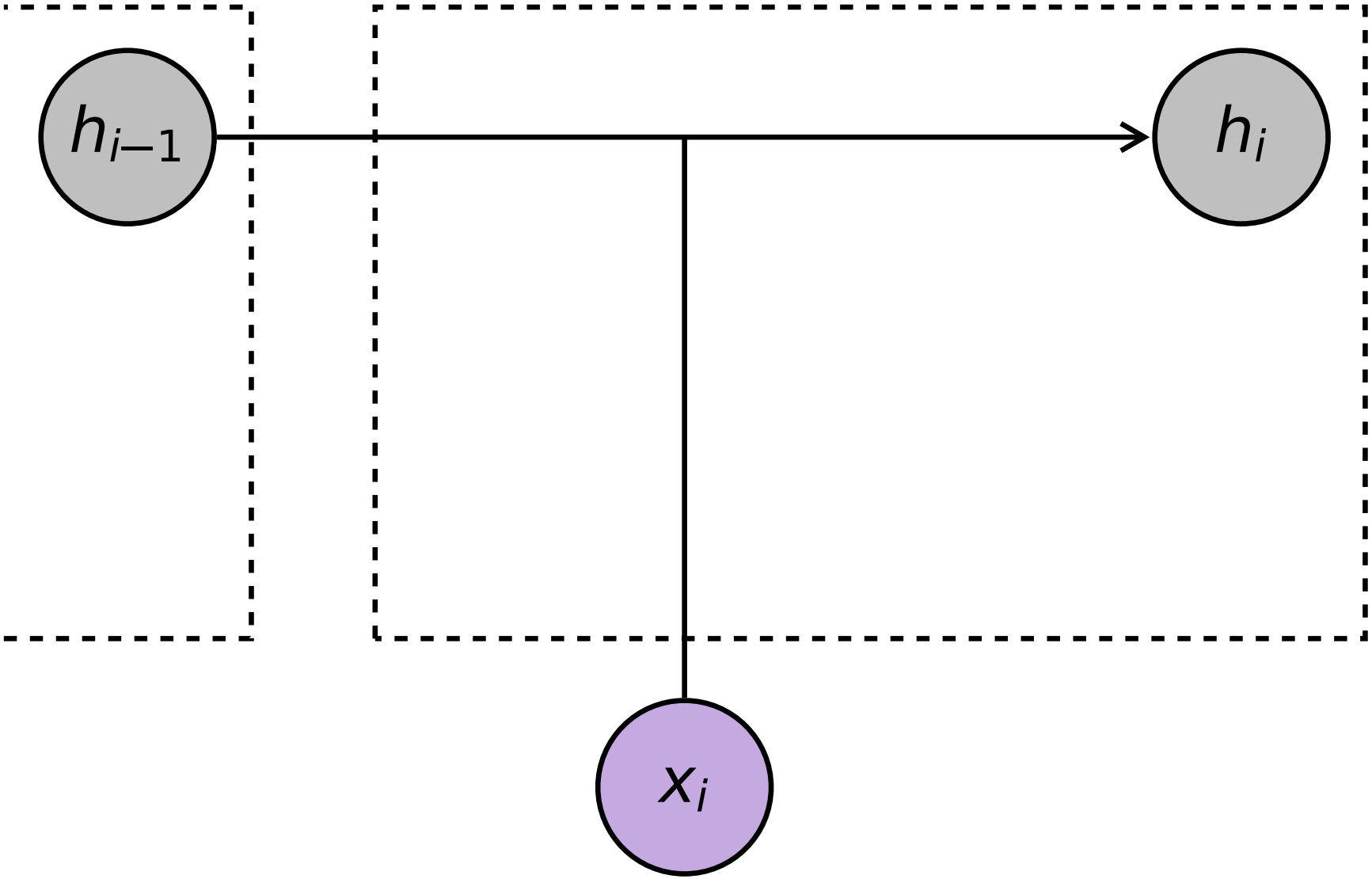
Exploding/Vanishing Gradients

- Gradient descent for vanilla RNNs:
  - Backprop through recurrent connections
  - Repeated multiplications
  - Exponential increase/decrease

