Statistical Natural Language Processing [COMP0087]

Introduction to neural networks and backpropagation

Vasileios Lampos

Computer Science, UCL

lampos.net

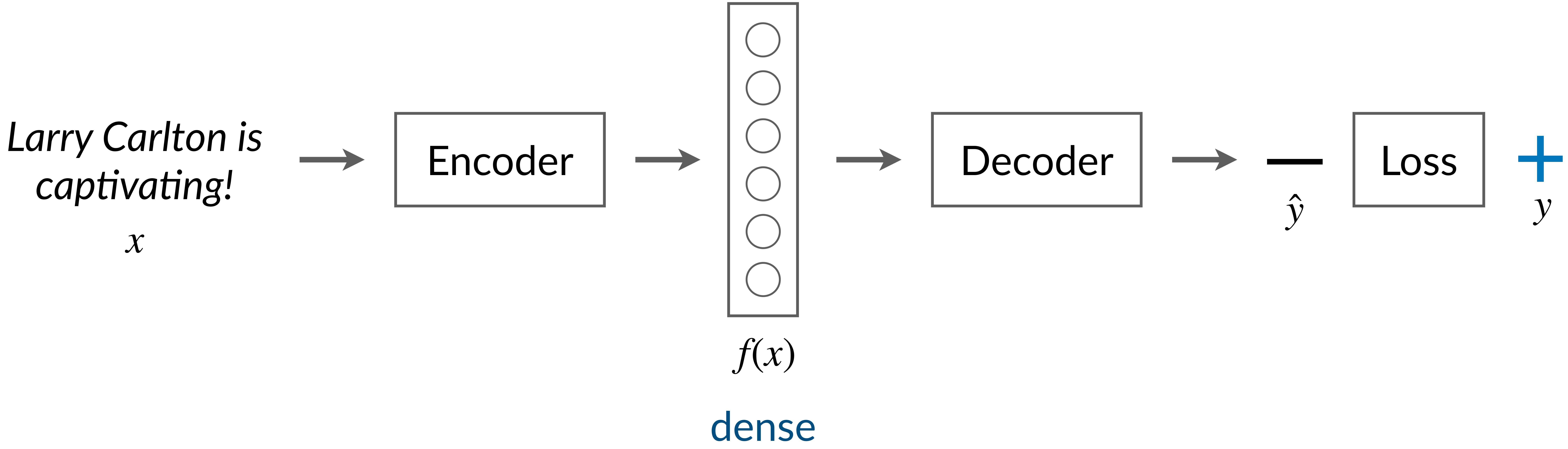
About this lecture

- In this lecture:
  - Introductory neural network concepts
  - Inference and training (*backpropagation*) with feedforward neural networks

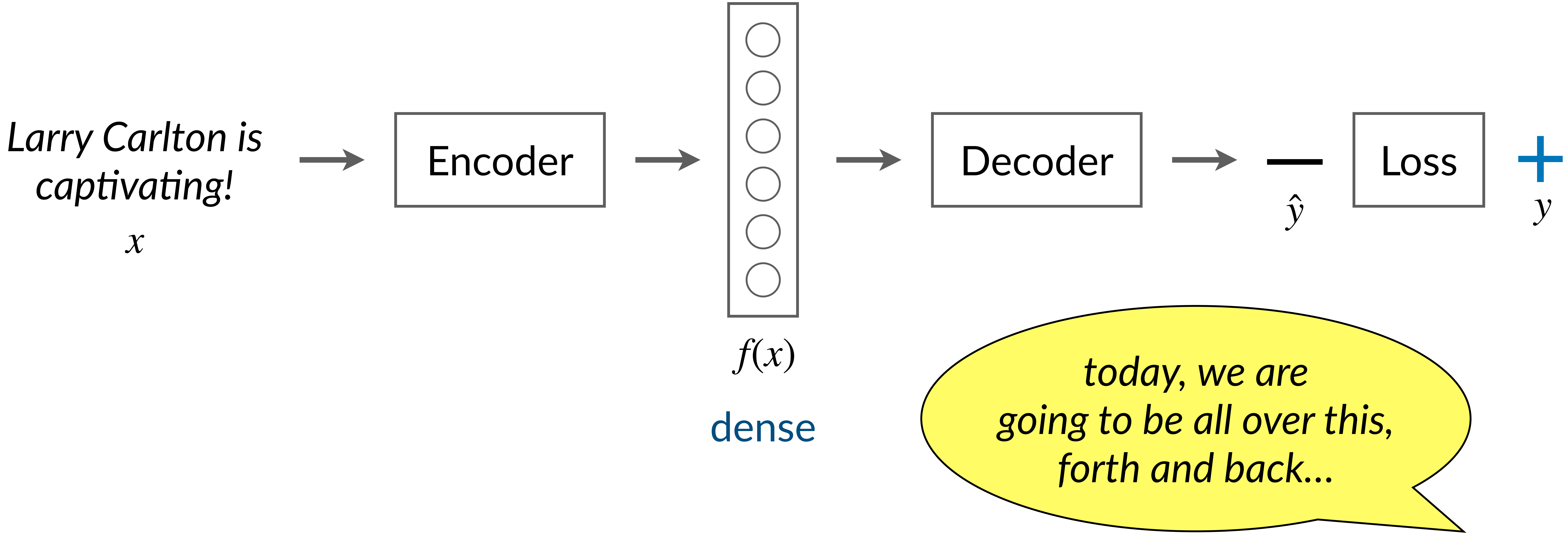
- **Reading / Lecture partly based on:** Chapter 7 of "*Speech and Language Processing*" (SLP) by Jurafsky and Martin (2023) — web.stanford.edu/~jurafsky/slp3/

- For those of you who want to have the slides in front of them during the lecture, there is a clipped / early version at lampos.net/teaching (*non clipped / slightly refined version will be added after the lecture*)

The NLP view (*for this lecture*)

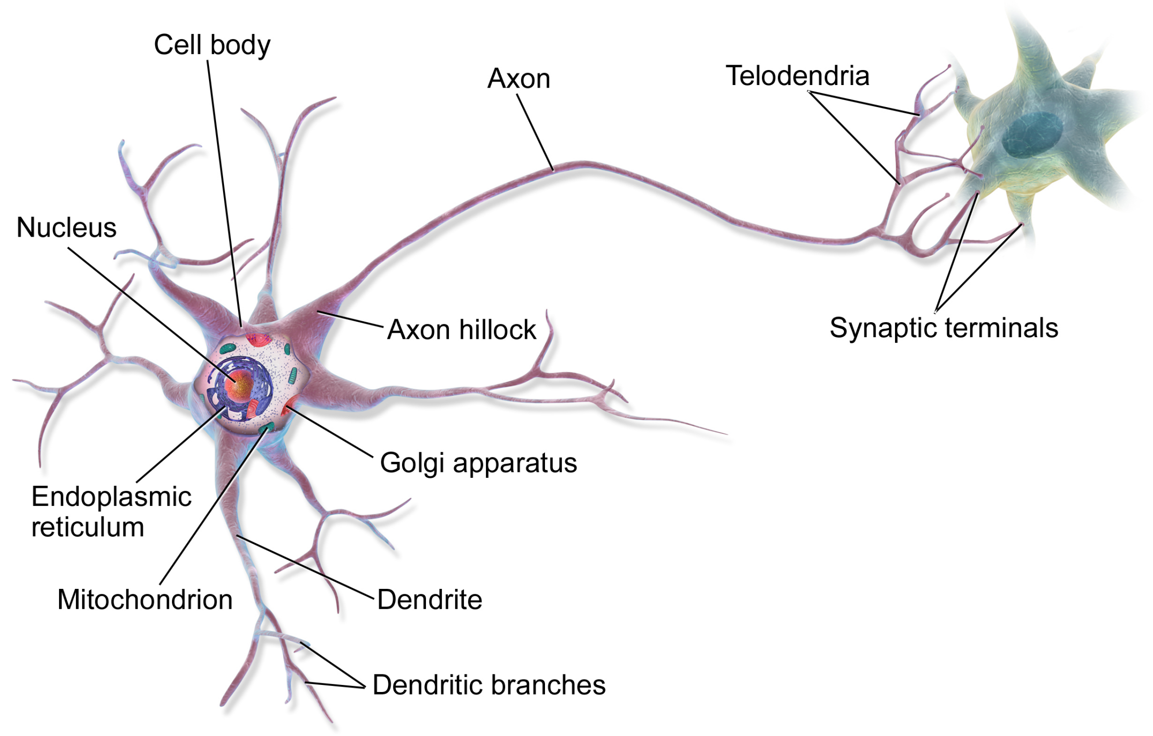


The NLP view (*for this lecture*)

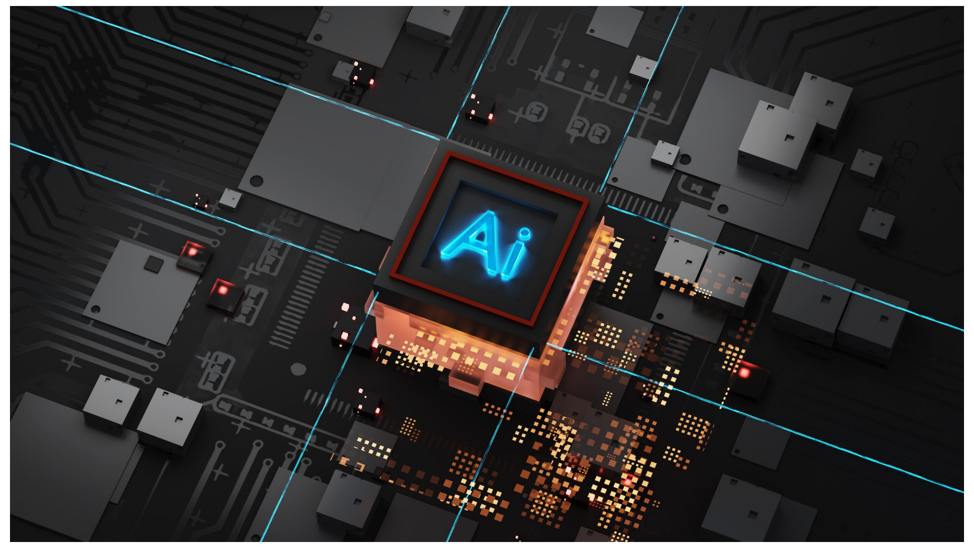


Artificial neural networks — A few introductory remarks

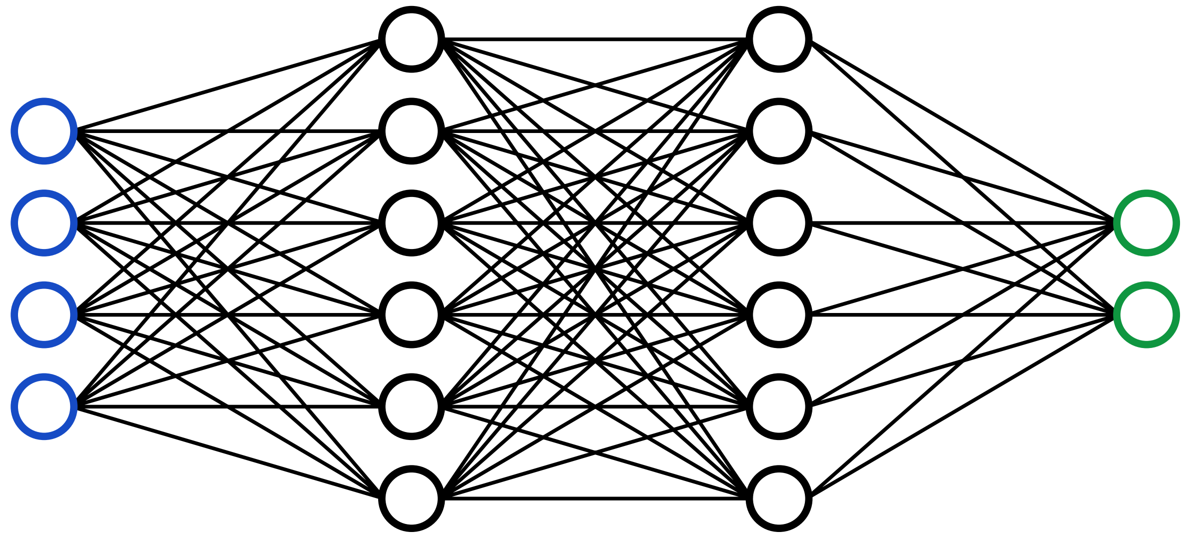
- Artificial Neural Networks (**NNs**) $\neq$ biological neural networks *until we actually obtain a* ***complete*** *understanding about how the human brain operates!*



- **NNs are powerful learning functions / universal approximators**, e.g. standard multi-layer feedforward networks with as few as one hidden layer are capable of approximating any (*Borel measurable*) function — and we are aware of this for almost 40 years (Hornik, Stinchcombe and White, 1989, doi.org/10.1016/0893-6080(89)90020-8)



- **NB**: Good understanding of **logistic regression**? Easy to understand today's lecture and fundamentals about NNs in a few seconds. *Otherwise it might take a few minutes.*

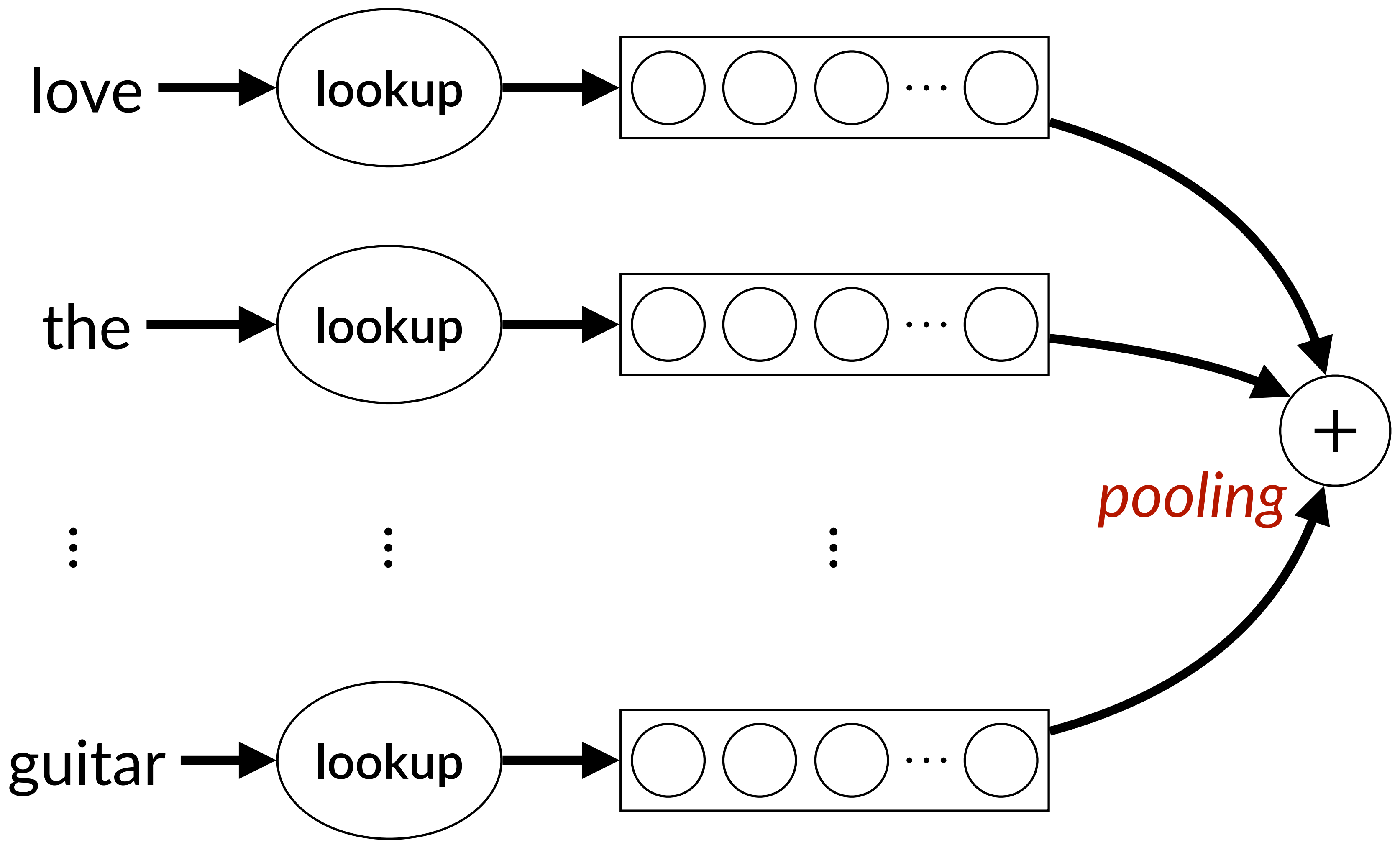


Background task — Sentiment classification

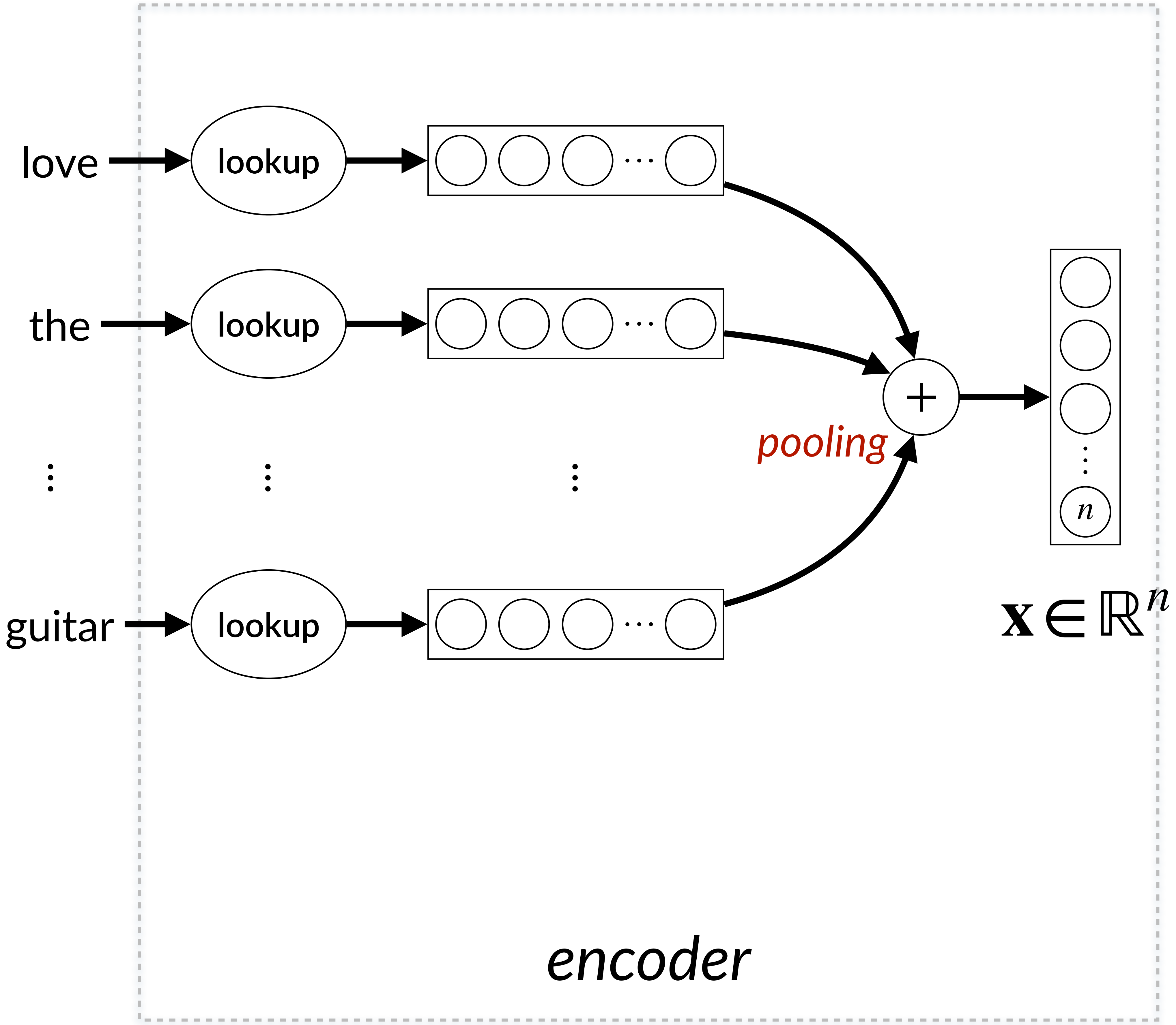
Sentiment?

Wow, I love the sound of this acoustic guitar!	$\longrightarrow$	$+$ *(positive)*
It was just another uneventful Marvel movie!	$\longrightarrow$	$-$ *(negative)*
Can't say I loved this performance, but I didn't dislike it either.	$\longrightarrow$	*neutral*

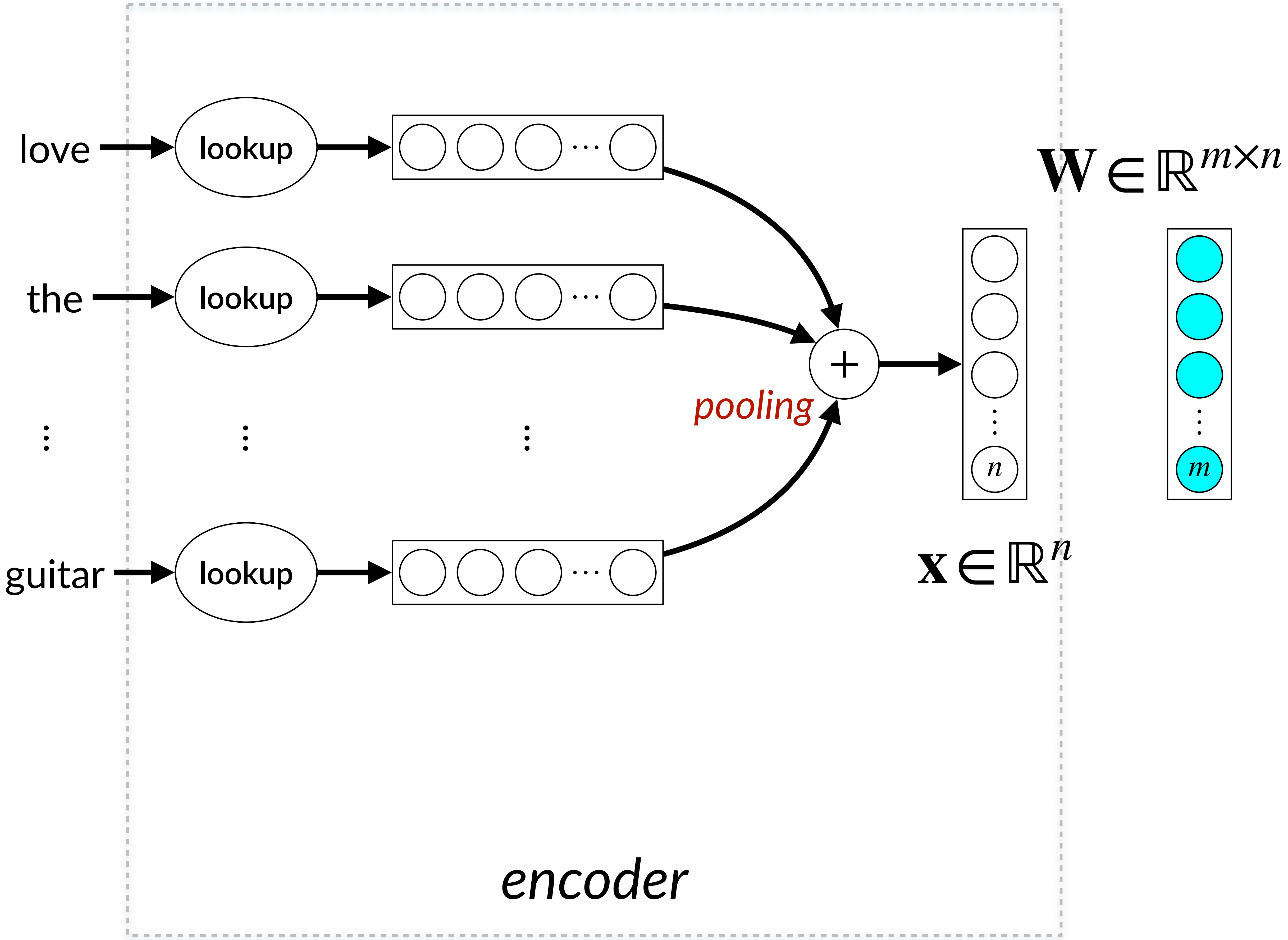
Neural network — A simplified encoder–decoder architecture



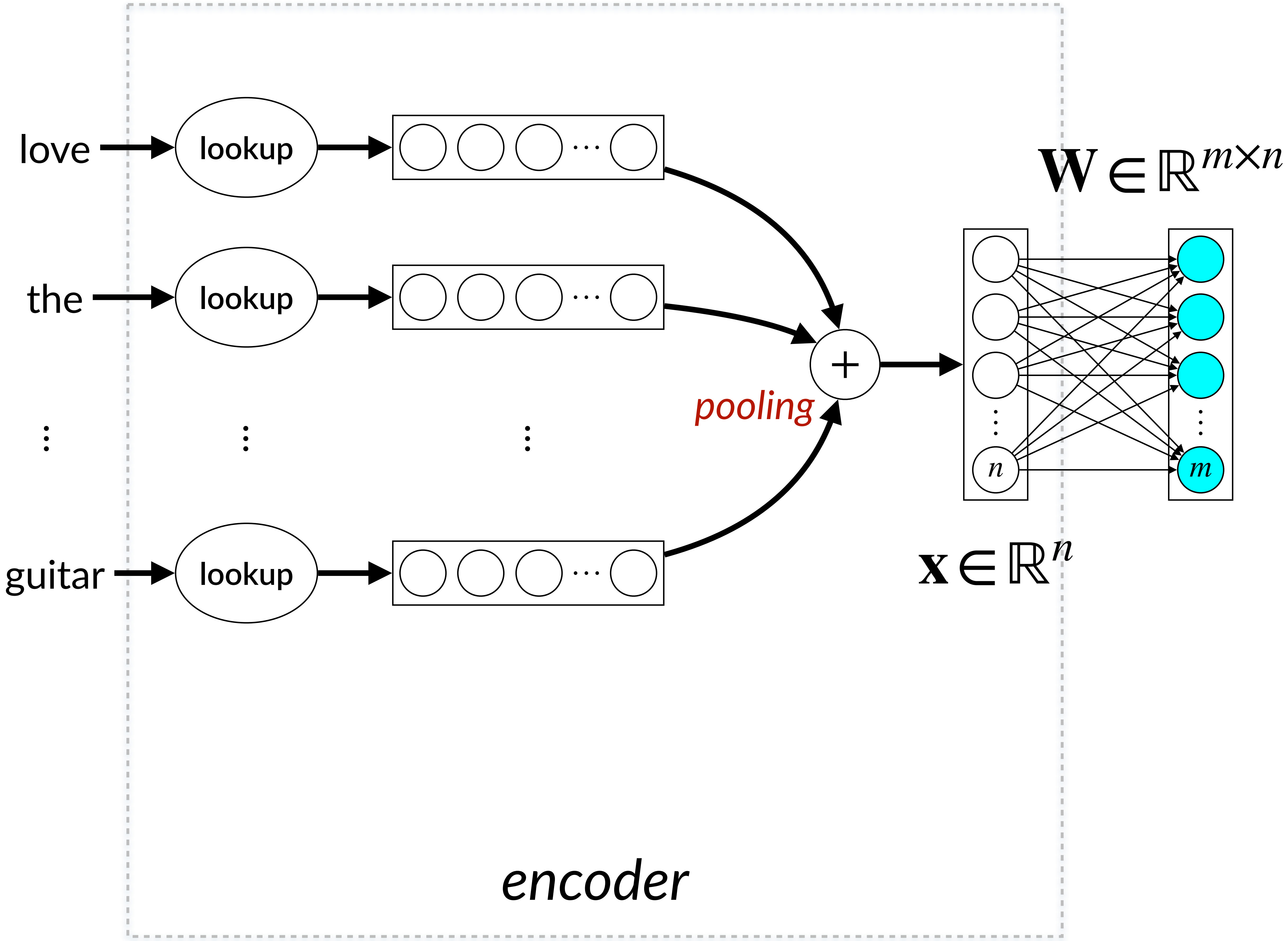
Neural network — A simplified encoder–decoder architecture



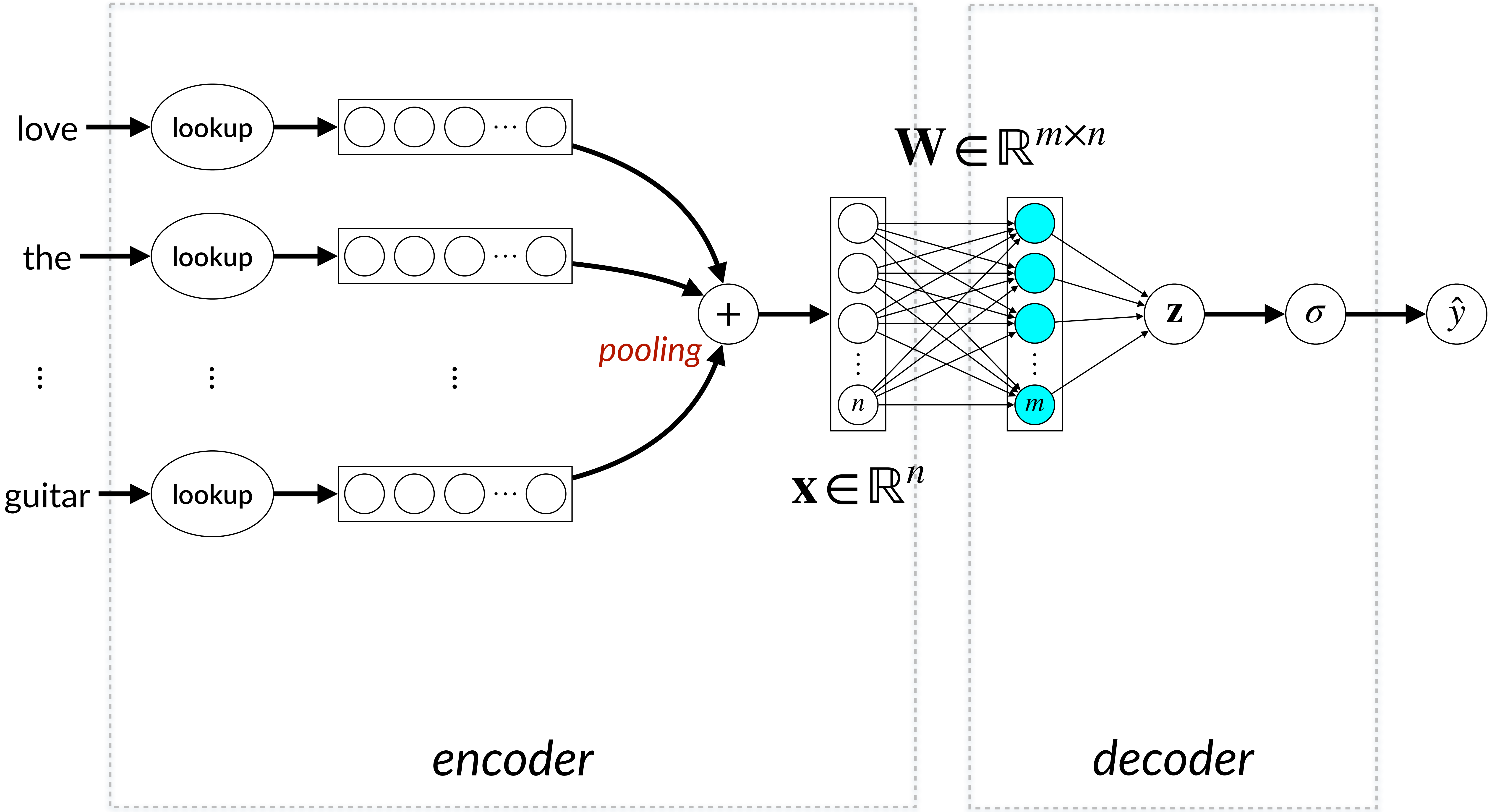
Neural network — A simplified encoder—decoder architecture



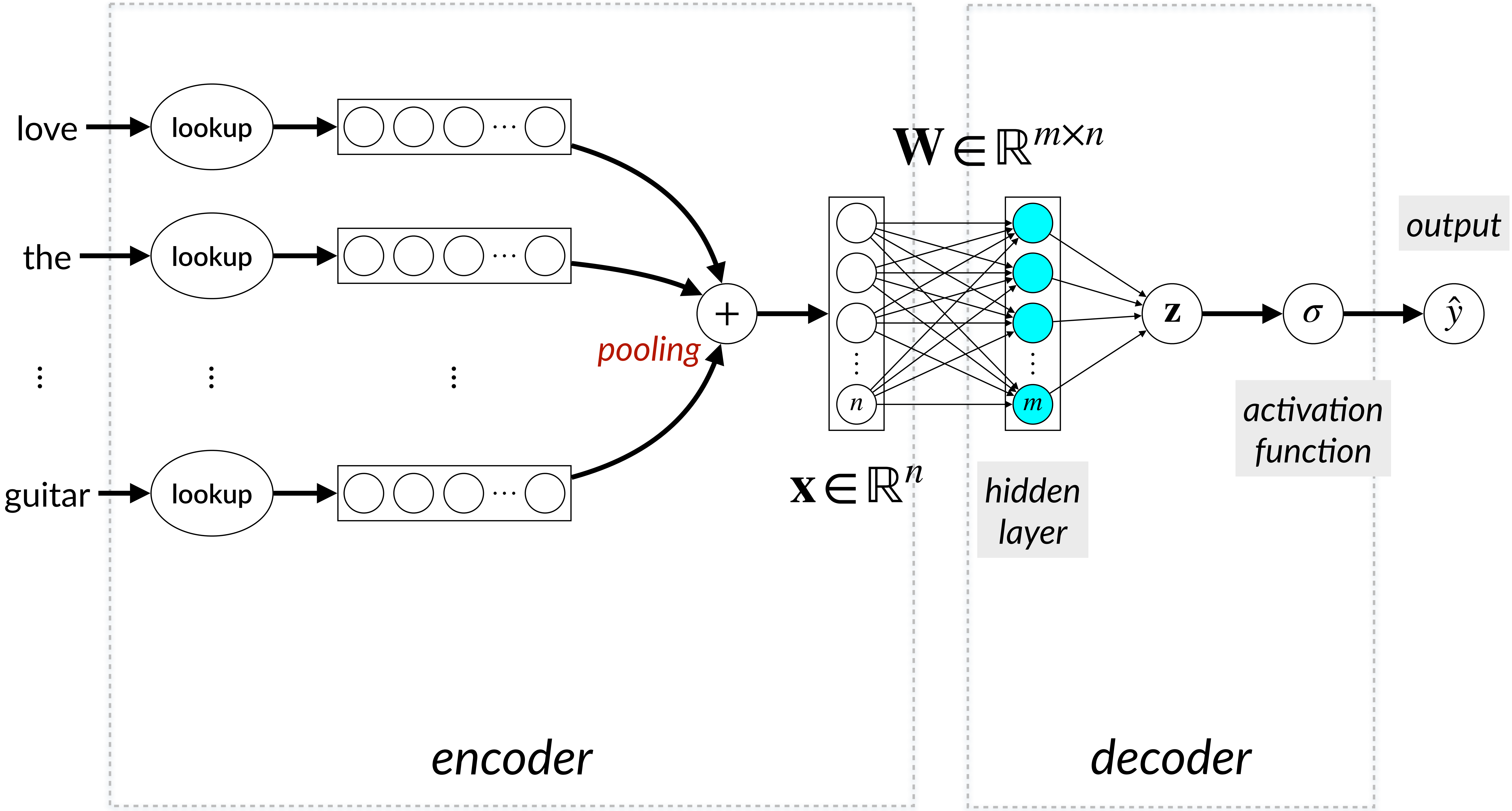
Neural network — A simplified encoder–decoder architecture



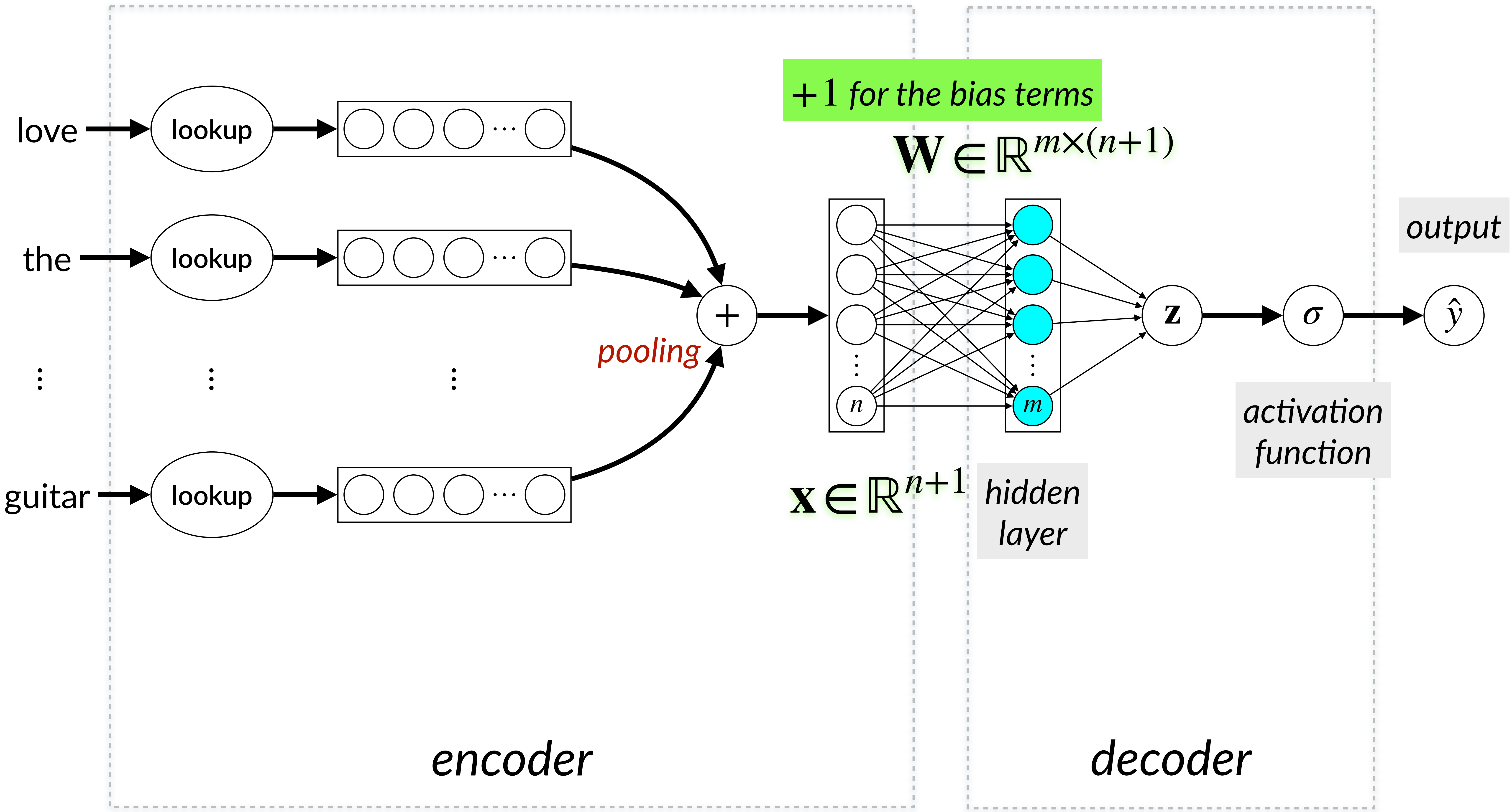
Neural network — A simplified encoder—decoder architecture



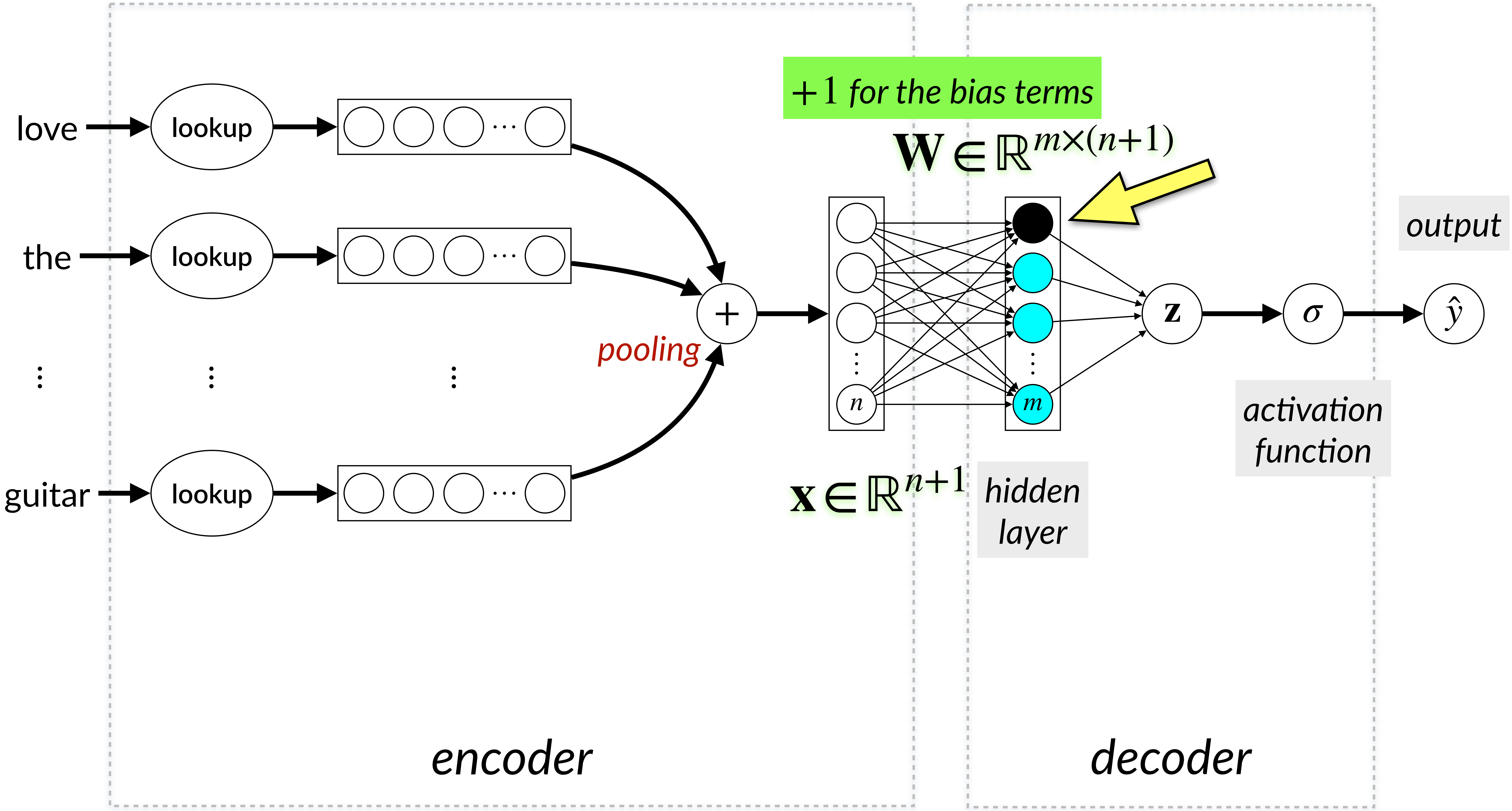
Neural network — A simplified encoder—decoder architecture