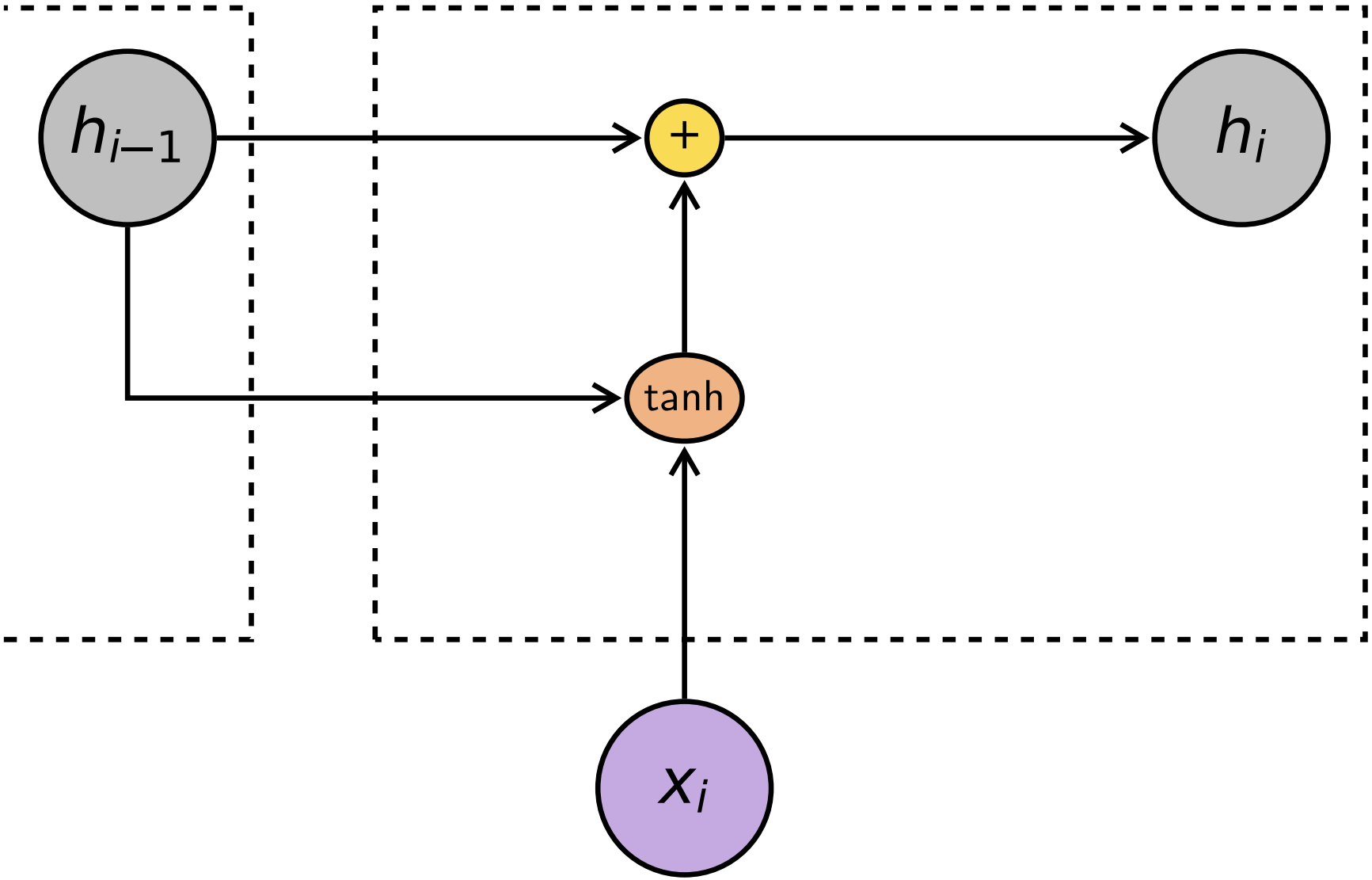
Exploding/Vanishing Gradients

- Gradient descent for vanilla RNNs:
  - Backprop through recurrent connections
  - Repeated multiplications
  - Exponential increase/decrease
- Long Short-Term Memory (LSTM):
  - Avoid repeated multiplications