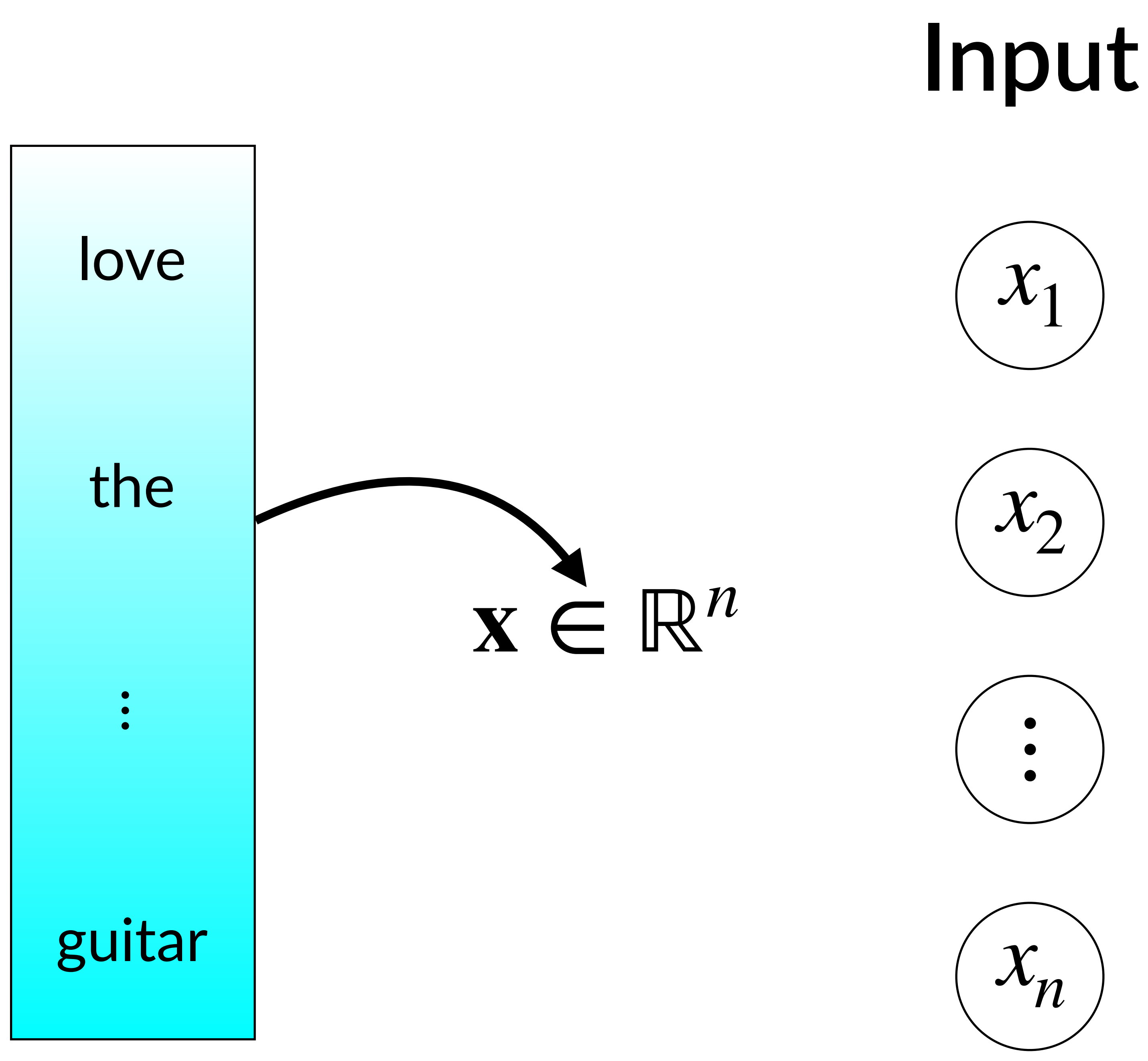


Neural network — A simplified encoder—decoder architecture



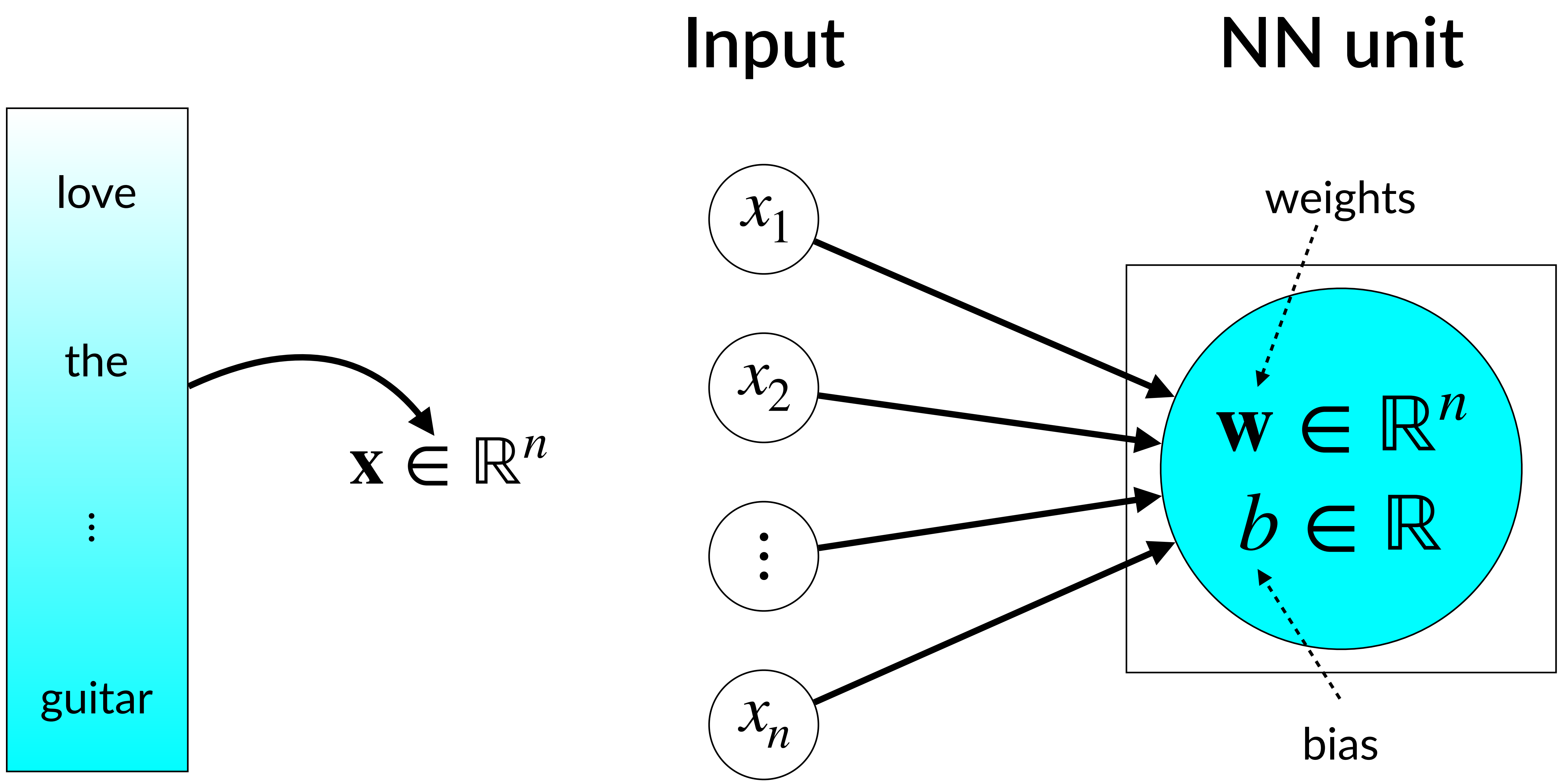
Neural network — A simplified encoder—decoder architecture





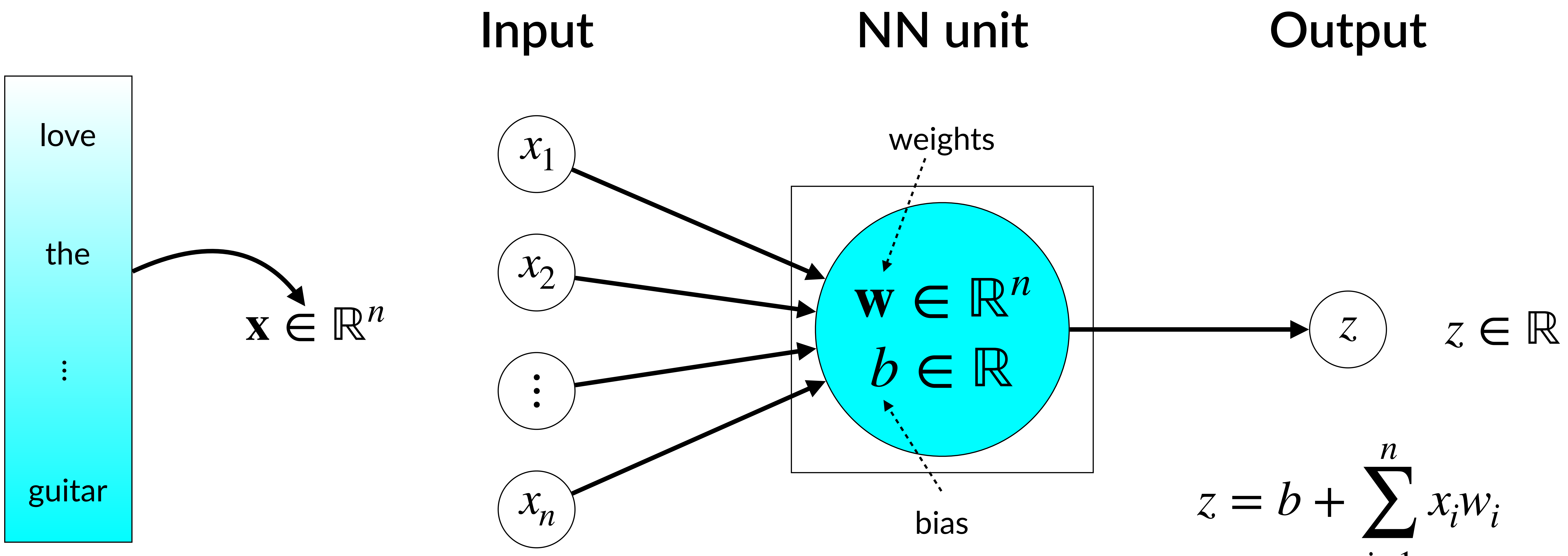
$$\mathbf{x} = [x_1 \quad \cdots \quad x_n]$$

One neural computation unit (*of a hidden layer*)



$$\mathbf{x} = [x_1 \quad \cdots \quad x_n] \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

One neural computation unit (*of a hidden layer*)

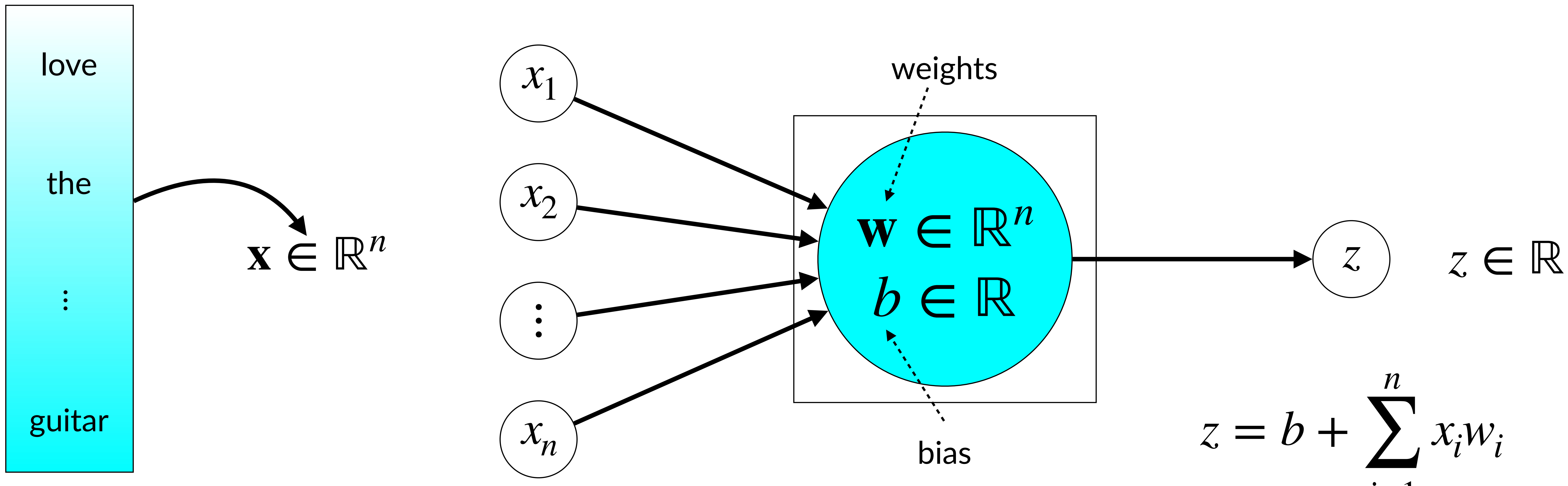


$$z = b + \sum_{i=1}^{n} x_i w_i$$

$$\mathbf{x} = [x_1 \quad \cdots \quad x_n] \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

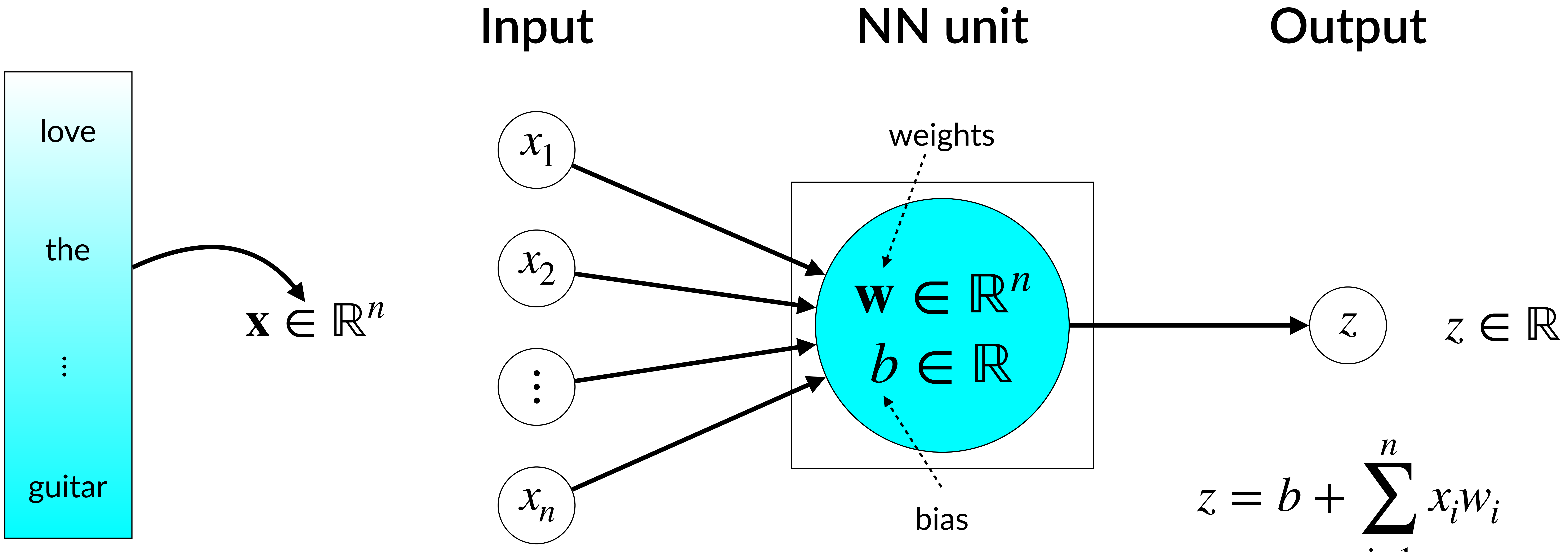
One neural computation unit (*of a hidden layer*)

Input **NN unit** **Output**



$$z = b + \sum_{i=1}^{n} x_i w_i$$

$$\mathbf{x} = [x_1 \quad \cdots \quad x_n] \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$



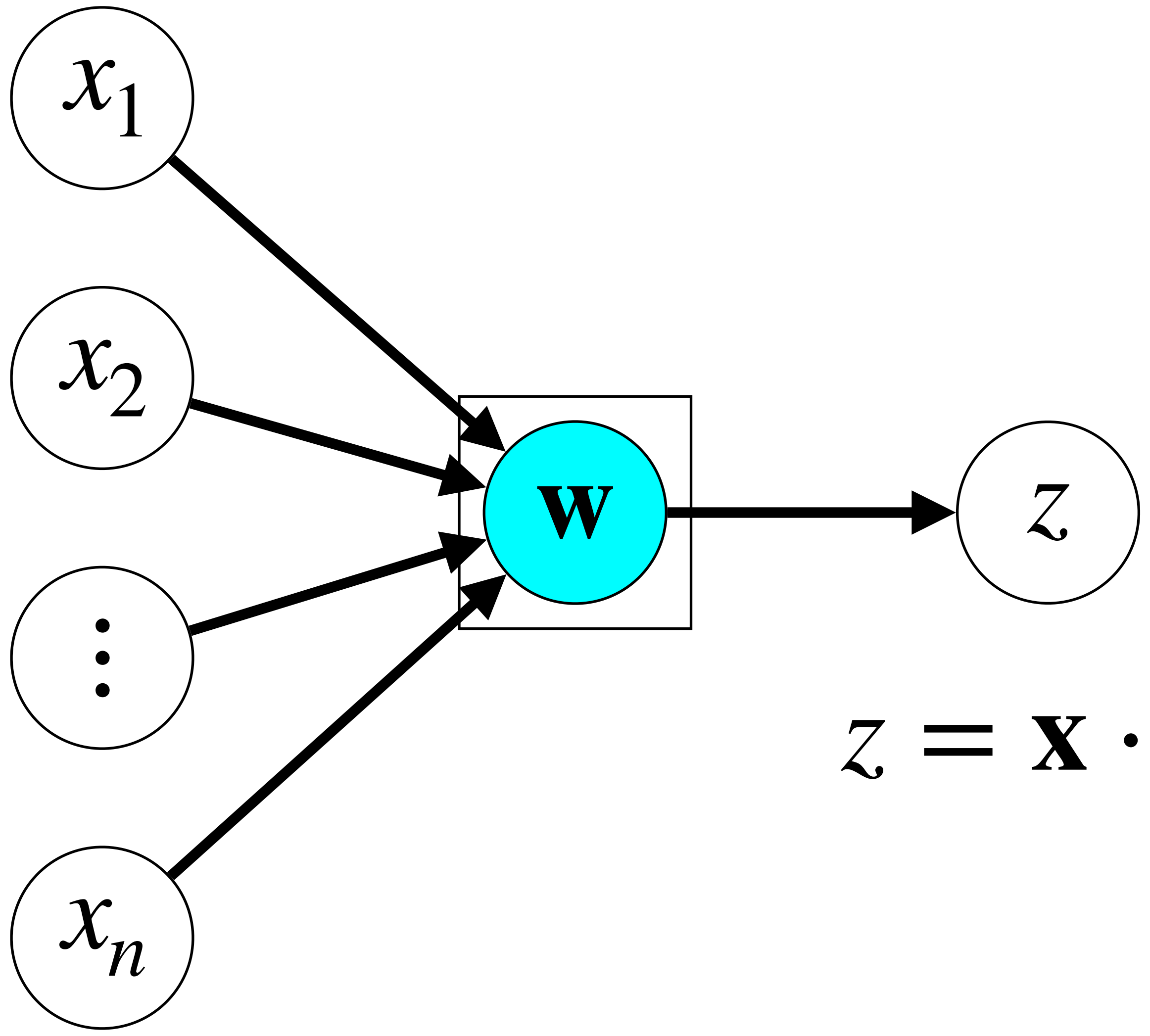
$$z = b + \sum_{i=1}^{n} x_i w_i$$

$$\mathbf{x} = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

simplifying notation

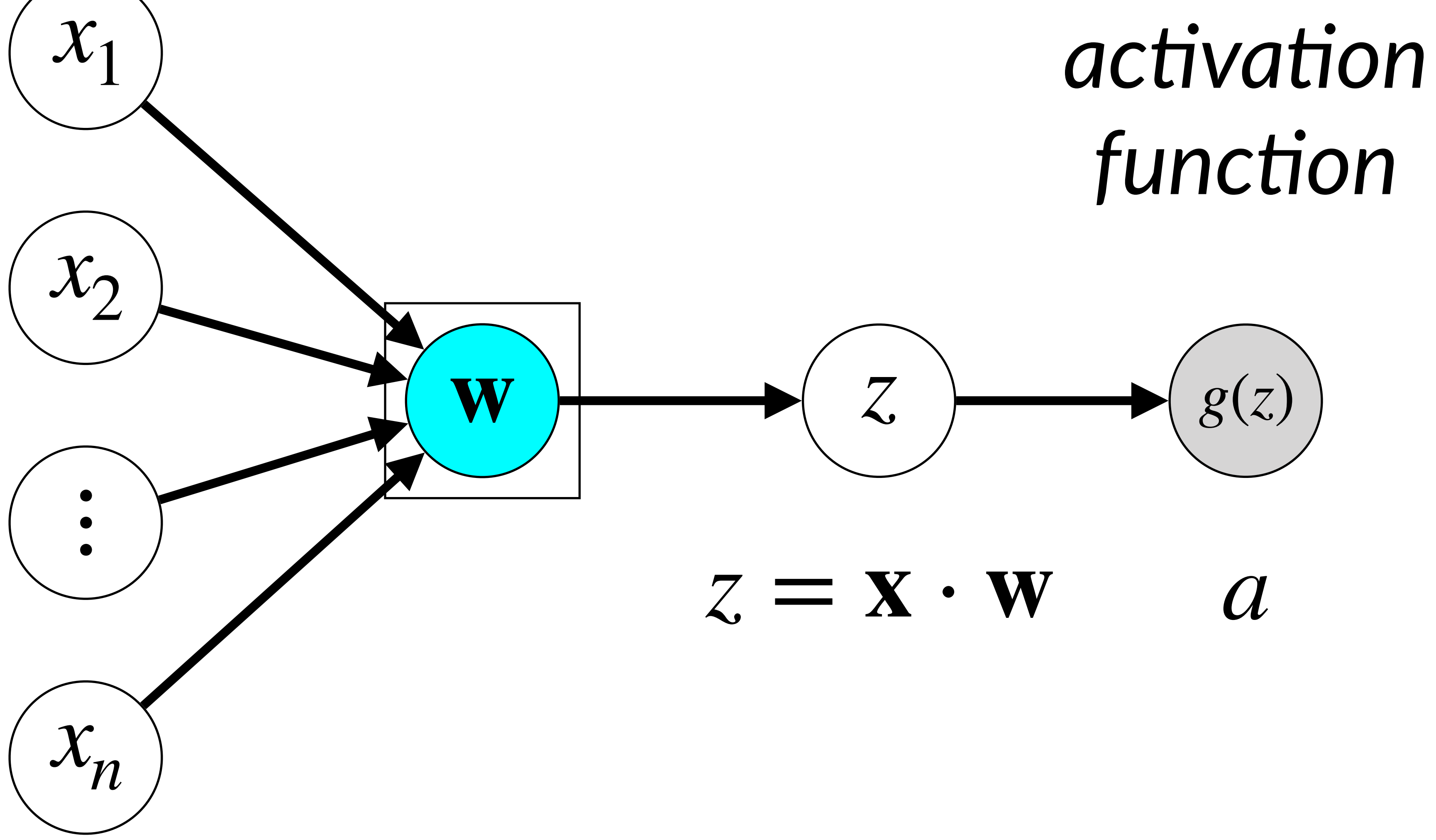
$$z = \underbrace{\begin{bmatrix} 1 & x_1 & \dots & x_n \end{bmatrix} \cdot \begin{bmatrix} b \\ w_1 \\ \vdots \\ w_n \end{bmatrix}}_{\mathbf{x} \cdot \mathbf{w}}$$

Activation functions

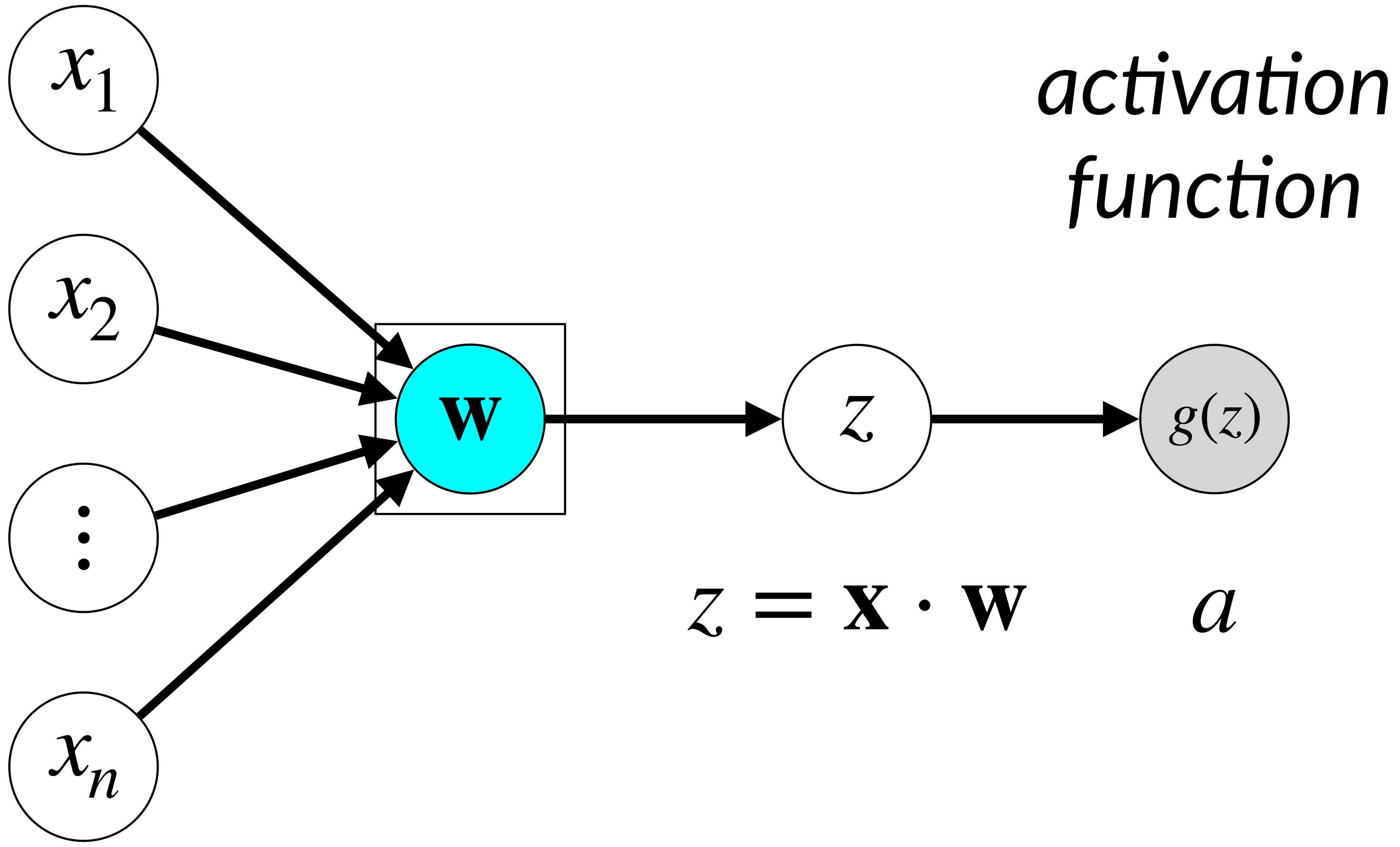


$$z = \mathbf{x} \cdot \mathbf{w}$$

Activation functions



Activation functions



$$\sigma(z) = \frac{1}{1 + \exp(-z)} = a$$

sigmoid logistic

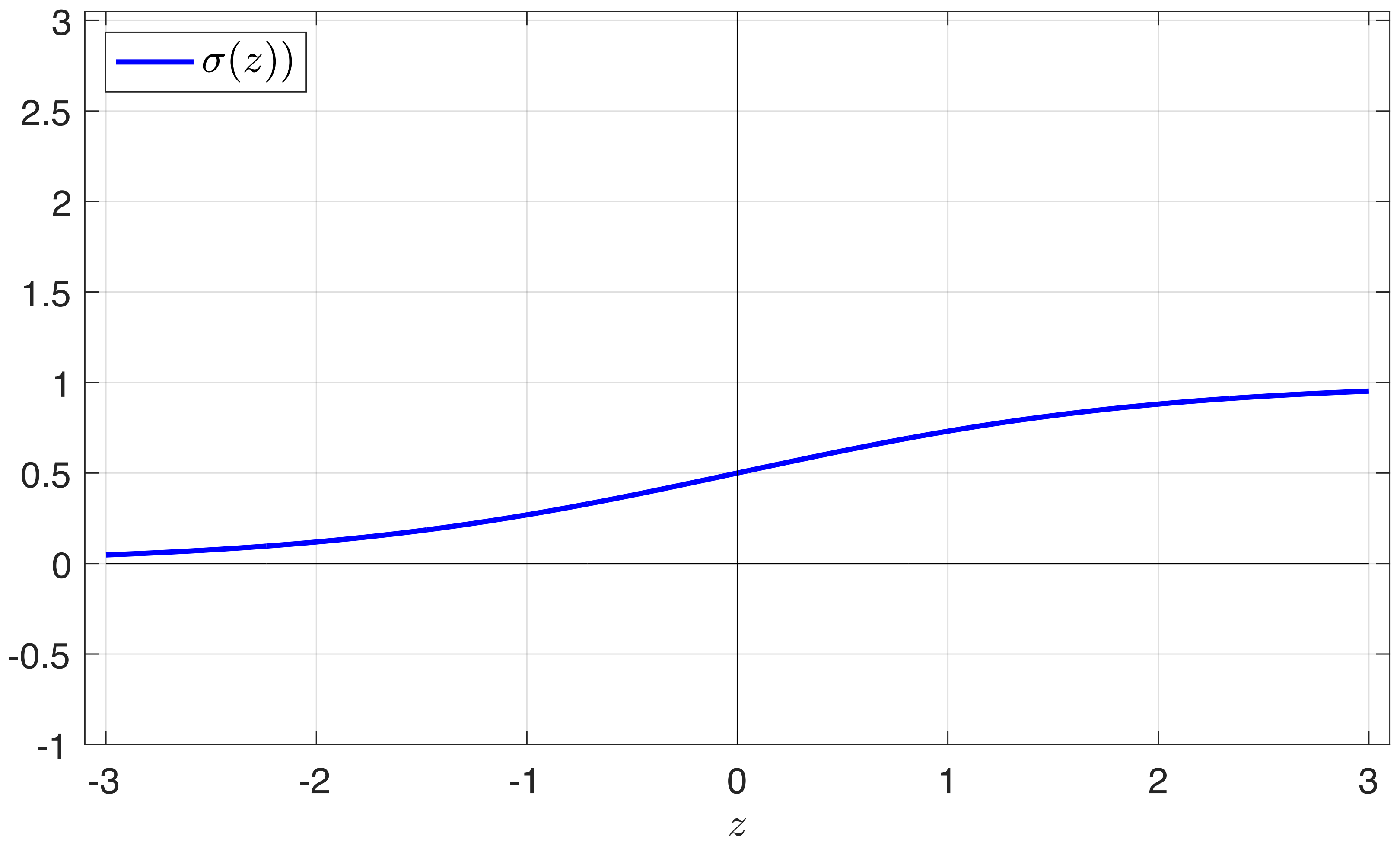
$$\tanh(z) = \frac{\exp(z) - \exp(-z)}{\exp(z) + \exp(-z)} = a$$

hyperbolic tangent

$$\mathrm{ReLU}(z) = \max\,(z, 0) = a$$

rectified linear unit

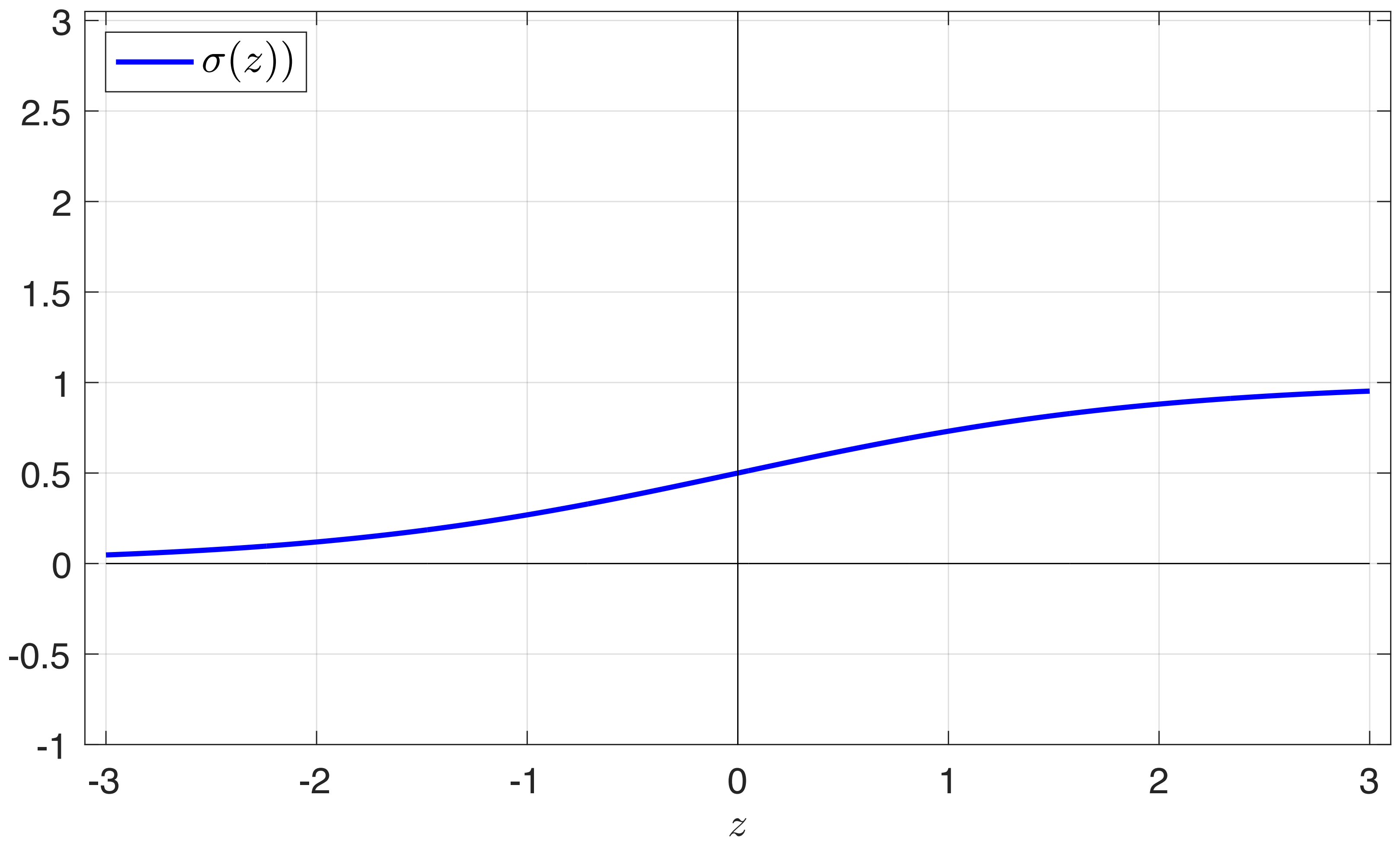
Activation functions

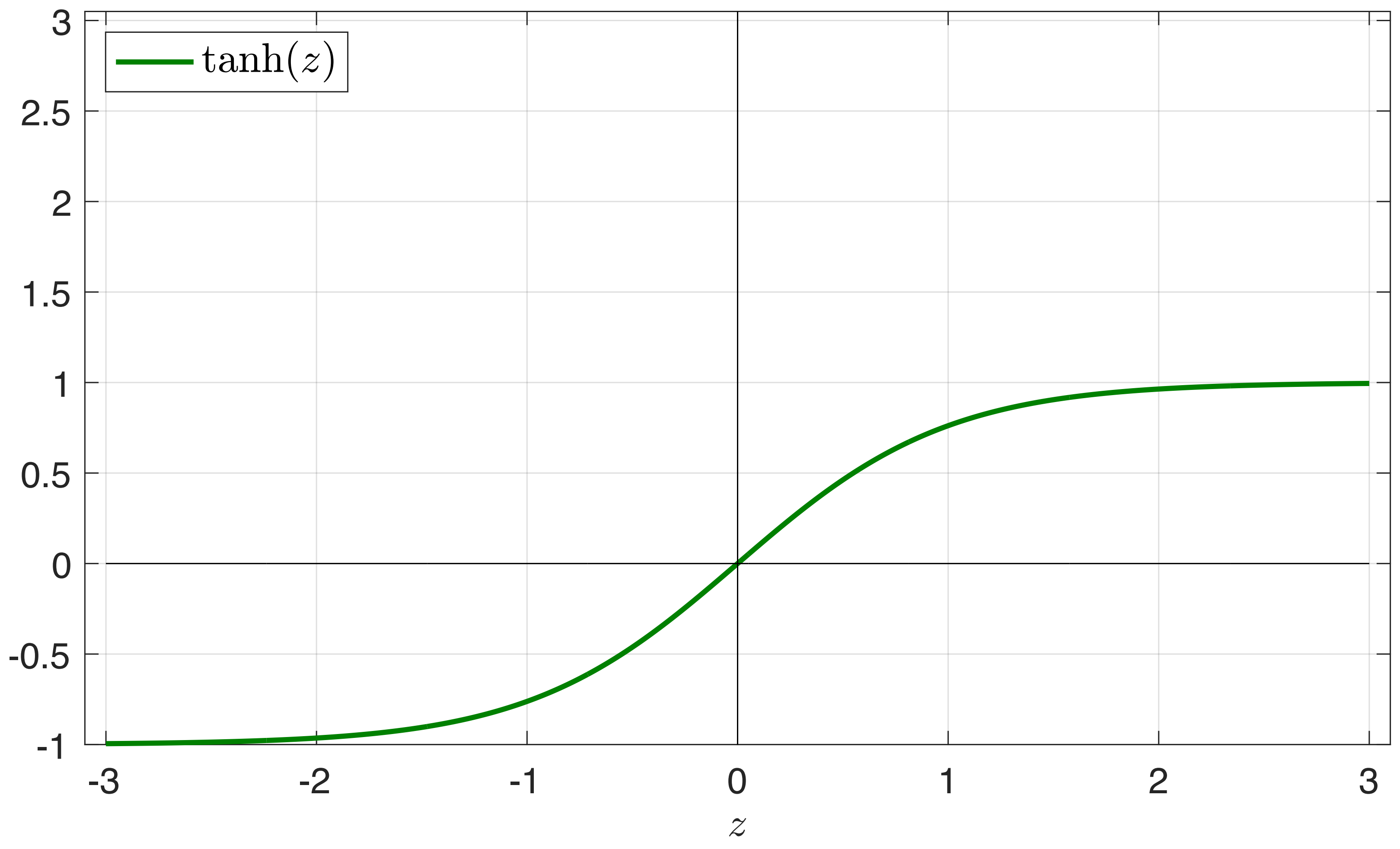


$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

$$\sigma \in (0,1)$$

Activation functions





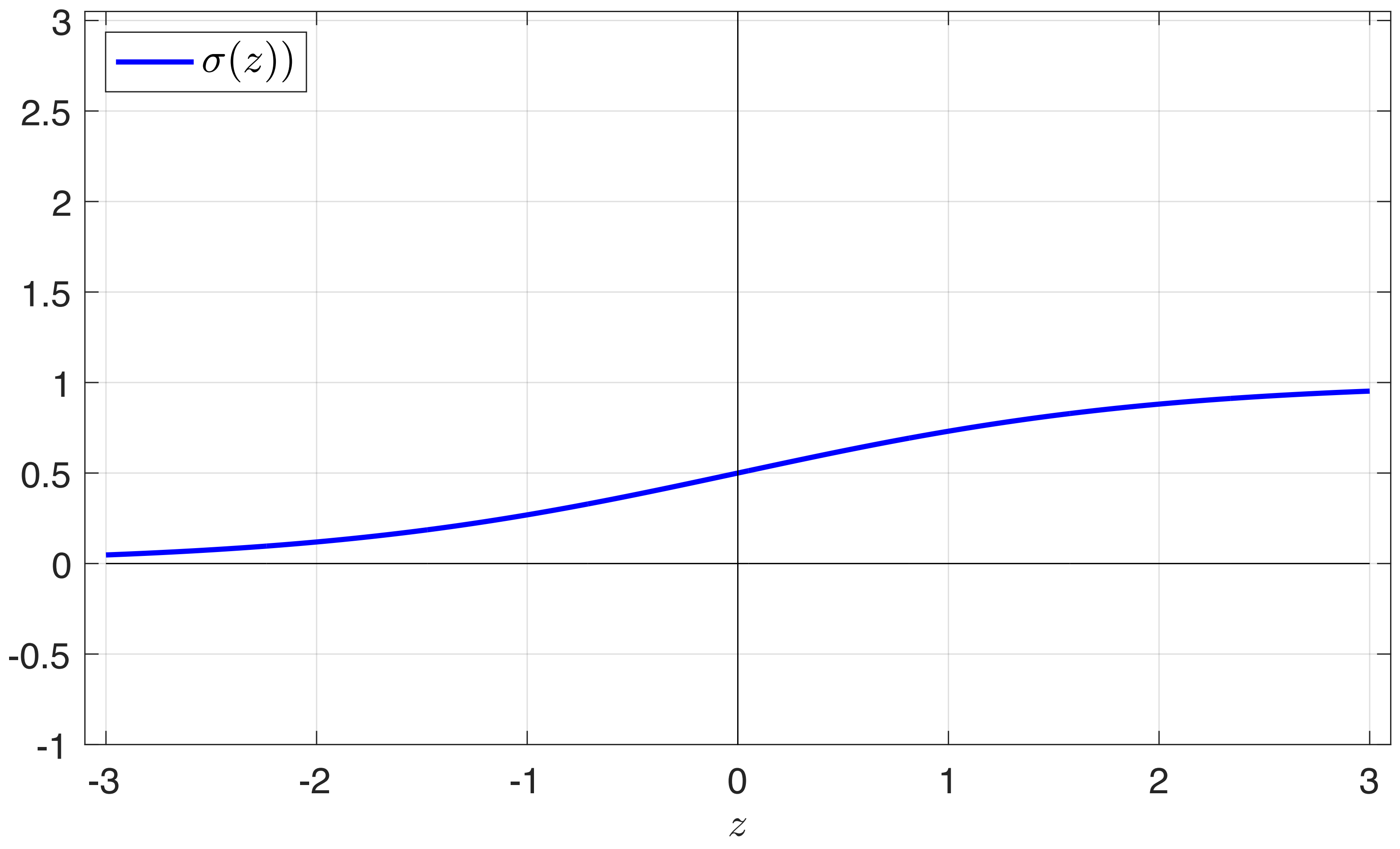
$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

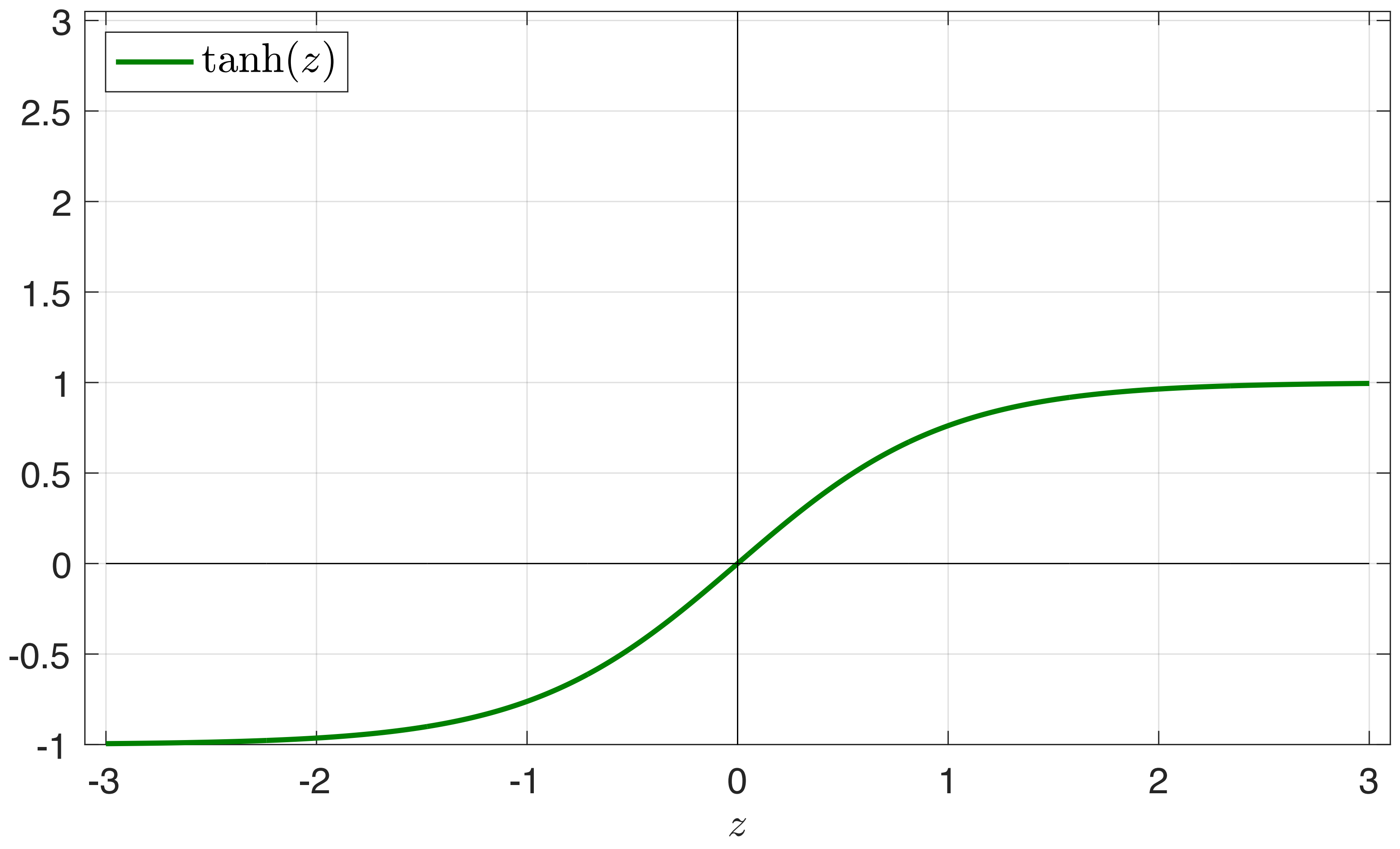
$$\tanh(z) = \frac{\exp(z) - \exp(-z)}{\exp(z) + \exp(-z)}$$