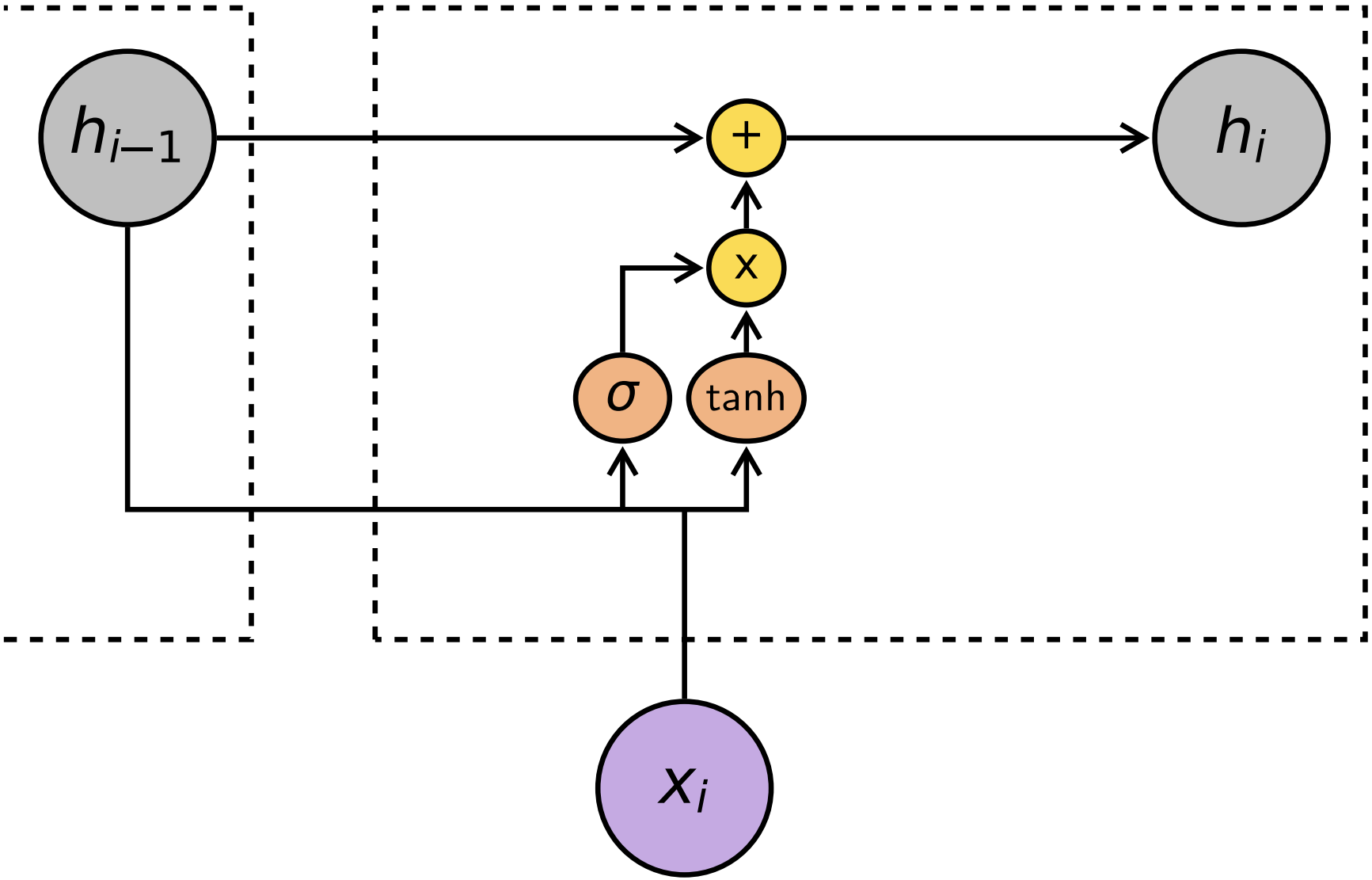
Long Short-Term Memory



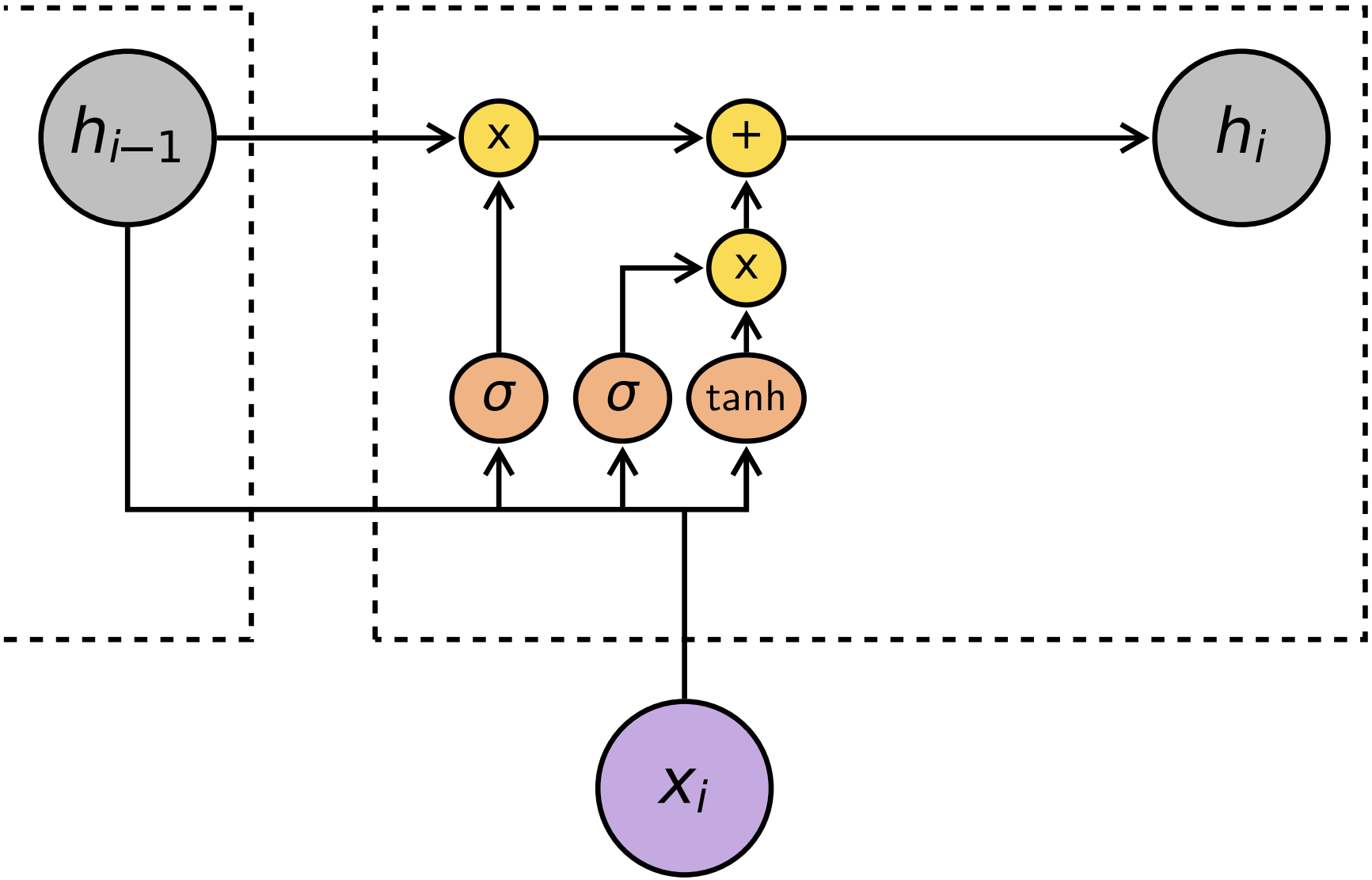
Long Short-Term Memory



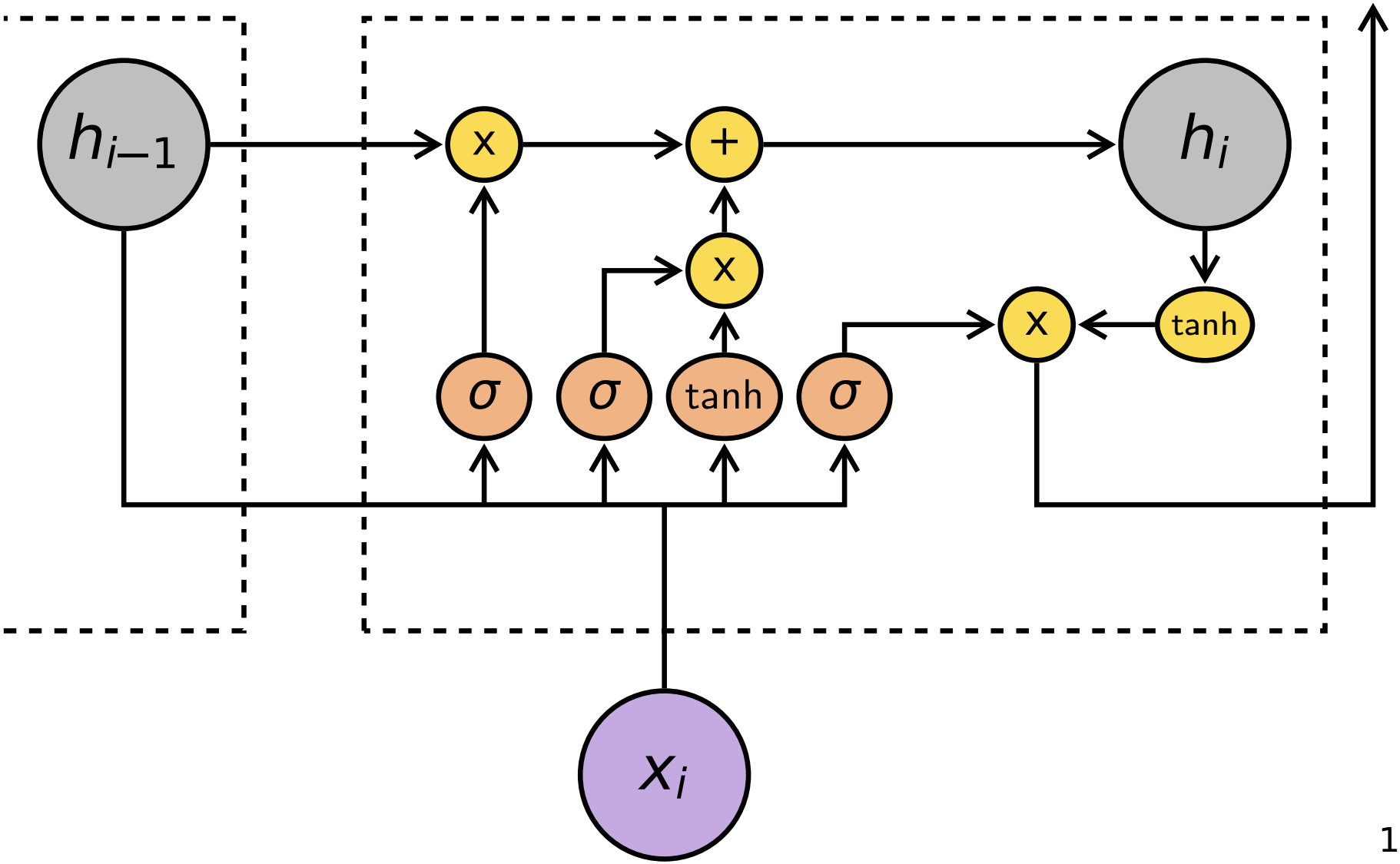
Long Short-Term Memory



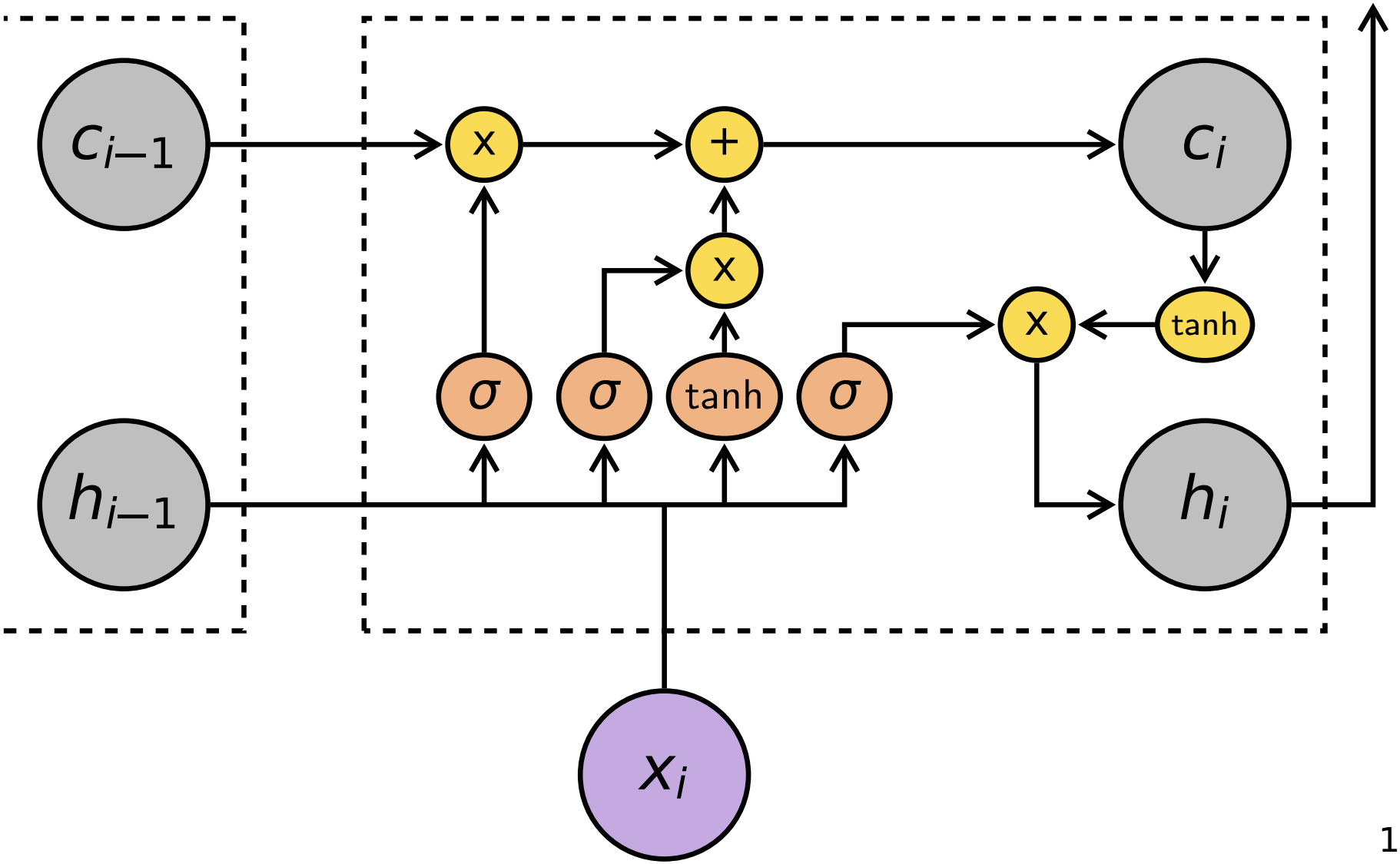
Long Short-Term Memory



Long Short-Term Memory



Long Short-Term Memory



The Devil's in the Hyperparameters

- A lot of details...
  - Activation function
  - Dimensionality
  - Descent algorithm
  - Learning rate
  - Batch size
  - Regularisation
  - No. training epochs
  - Initialisation
  - etc...

“Black Boxes”

- Interpretation of features?

“Black Boxes”

- Interpretation of features?
- No pre-defined interpretation (unlike e.g. LDA)

“Black Boxes”

- Interpretation of features?
- No pre-defined interpretation (unlike e.g. LDA)
- Can measure correlations
- Can measure effects on predictions

“Black Boxes”

- Interpretation of features?
- No pre-defined interpretation (unlike e.g. LDA)
- Can measure correlations
- Can measure effects on predictions
- Open area of research...

Summary

- Feedforward networks
  - CNNs
  - RNNs
  - LSTMs
- Hyperparameter tuning
- Challenge: interpreting a model