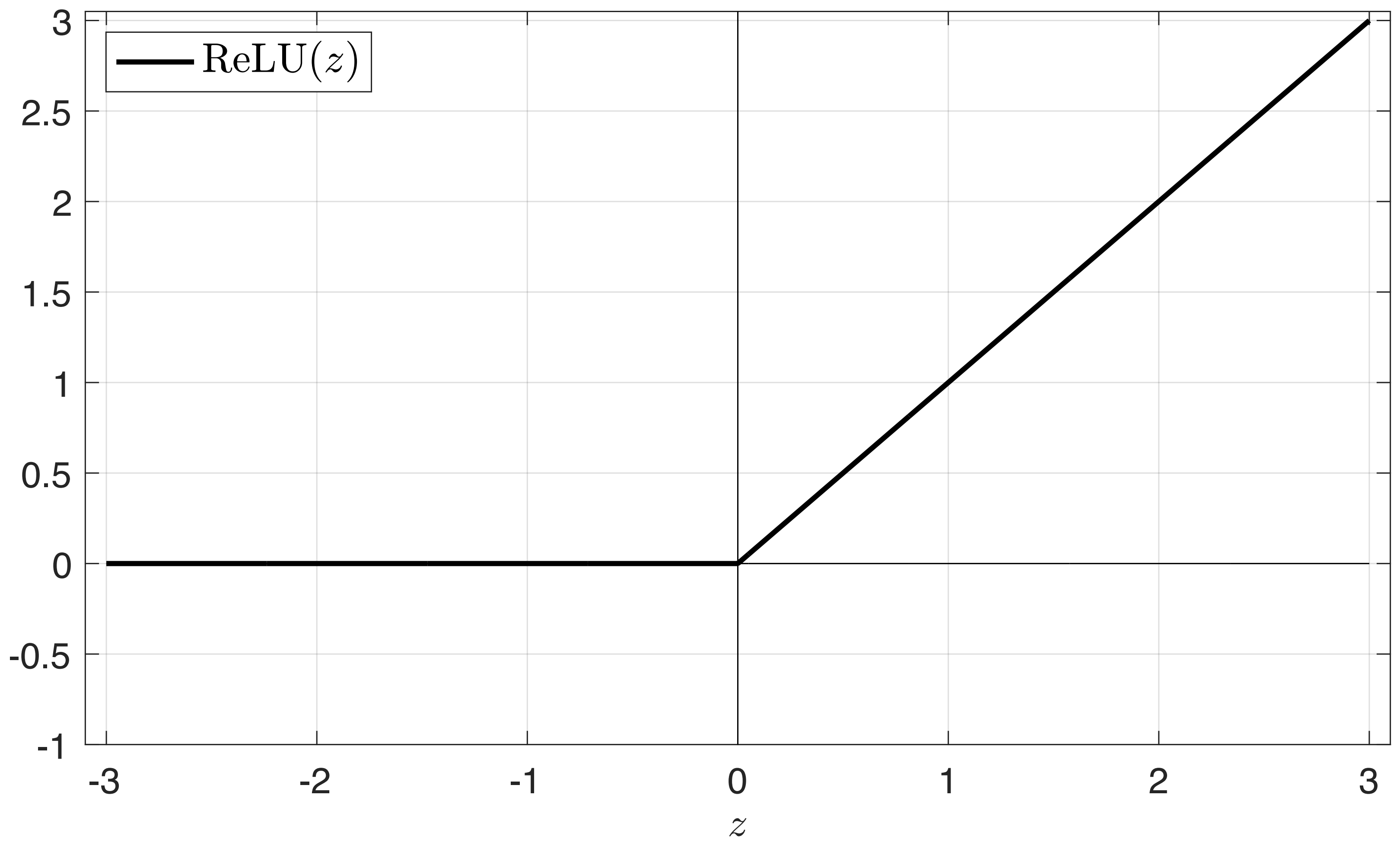
$$\sigma \in (0,1)$$

$$\tanh \in (-1,1)$$

Activation functions







$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

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

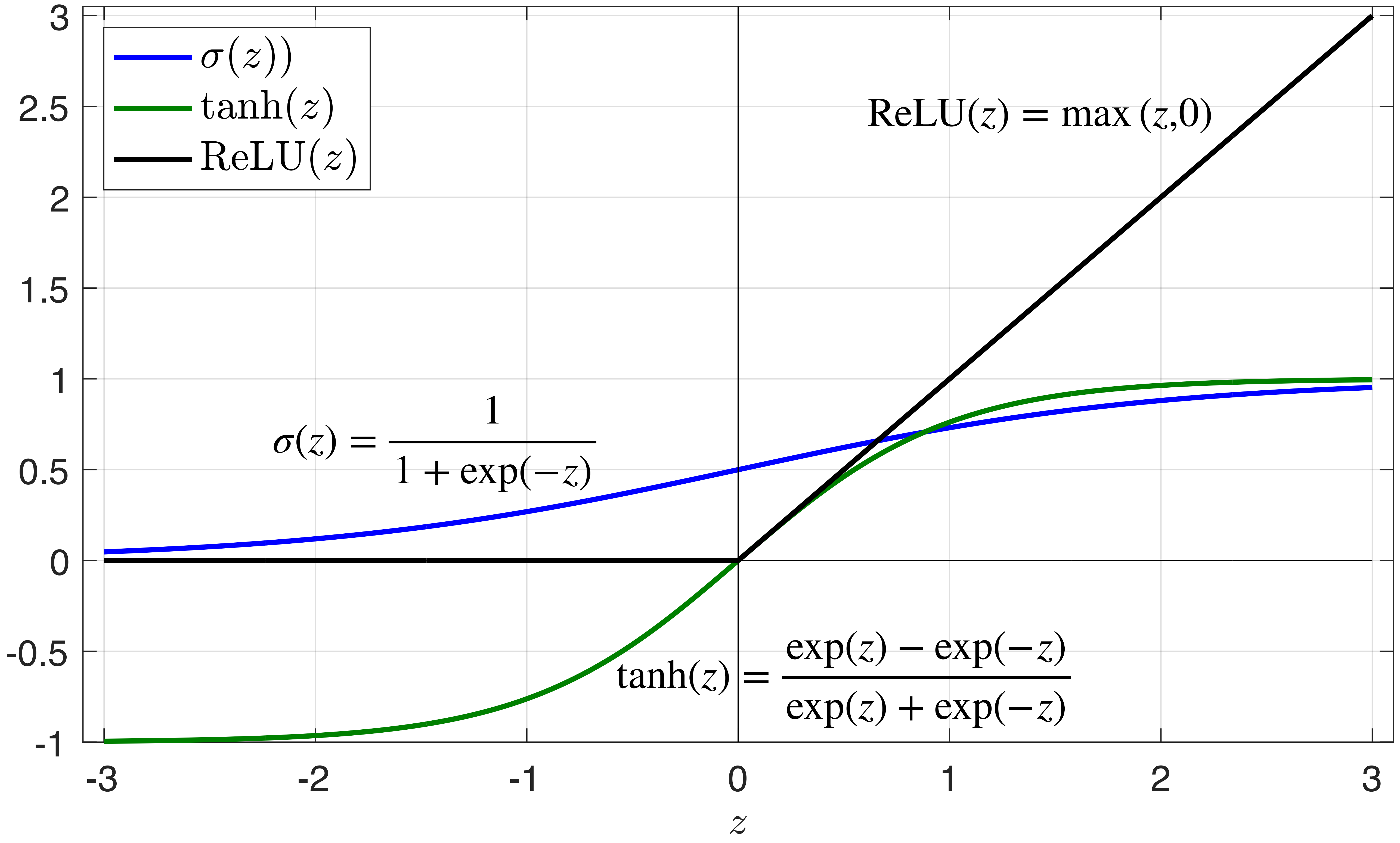
$$\mathrm{ReLU}(z) = \max(z, 0)$$

$$\sigma \in (0,1)$$

$$\tanh \in (-1,1)$$

$$\mathrm{ReLU} \in (0, +\infty)$$

Activation functions — Vanishing gradient

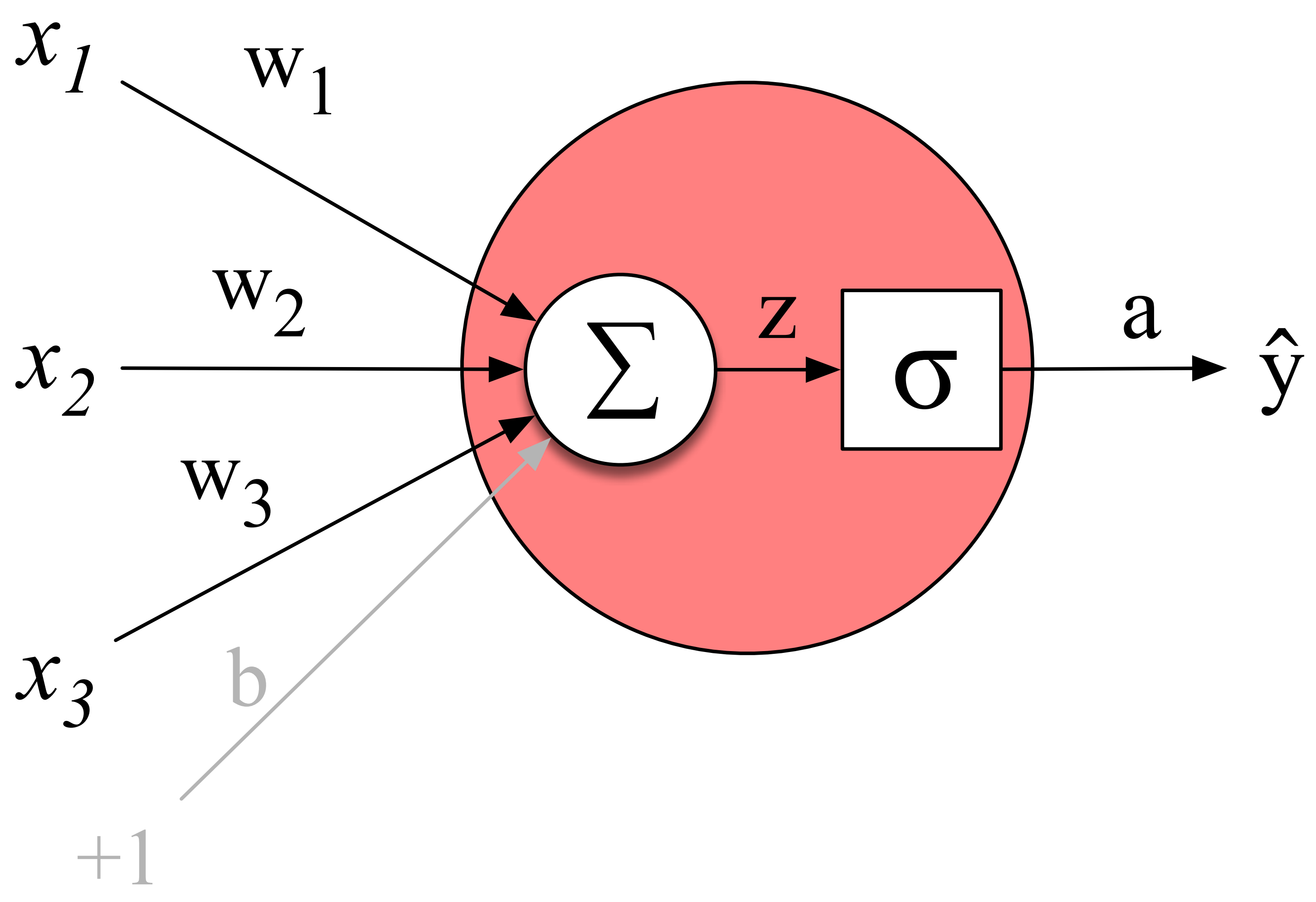


$\sigma \in (0,1)$

$\tanh \in (-1,1)$

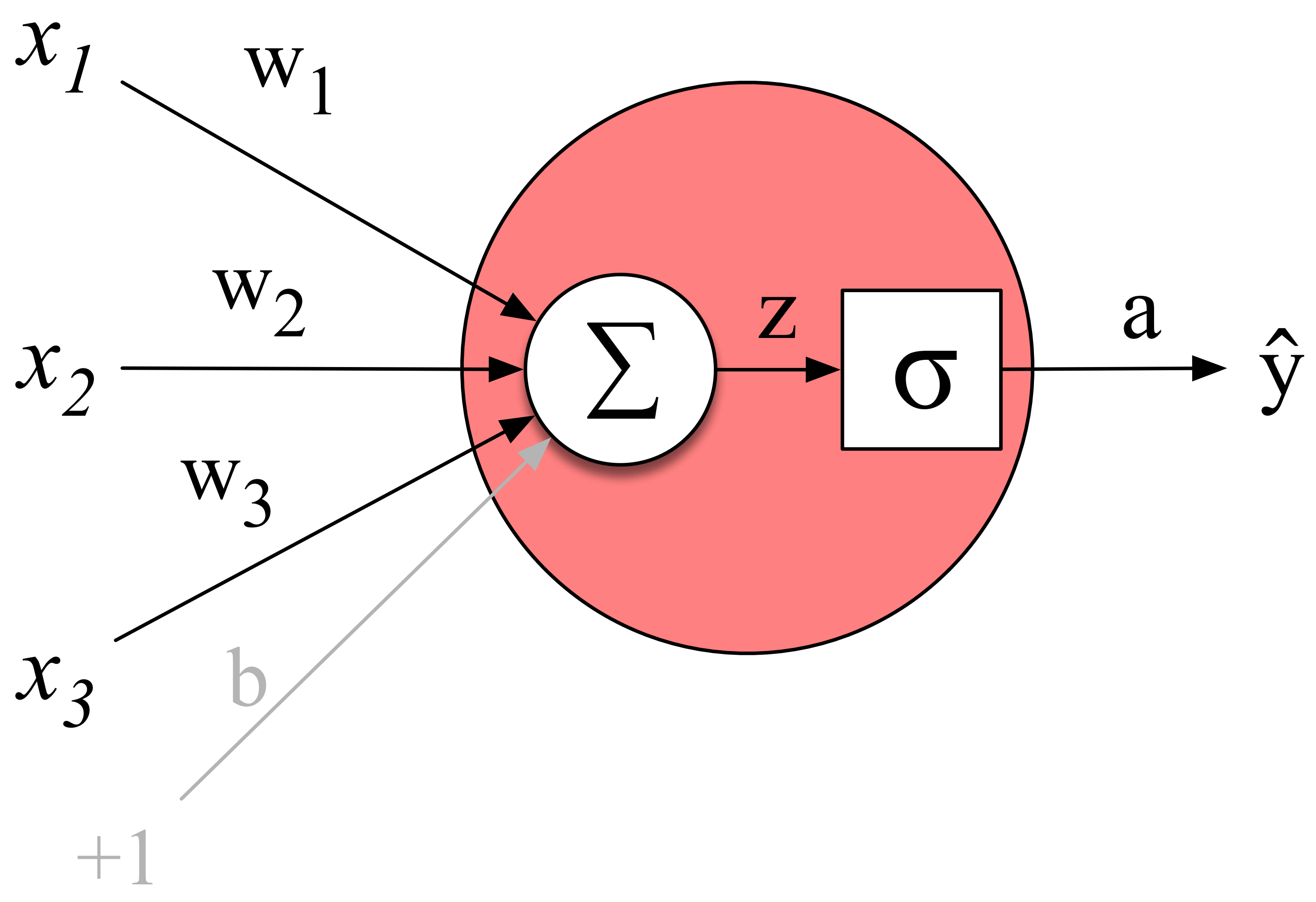
$\mathrm{ReLU} \in (0, +\infty)$

- σ, $\tanh$ are differentiable, ReLU not differentiable at 0
- $\tanh$ is almost always preferred to σ, more expansive mapping
- if $z >> 0$, σ and $\tanh$ become saturated, i.e. ≈ 1 with derivatives $\approx 0 \rightarrow$ gradient updates ≈ 0 (*no more learning*), **vanishing gradient issue**
- ReLU ~ linear / does not have this vanishing gradient issue



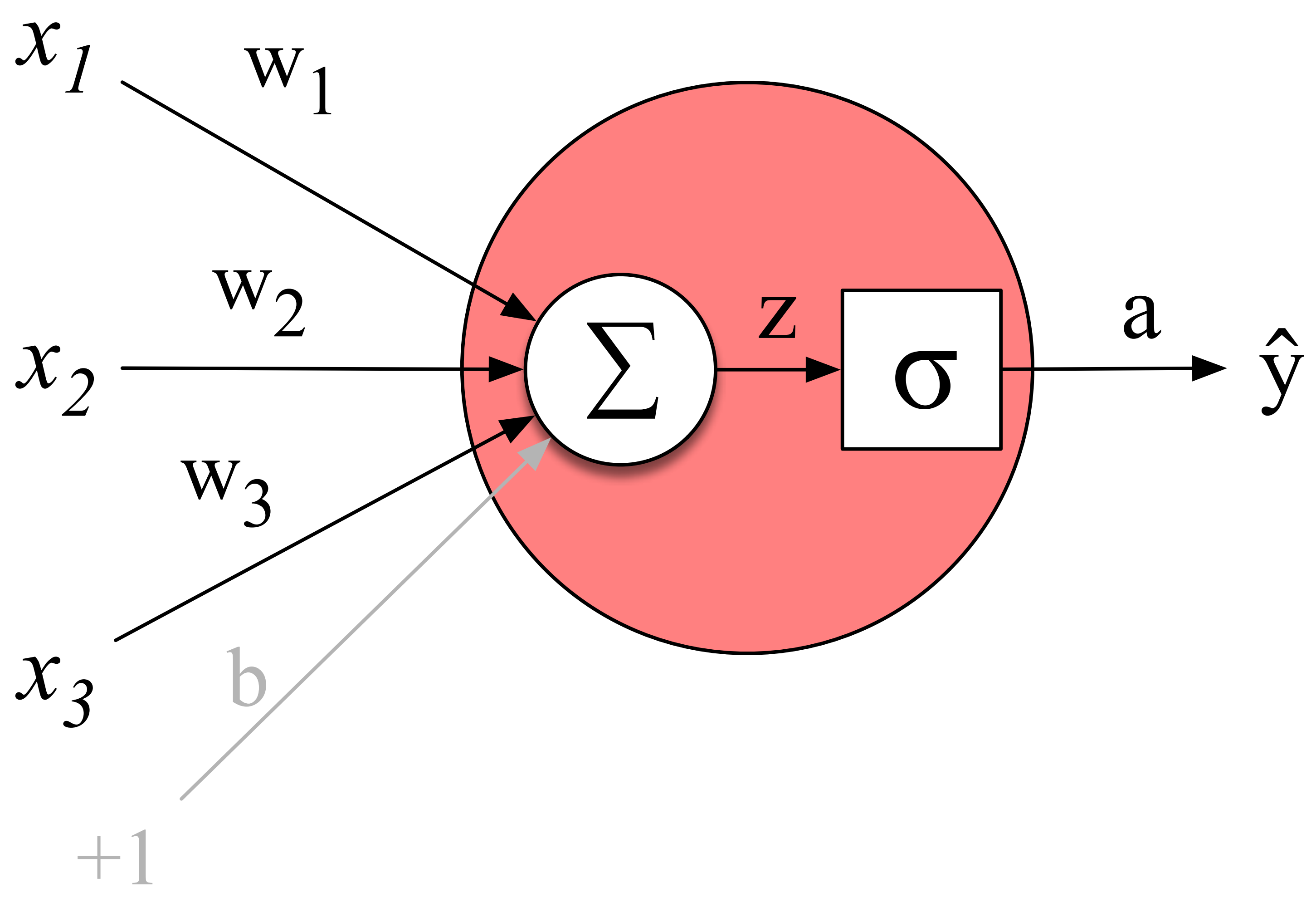
$$\mathbf{x} = [0.5\ 0.6\ 0.1]$$
$$\mathbf{w} = [0.2\ 0.3\ 0.9]$$
$$b = 0.5$$

One neural unit — Example



$$\mathbf{x} = [0.5\ 0.6\ 0.1]$$
$$\mathbf{w} = [0.2\ 0.3\ 0.9]$$
$$b = 0.5$$

$z = ?$



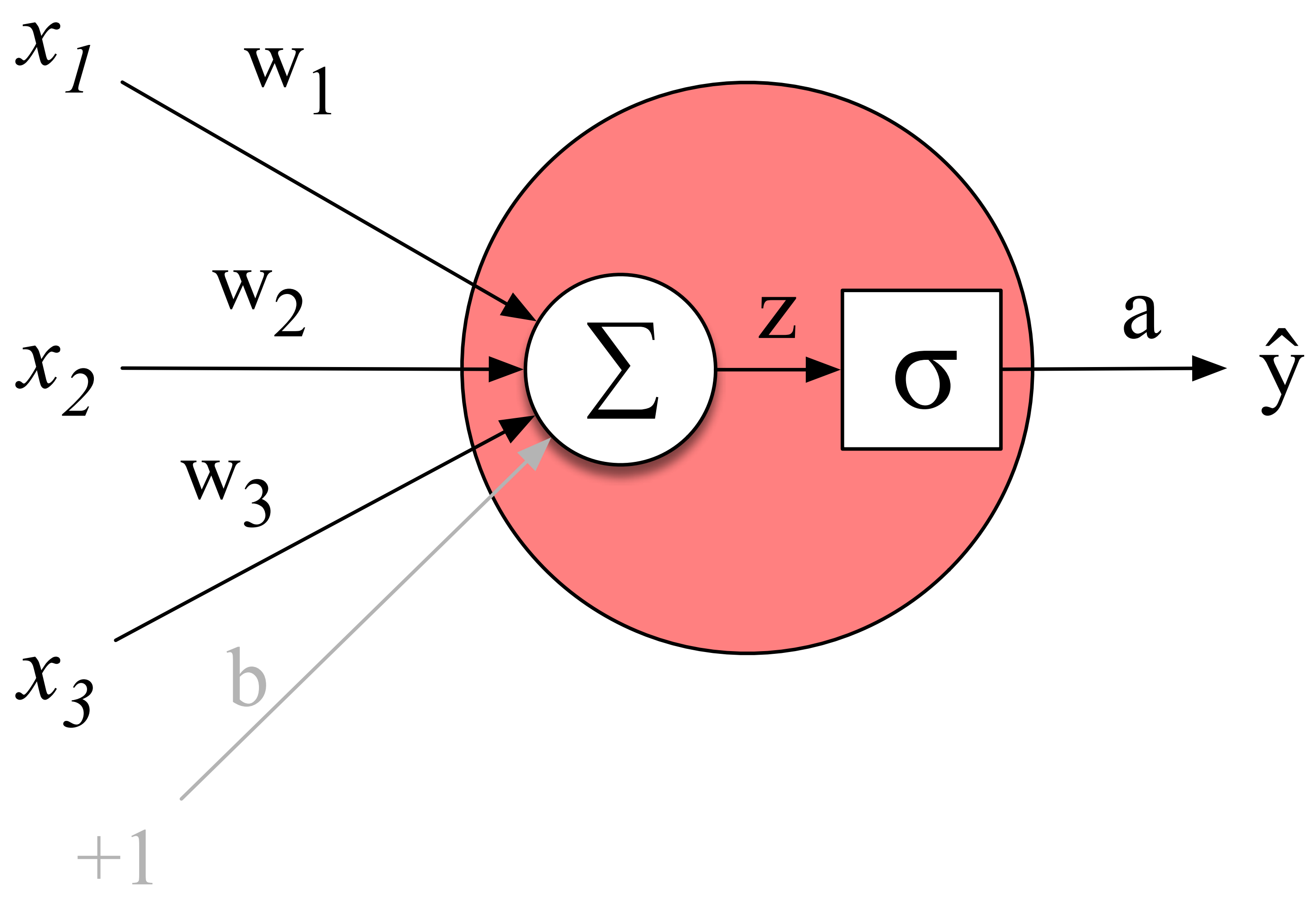
$$\mathbf{x} = [0.5\ 0.6\ 0.1]$$
$$\mathbf{w} = [0.2\ 0.3\ 0.9]$$
$$b = 0.5$$

$\xrightarrow{\text{bias}}$

$$\mathbf{x} = [1\ 0.5\ 0.6\ 0.1]$$
$$\mathbf{w} = [0.5\ 0.2\ 0.3\ 0.9]$$

$$z = ?$$

One neural unit — Example



$$\mathbf{x} = [0.5\ 0.6\ 0.1]$$
$$\mathbf{w} = [0.2\ 0.3\ 0.9]$$
$$b = 0.5$$

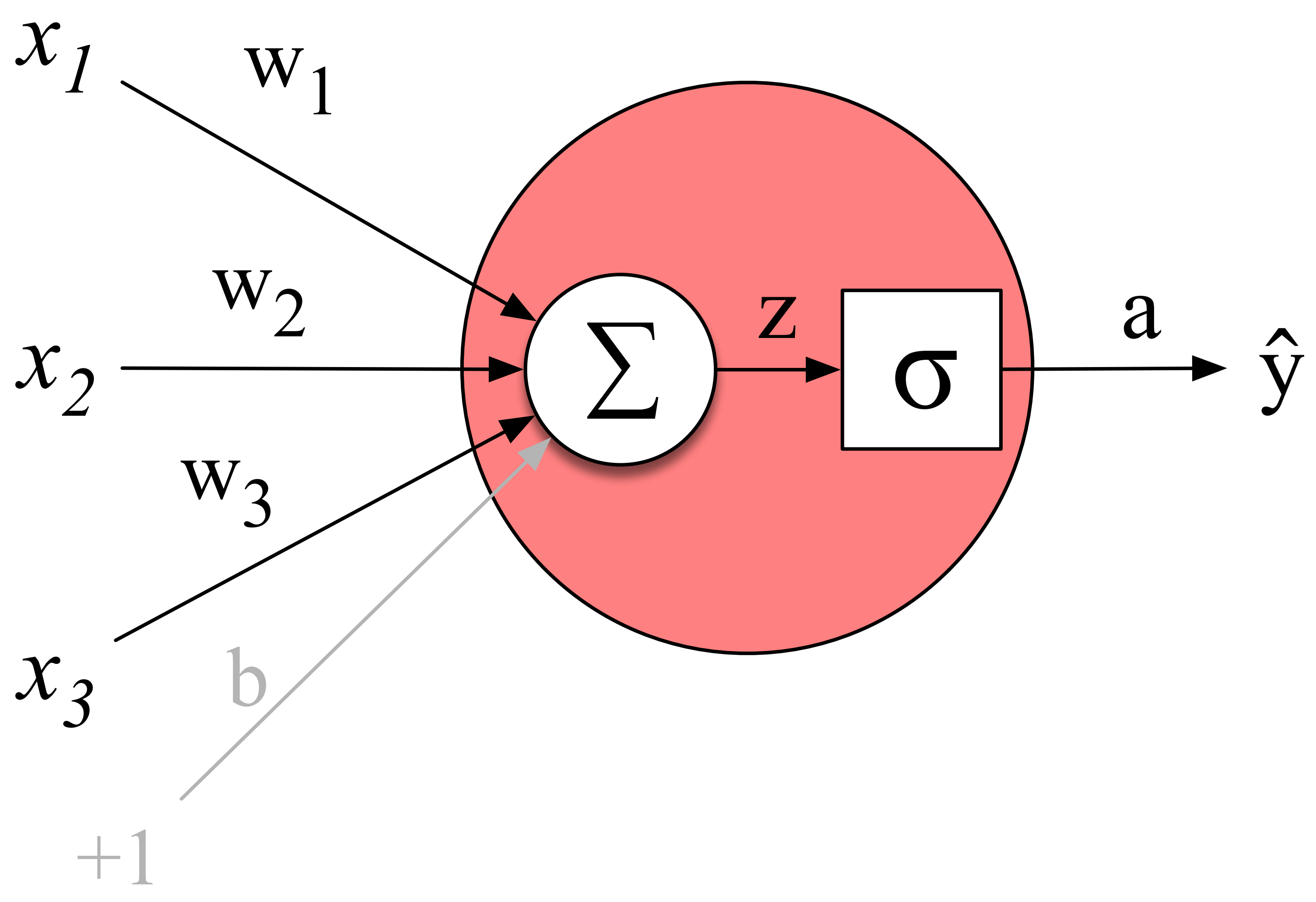
bias

$$\mathbf{x} = [1\ 0.5\ 0.6\ 0.1]$$
$$\mathbf{w} = [0.5\ 0.2\ 0.3\ 0.9]$$

$$\boxed{z = ?}$$

$$z = \mathbf{x} \cdot \mathbf{w} = 1 \cdot 0.5 + 0.5 \cdot 0.2 + \ldots + 0.1 \cdot 0.9 = 0.87$$

One neural unit — Example



$$\mathbf{x} = [0.5\ 0.6\ 0.1]$$
$$\mathbf{w} = [0.2\ 0.3\ 0.9]$$
$$b = 0.5$$

bias

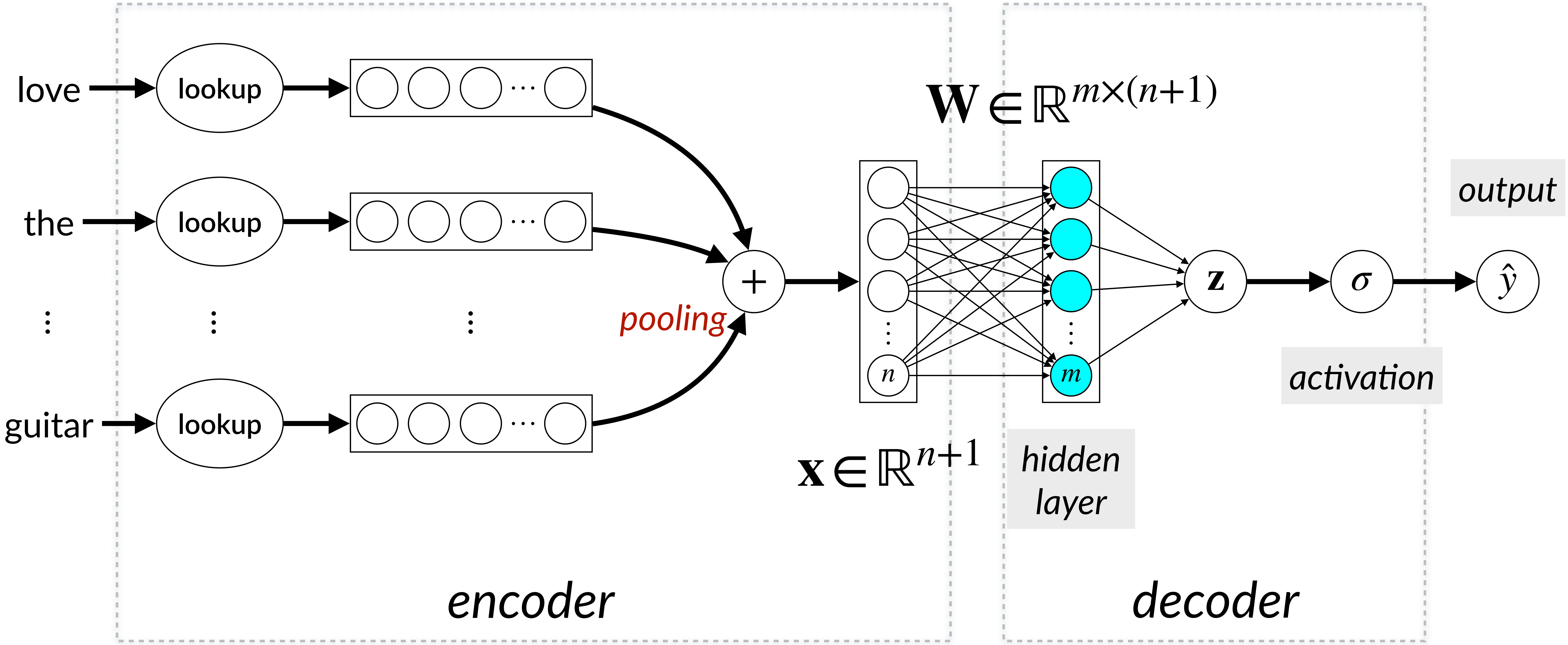
$$\mathbf{x} = [1\ 0.5\ 0.6\ 0.1]$$
$$\mathbf{w} = [0.5\ 0.2\ 0.3\ 0.9]$$

$$\boxed{z = ?}$$

$$z = \mathbf{x} \cdot \mathbf{w} = 1 \cdot 0.5 + 0.5 \cdot 0.2 + \ldots + 0.1 \cdot 0.9 = 0.87$$

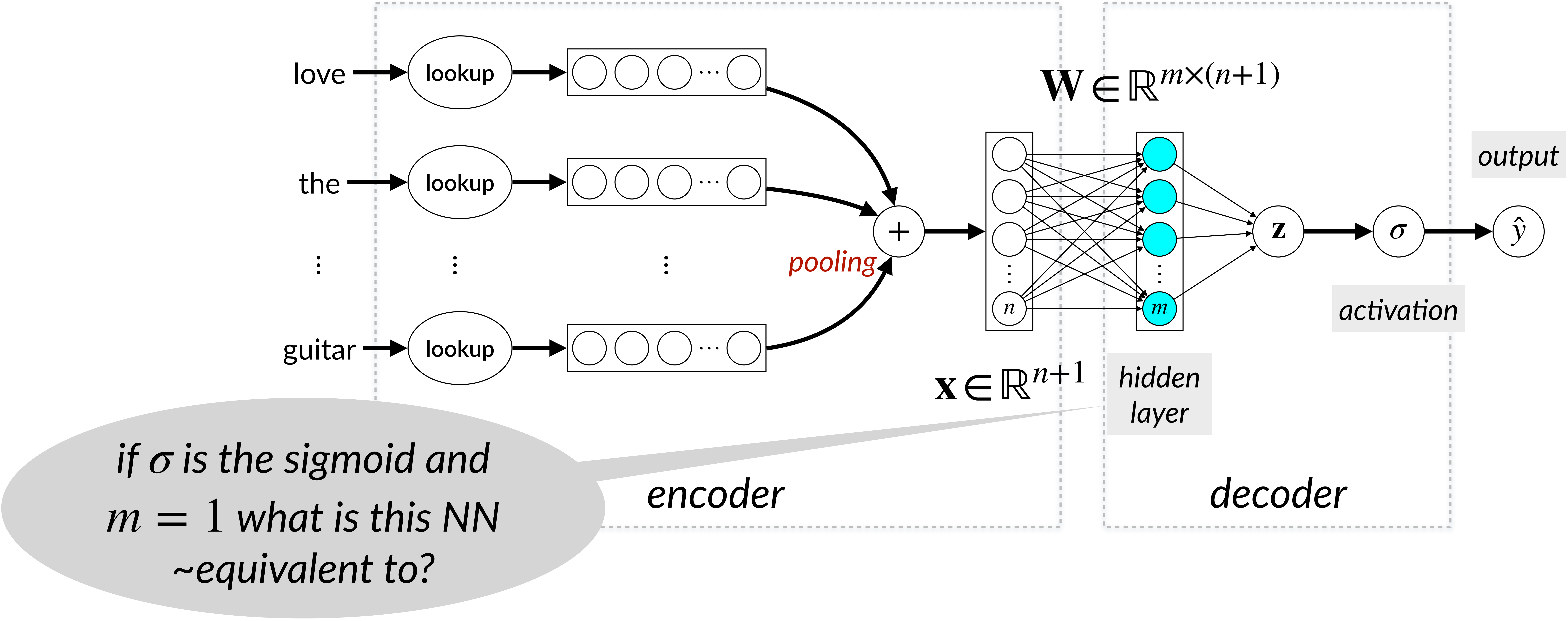
$$\hat{y} = a = \sigma(z) = \frac{1}{1 + \exp(-z)} = \frac{1}{1 + \exp(-0.87)} = 0.705$$

Feedforward neural network (1 layer)



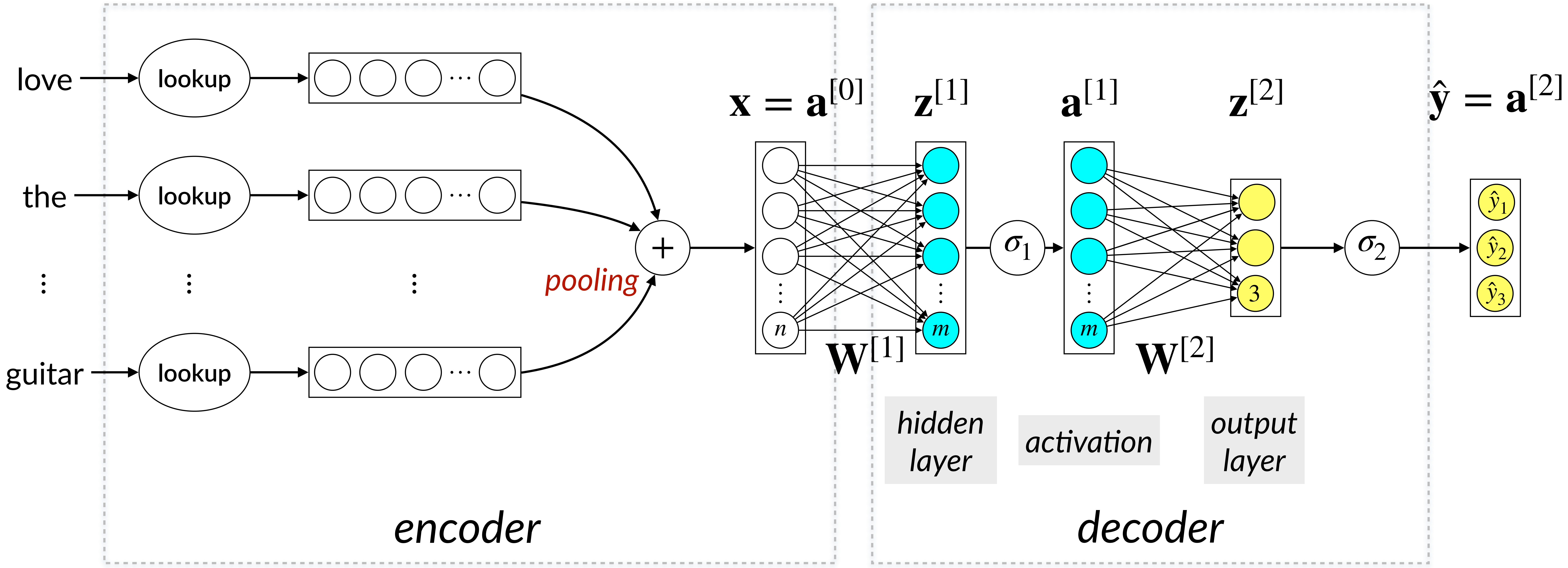
– A feedforward NN has: input units, hidden units, and output units
– Fully connected (*standard version*)
– This NN has 1 layer (*input layer does not count*)

Feedforward neural network (1 layer)



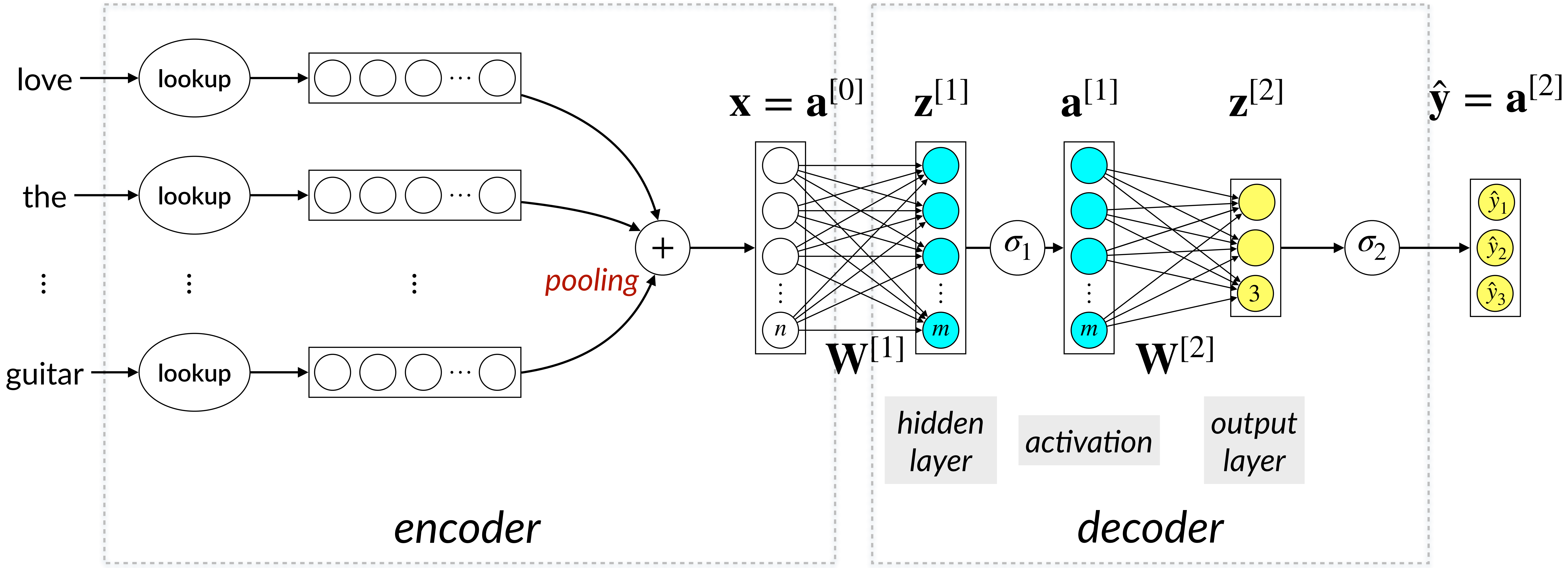
– A feedforward NN has: input units, hidden units, and output units
– Fully connected (*standard version*)
– This NN has 1 layer (*input layer does not count*)

Feedforward neural network (2 layers, multiple outputs)



$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$$
$$\mathbf{a}^{[1]} = \sigma_1\left(\mathbf{z}^{[1]}\right)$$
$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{a}^{[1]}$$
$$\mathbf{a}^{[2]} = \sigma_2\left(\mathbf{z}^{[2]}\right)$$
$$\hat{\mathbf{y}} = \mathbf{a}^{[2]}$$

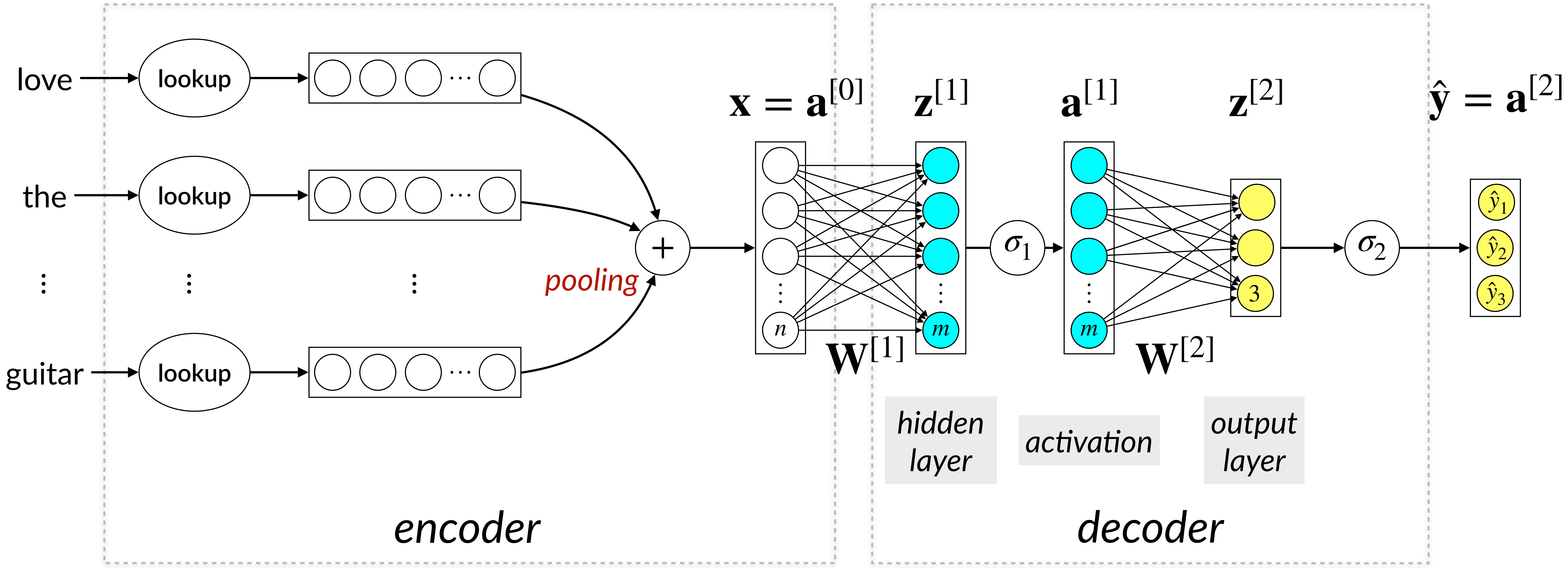
Feedforward neural network (2 layers, multiple outputs)



$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$$
$$\mathbf{a}^{[1]} = \sigma_1\left(\mathbf{z}^{[1]}\right)$$
$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{a}^{[1]}$$
$$\mathbf{a}^{[2]} = \sigma_2\left(\mathbf{z}^{[2]}\right)$$
$$\hat{\mathbf{y}} = \mathbf{a}^{[2]}$$

Dimensionalities?

Feedforward neural network (2 layers, multiple outputs)

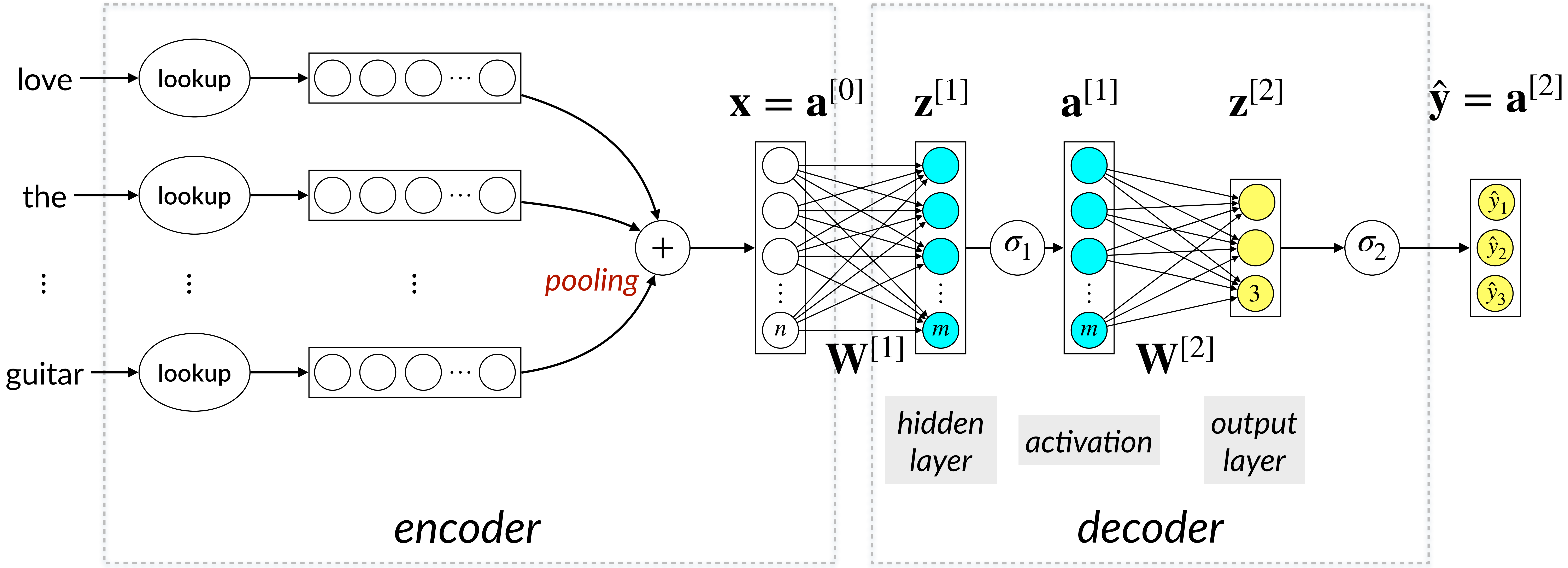


$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$$
$$\mathbf{a}^{[1]} = \sigma_1\left(\mathbf{z}^{[1]}\right)$$
$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{a}^{[1]}$$
$$\mathbf{a}^{[2]} = \sigma_2\left(\mathbf{z}^{[2]}\right)$$
$$\hat{\mathbf{y}} = \mathbf{a}^{[2]}$$

$\mathbf{x}, \mathbf{a}^{[0]} \in \mathbb{R}^{n+1}$

Dimensionalities?

Feedforward neural network (2 layers, multiple outputs)



$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$$
$$\mathbf{a}^{[1]} = \sigma_1\left(\mathbf{z}^{[1]}\right)$$
$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{a}^{[1]}$$
$$\mathbf{a}^{[2]} = \sigma_2\left(\mathbf{z}^{[2]}\right)$$
$$\hat{\mathbf{y}} = \mathbf{a}^{[2]}$$

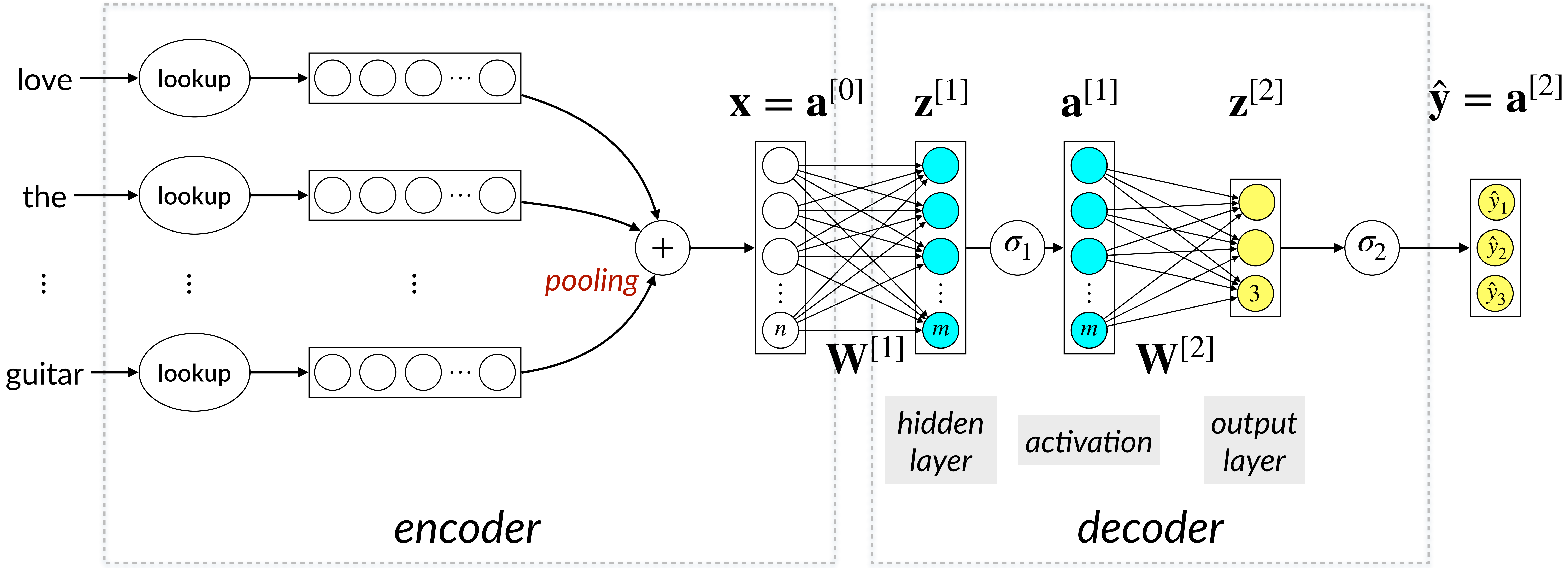
$\mathbf{x}, \mathbf{a}^{[0]} \in \mathbb{R}^{n+1}$

$\mathbf{W}^{[1]} \in \mathbb{R}^{m\times(n+1)}$

bias term (+1)

Dimensionalities?

Feedforward neural network (2 layers, multiple outputs)



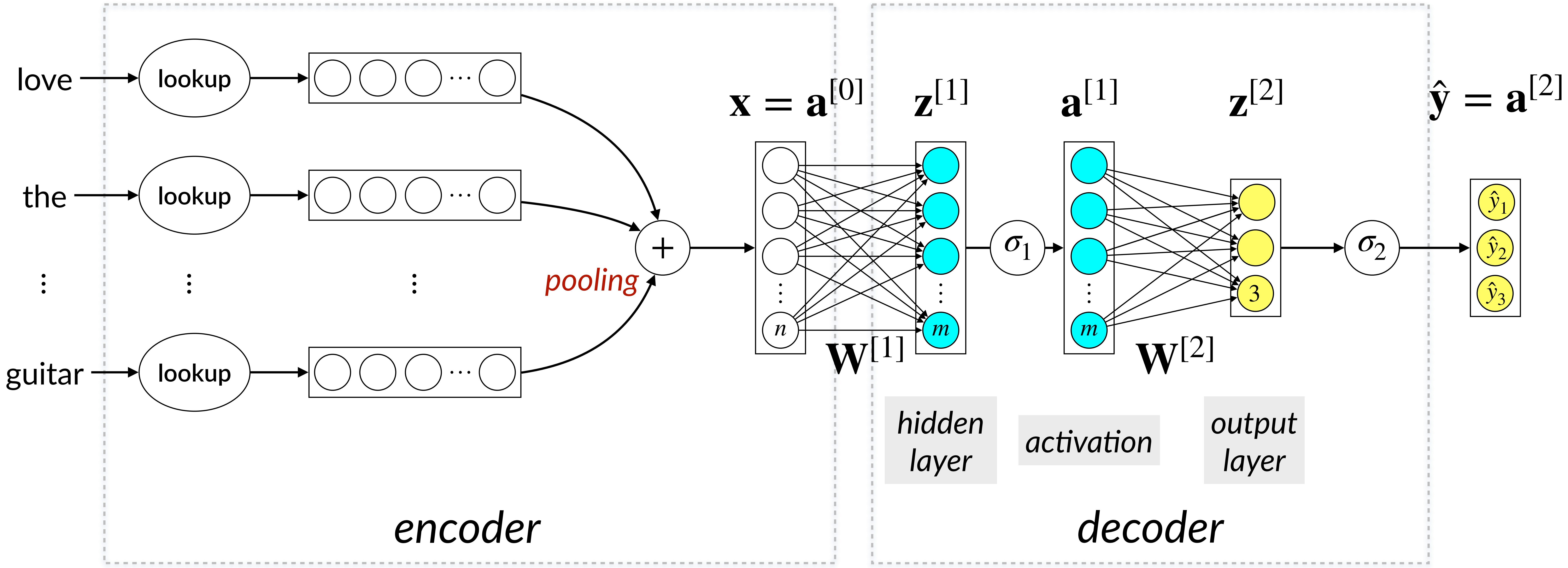
$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$$
$$\mathbf{a}^{[1]} = \sigma_1\left(\mathbf{z}^{[1]}\right)$$
$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{a}^{[1]}$$
$$\mathbf{a}^{[2]} = \sigma_2\left(\mathbf{z}^{[2]}\right)$$
$$\hat{\mathbf{y}} = \mathbf{a}^{[2]}$$

$$\mathbf{x}, \mathbf{a}^{[0]} \in \mathbb{R}^{n+1}$$
$$\mathbf{W}^{[1]} \in \mathbb{R}^{m\times(n+1)}$$
$$\mathbf{a}^{[1]}, \mathbf{z}^{[1]} \in \mathbb{R}^{m}$$

bias term (+1)

Dimensionalities?

Feedforward neural network (2 layers, multiple outputs)



$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$$
$$\mathbf{a}^{[1]} = \sigma_1\left(\mathbf{z}^{[1]}\right)$$
$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{a}^{[1]}$$
$$\mathbf{a}^{[2]} = \sigma_2\left(\mathbf{z}^{[2]}\right)$$
$$\hat{\mathbf{y}} = \mathbf{a}^{[2]}$$

$\mathbf{x}, \mathbf{a}^{[0]} \in \mathbb{R}^{n+1}$ *bias term* (+1)

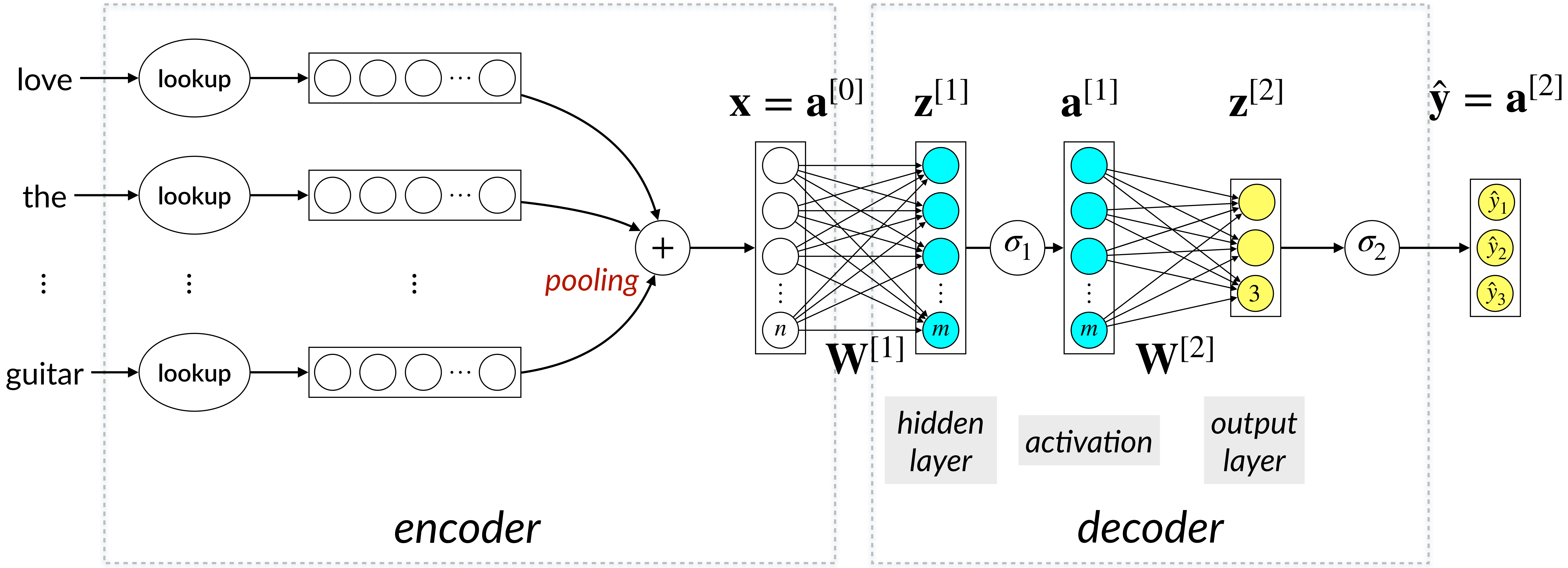
$\mathbf{W}^{[1]} \in \mathbb{R}^{m\times(n+1)}$

$\mathbf{a}^{[1]}, \mathbf{z}^{[1]} \in \mathbb{R}^{m}$

$\mathbf{W}^{[2]} \in \mathbb{R}^{3\times m}$

Dimensionalities?

Feedforward neural network (2 layers, multiple outputs)



$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$$
$$\mathbf{a}^{[1]} = \sigma_1\left(\mathbf{z}^{[1]}\right)$$
$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{a}^{[1]}$$
$$\mathbf{a}^{[2]} = \sigma_2\left(\mathbf{z}^{[2]}\right)$$
$$\hat{\mathbf{y}} = \mathbf{a}^{[2]}$$

$\mathbf{x}, \mathbf{a}^{[0]} \in \mathbb{R}^{n+1}$ *bias term* (+1)

$\mathbf{W}^{[1]} \in \mathbb{R}^{m\times(n+1)}$

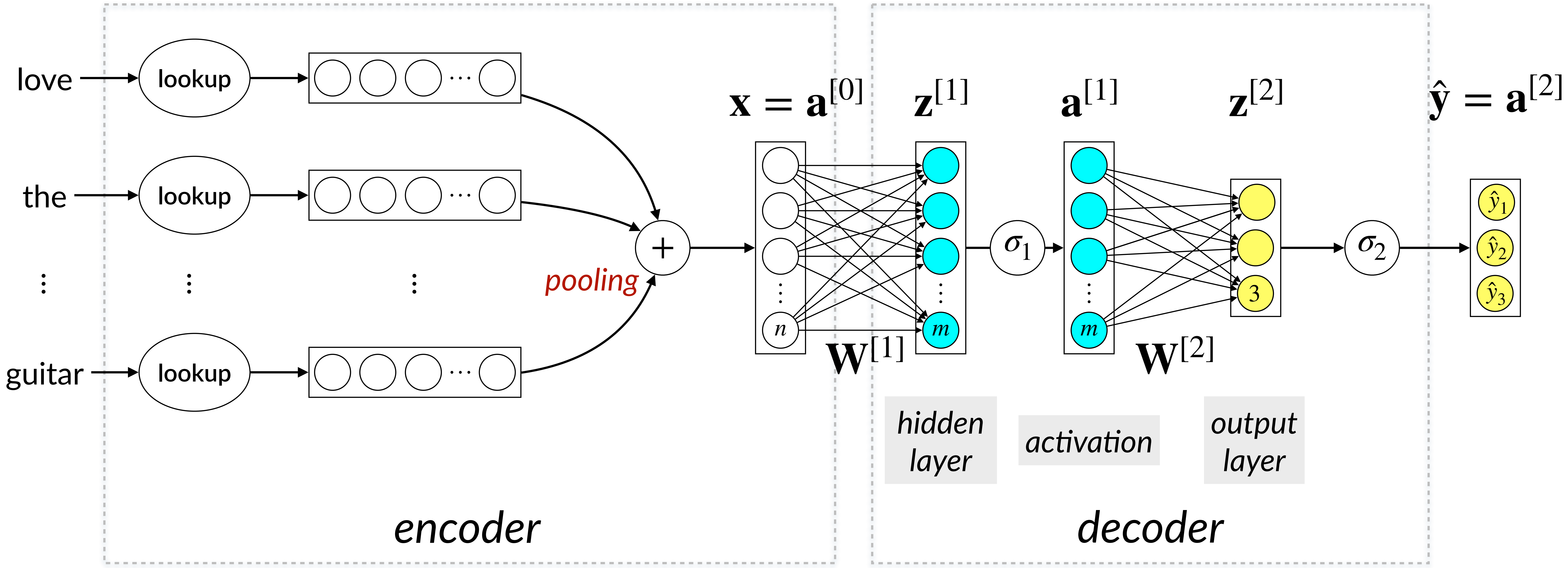
$\mathbf{a}^{[1]}, \mathbf{z}^{[1]} \in \mathbb{R}^{m}$

$\mathbf{W}^{[2]} \in \mathbb{R}^{3\times m}$ *no bias term used in the output layer*

$\mathbf{a}^{[2]}, \mathbf{z}^{[2]}, \hat{\mathbf{y}} \in \mathbb{R}^{3}$

Dimensionalities?

Feedforward neural network (2 layers, multiple outputs)



$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$$
$$\mathbf{a}^{[1]} = \sigma_1\left(\mathbf{z}^{[1]}\right)$$
$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{a}^{[1]}$$
$$\mathbf{a}^{[2]} = \sigma_2\left(\mathbf{z}^{[2]}\right)$$
$$\hat{\mathbf{y}} = \mathbf{a}^{[2]}$$

$$\mathbf{x}, \mathbf{a}^{[0]} \in \mathbb{R}^{n+1}$$

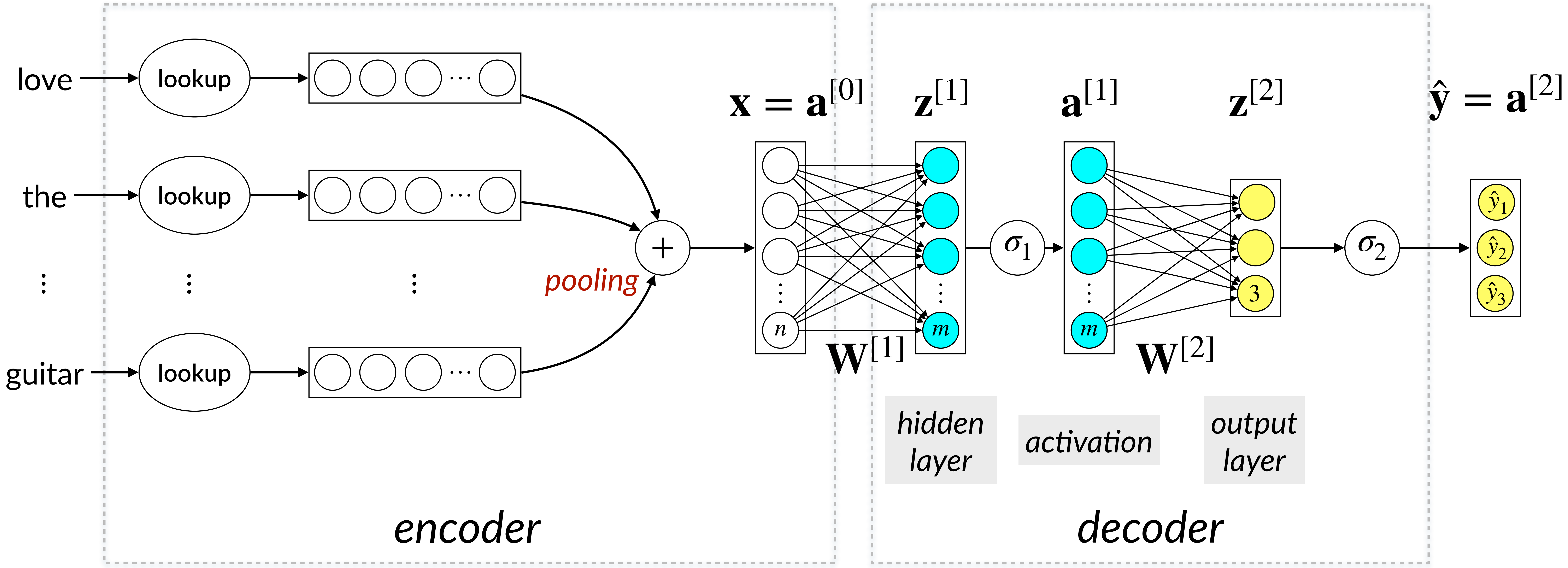
bias term (+1)

$$\mathbf{W}^{[1]} \in \mathbb{R}^{m\times(n+1)}$$
$$\mathbf{a}^{[1]}, \mathbf{z}^{[1]} \in \mathbb{R}^{m}$$
$$\mathbf{W}^{[2]} \in \mathbb{R}^{3\times m}$$
$$\mathbf{a}^{[2]}, \mathbf{z}^{[2]}, \hat{\mathbf{y}} \in \mathbb{R}^{3}$$

no bias term used in the output layer

$$\boxed{\hat{\mathbf{y}} = \sigma_2\Big(\mathbf{W}^{[2]}\,\sigma_1\big(\mathbf{W}^{[1]}\mathbf{x}\big)\Big)}$$

Feedforward neural network (2 layers, multiple outputs)



$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$$
$$\mathbf{a}^{[1]} = \sigma_1\left(\mathbf{z}^{[1]}\right)$$
$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{a}^{[1]}$$
$$\mathbf{a}^{[2]} = \sigma_2\left(\mathbf{z}^{[2]}\right)$$
$$\hat{\mathbf{y}} = \mathbf{a}^{[2]}$$

$$\mathbf{x}, \mathbf{a}^{[0]} \in \mathbb{R}^{n+1}$$
$$\mathbf{W}^{[1]} \in \mathbb{R}^{m\times(n+1)}$$
$$\mathbf{a}^{[1]}, \mathbf{z}^{[1]} \in \mathbb{R}^{m}$$
$$\mathbf{W}^{[2]} \in \mathbb{R}^{3\times m}$$
$$\mathbf{a}^{[2]}, \mathbf{z}^{[2]}, \hat{\mathbf{y}} \in \mathbb{R}^{3}$$

bias term (+1)

no bias term used in the output layer

How many parameters do we need to learn?

Are nonlinear (σ) activation functions necessary?

If our activation functions were linear in $\hat{\mathbf{y}} = \sigma_2\Big(\mathbf{W}^{[2]}\ \sigma_1\big(\mathbf{W}^{[1]}\mathbf{x}\big)\Big)$ ...

Are nonlinear (σ) activation functions necessary?

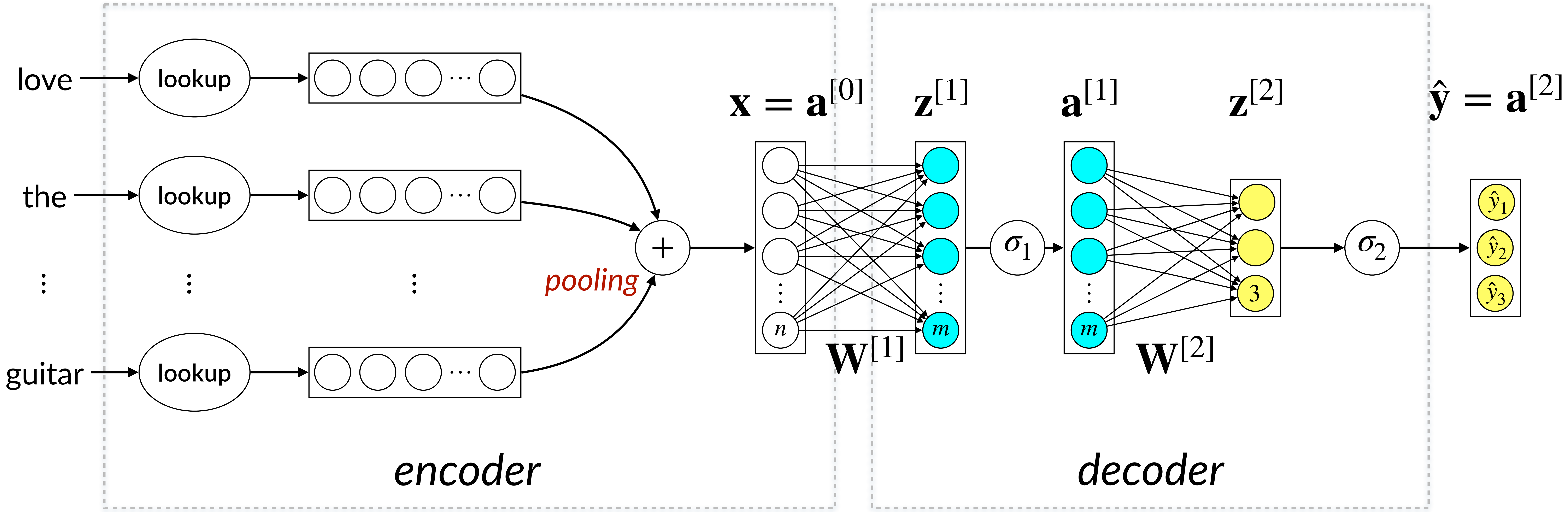
If our activation functions were linear in $\hat{\mathbf{y}} = \sigma_2\Big(\mathbf{W}^{[2]}\ \sigma_1\big(\mathbf{W}^{[1]}\mathbf{x}\big)\Big)$ ...

then we can simply omit the non-linear activations σ_1 and σ_2:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{z}^{[2]}\\ &= \mathbf{W}^{[2]}\mathbf{z}^{[1]}\\ &= \mathbf{W}^{[2]}\mathbf{W}^{[1]}\mathbf{x}\\ &= \mathbf{W}'\mathbf{x}\end{aligned}$$

Hence, we have reduced 2 layers back to 1 with altered parameters ($\mathbf{W}'$).
This generalises to **any** number of layers.

Inference with a feedforward neural network — Softmax



Need to convert outputs to pseudo-probabilities

➡ common setting for σ_2 is the softmax function

$$y_i = \text{softmax}\left(z_i\right) = \frac{\exp\left(z_i\right)}{\sum_{j=1}^{d} \exp\left(z_j\right)}, \quad 1 \leq i \leq d$$

Softmax example

$$y_i = \text{softmax}\left(z_i\right) = \frac{\exp\left(z_i\right)}{\sum_{j=1}^{d} \exp\left(z_j\right)}, \ \ 1 \le i \le d$$

$\sum_i y_i = 1$ and $y_i \in [0,1]$ *pseudo-probabilities*

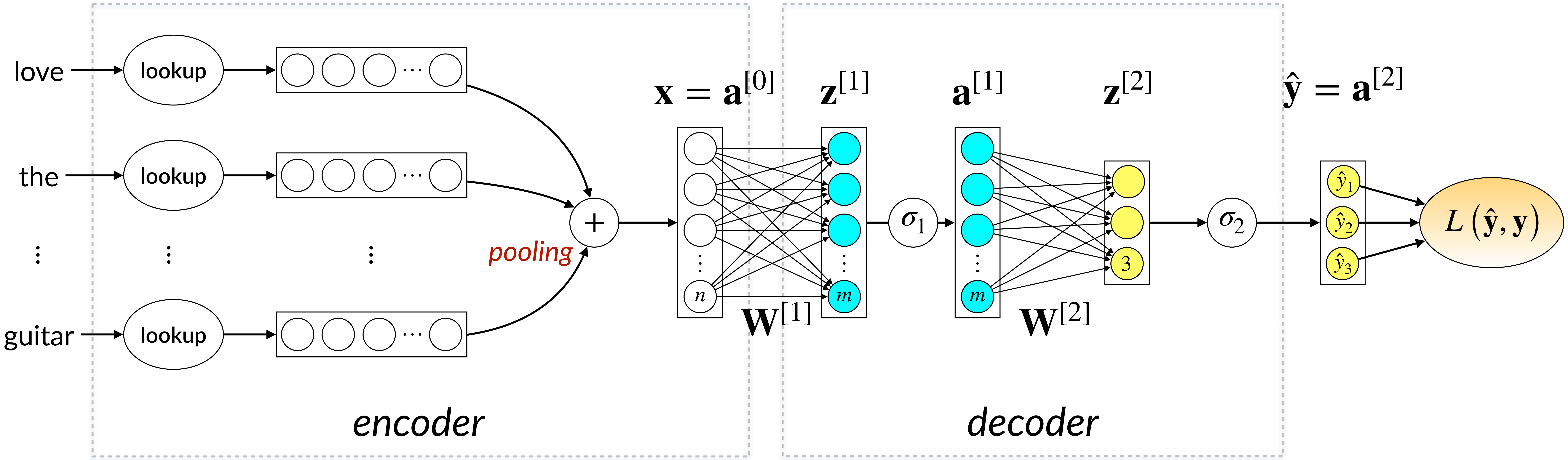
So, in our example if

$$\mathbf{z}^{[2]} = [2 \quad -1.99 \quad -0.01]$$

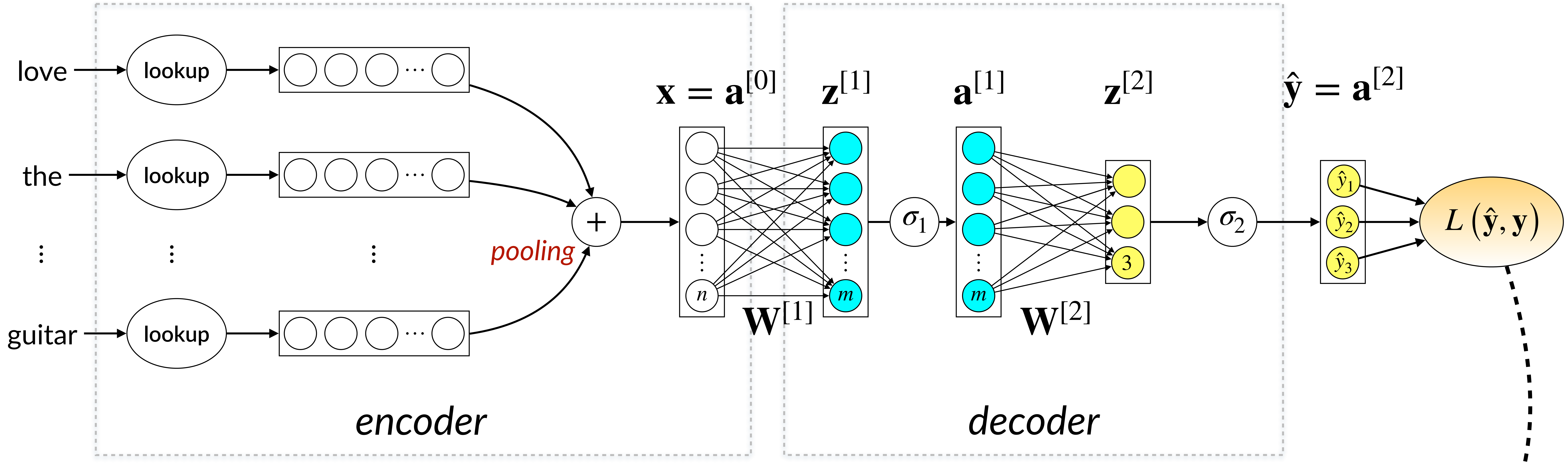
then

$$\hat{\mathbf{y}} = \text{softmax}\left(\mathbf{z}^{[2]}\right) = [0.868 \quad 0.016 \quad 0.116]$$

Training a feedforward neural network

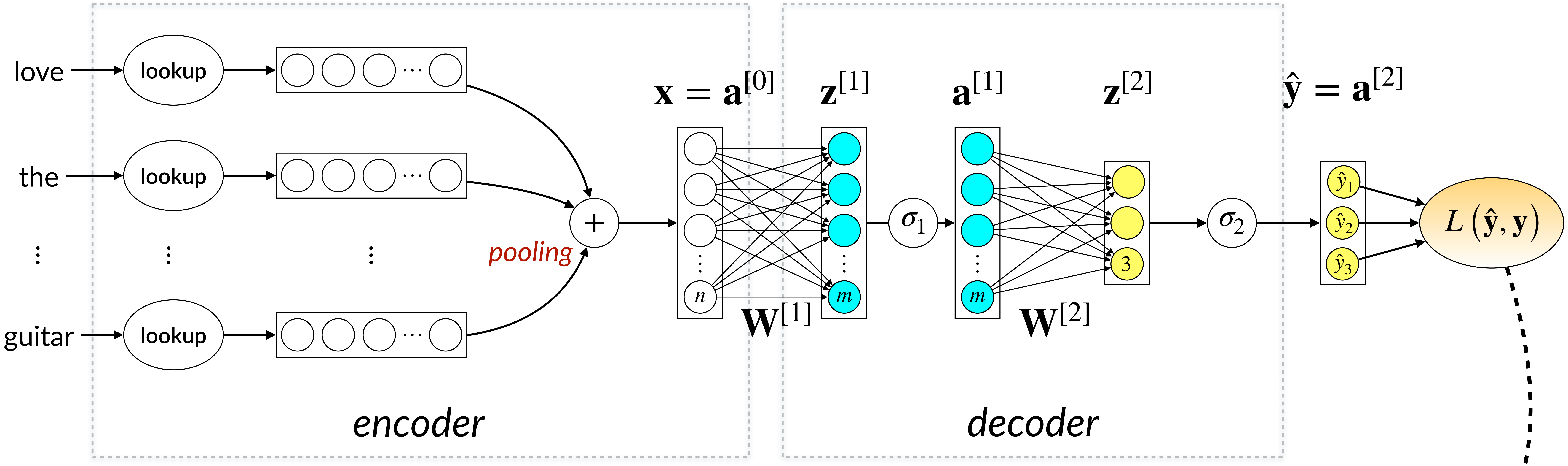


Training a feedforward neural network



How would I use a loss function $L(\hat{\mathbf{y}}, \mathbf{y})$
to update *efficiently* my NN parameters
in $\mathbf{W}^{[1]}$ and $\mathbf{W}^{[2]}$?

Training a feedforward neural network



Backpropagation
a.k.a.
"backprop"

How would I use a loss function $L(\hat{\mathbf{y}}, \mathbf{y})$ to update *efficiently* my NN parameters in $\mathbf{W}^{[1]}$ and $\mathbf{W}^{[2]}$?

Cross-entropy loss function

Cross-entropy loss

$$L_{\text{ce}}\left(\hat{\mathbf{y}}, \mathbf{y}\right) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

where K is the number of output classes

Cross-entropy loss function

Cross-entropy loss $$L_{\text{ce}}\left(\hat{\mathbf{y}}, \mathbf{y}\right) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

where K is the number of output classes

Only one of the K y_k's will be equal to 1. The rest will be 0.
If, say, $y_c = 1, c = \{1,\ldots,K\}$, i.e. c is the correct class,
the loss can be simplified as:

$$L_{\text{ce}}\left(\hat{\mathbf{y}}, \mathbf{y}\right) = -\, y_c \cdot \log \hat{y}_c = -\log \hat{y}_c$$

Cross-entropy loss function

Cross-entropy loss $$L_{\text{ce}}\left(\hat{\mathbf{y}}, \mathbf{y}\right) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

where K is the number of output classes

Only one of the K y_k's will be equal to 1. The rest will be 0.
If, say, $y_c = 1, c = \{1,\ldots,K\}$, i.e. c is the correct class,
the loss can be simplified as:

$$L_{\text{ce}}\left(\hat{\mathbf{y}}, \mathbf{y}\right) = -y_c \cdot \log \hat{y}_c = -\log \hat{y}_c$$

$$= -\log \underbrace{\frac{\exp\left(z_c\right)}{\sum_{j=1}^{K} \exp\left(z_j\right)}}_{\text{softmax}}$$

Backpropagation uses the chain rule

$$f(x) = g\big(h(x)\big)$$

Chain rule: $$\frac{df}{dx} = \frac{dg}{dh} \cdot \frac{dh}{dx}$$

Backpropagation uses the chain rule

$$f(x) = g\big(h(x)\big)$$

Chain rule: $$\frac{df}{dx} = \frac{dg}{dh} \cdot \frac{dh}{dx}$$

$$f(x) = (x^2 + 1)^2 = g\big(h(x)\big) \quad ??$$

Backpropagation uses the chain rule

$$f(x) = g\big(h(x)\big)$$

Chain rule: $$\frac{df}{dx} = \frac{dg}{dh} \cdot \frac{dh}{dx}$$

$$f(x) = (x^2 + 1)^2 = g\big(h(x)\big) \quad ??$$

$h(x) = x^2 + 1$ and $g(x) = x^2$

Backpropagation uses the chain rule

$$f(x) = g\big(h(x)\big)$$

Chain rule: $$\frac{df}{dx} = \frac{dg}{dh} \cdot \frac{dh}{dx}$$

$$f(x) = (x^2 + 1)^2 = g\big(h(x)\big) \quad ??$$

$h(x) = x^2 + 1$ and $g(x) = x^2$

$$\frac{df}{dx} = 2(x^2 + 1) \cdot 2x$$

Backpropagation uses the chain rule

$$f(x) = g\big(h(x)\big)$$

Chain rule: $$\frac{df}{dx} = \frac{dg}{dh} \cdot \frac{dh}{dx}$$

$$f(x) = (x^2 + 1)^2 = g\big(h(x)\big) \quad ??$$

$h(x) = x^2 + 1$ and $g(x) = x^2$

$$\frac{df}{dx} = 2(x^2 + 1) \cdot 2x$$

$$f(x) = \ln(ax) = g\big(h(x)\big)$$

$h(x) = ax$ and $g(x) = \ln(x)$

Backpropagation uses the chain rule

$$f(x) = g\big(h(x)\big)$$

Chain rule: $$\frac{df}{dx} = \frac{dg}{dh} \cdot \frac{dh}{dx}$$

$$f(x) = (x^2 + 1)^2 = g\big(h(x)\big) \quad ??$$

$h(x) = x^2 + 1$ and $g(x) = x^2$

$$\frac{df}{dx} = 2(x^2 + 1) \cdot 2x$$

$$f(x) = \ln(ax) = g\big(h(x)\big)$$

$h(x) = ax$ and $g(x) = \ln(x)$

$$\frac{df}{dx} = \frac{1}{ax} \cdot a = \frac{1}{x}$$

Backpropagation uses the chain rule

$$f(x) = g\big(h(x)\big)$$

Chain rule: $$\frac{df}{dx} = \frac{dg}{dh} \cdot \frac{dh}{dx}$$

$$f(x) = (x^2 + 1)^2 = g\big(h(x)\big) \quad ??$$

$h(x) = x^2 + 1$ and $g(x) = x^2$

$$\frac{df}{dx} = 2(x^2 + 1) \cdot 2x$$

$$f(x) = \ln(ax) = g\big(h(x)\big)$$

$h(x) = ax$ and $g(x) = \ln(x)$

$$\frac{df}{dx} = \frac{1}{ax} \cdot a = \frac{1}{x}$$

$$\ln(ax) = \ln(a) + \ln(x)$$

Multidimensional chain rule

$$\mathbf{x} \in \mathbb{R}^{\ell}$$
$$\mathbf{a} = h(\mathbf{x}), \quad \mathbb{R}^{\ell} \to \mathbb{R}^{n}$$
$$\mathbf{b} = g(\mathbf{a}), \quad \mathbb{R}^{n} \to \mathbb{R}^{m}$$

$$\underset{\mathbb{R}^{\ell \times m}}{\frac{\partial \mathbf{b}}{\partial \mathbf{x}}} = \underset{\mathbb{R}^{\ell \times n}}{\boxed{\frac{\partial \mathbf{a}}{\partial \mathbf{x}}}} \cdot \underset{\mathbb{R}^{n \times m}}{\boxed{\frac{\partial \mathbf{b}}{\partial \mathbf{a}}}}$$

Multidimensional chain rule

$$\mathbf{x} \in \mathbb{R}^{\ell}$$

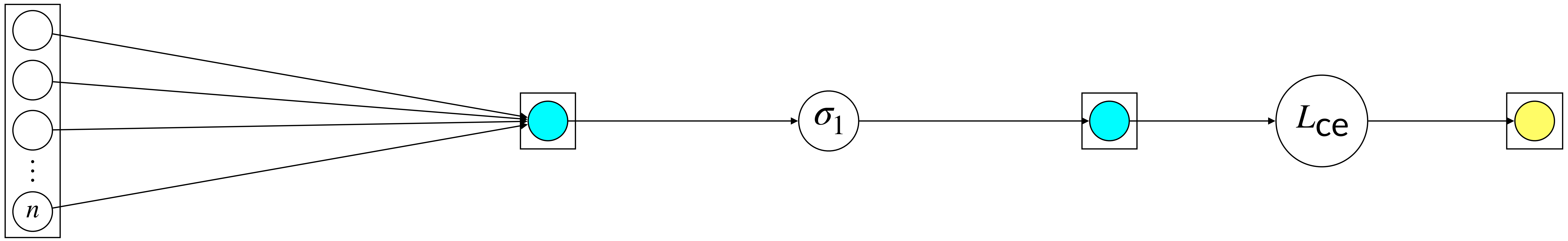
$$\mathbf{a} = h(\mathbf{x}), \quad \mathbb{R}^{\ell} \to \mathbb{R}^{n}$$

$$\mathbf{b} = g(\mathbf{a}), \quad \mathbb{R}^{n} \to \mathbb{R}^{m}$$

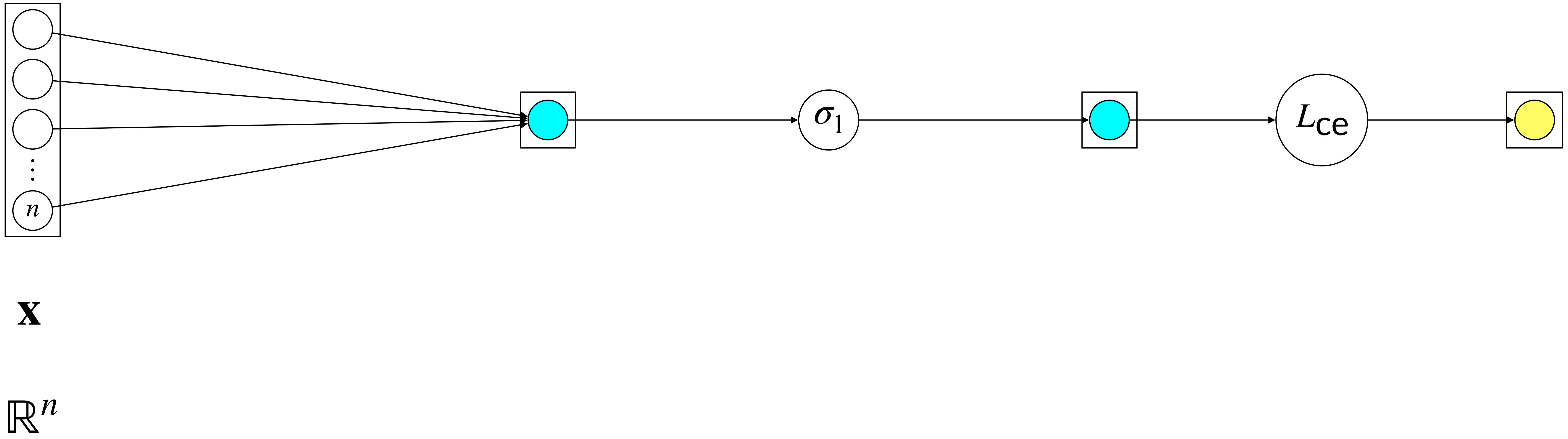
$$\begin{array}{ccccc} \mathbb{R}^{\ell\times m} & & \mathbb{R}^{\ell\times n} & & \mathbb{R}^{n\times m} \\ \dfrac{\partial \mathbf{b}}{\partial \mathbf{x}} & = & \boxed{\dfrac{\partial \mathbf{a}}{\partial \mathbf{x}}} & \cdot & \boxed{\dfrac{\partial \mathbf{b}}{\partial \mathbf{a}}} \end{array}$$

$$\begin{bmatrix} \dfrac{\partial b_1}{\partial a_1} & \dfrac{\partial b_2}{\partial a_1} & \dfrac{\partial b_3}{\partial a_1} & \cdots & \dfrac{\partial b_m}{\partial a_1} \\ \dfrac{\partial b_1}{\partial a_2} & \dfrac{\partial b_2}{\partial a_2} & \dfrac{\partial b_3}{\partial a_2} & \cdots & \dfrac{\partial b_m}{\partial a_2} \\ \dfrac{\partial b_1}{\partial a_3} & \dfrac{\partial b_2}{\partial a_3} & \dfrac{\partial b_3}{\partial a_3} & \cdots & \dfrac{\partial b_m}{\partial a_3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial b_1}{\partial a_n} & \dfrac{\partial b_2}{\partial a_n} & \dfrac{\partial b_3}{\partial a_n} & \cdots & \dfrac{\partial b_m}{\partial a_n} \end{bmatrix}$$

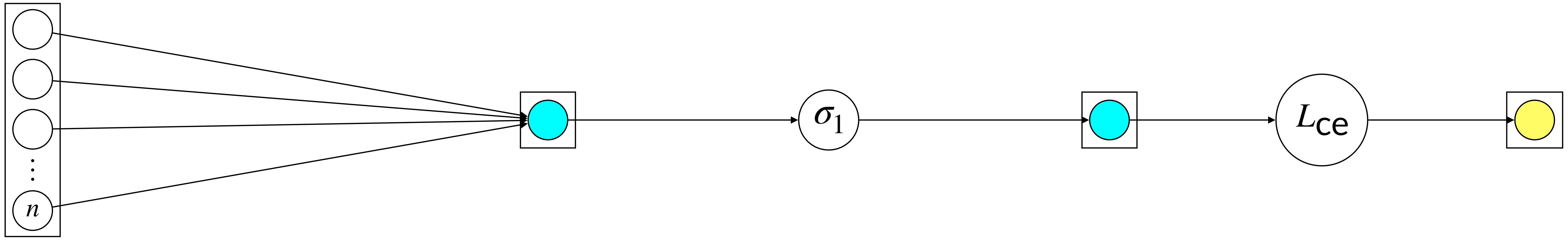
Backprop and the chain rule in a 1-dimensional NN



Backprop and the chain rule in a 1-dimensional NN



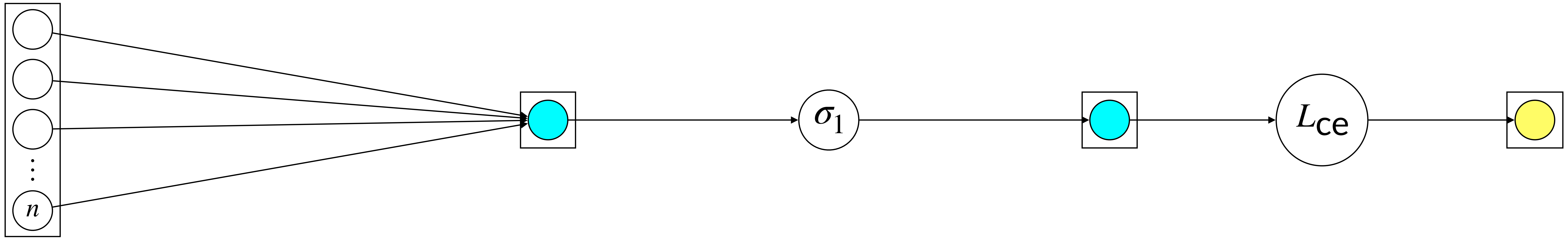
Backprop and the chain rule in a 1-dimensional NN



$\mathbf{x}$ → $z = \mathbf{w} \cdot \mathbf{x}$

$\mathbb{R}^n$ $\mathbb{R}$

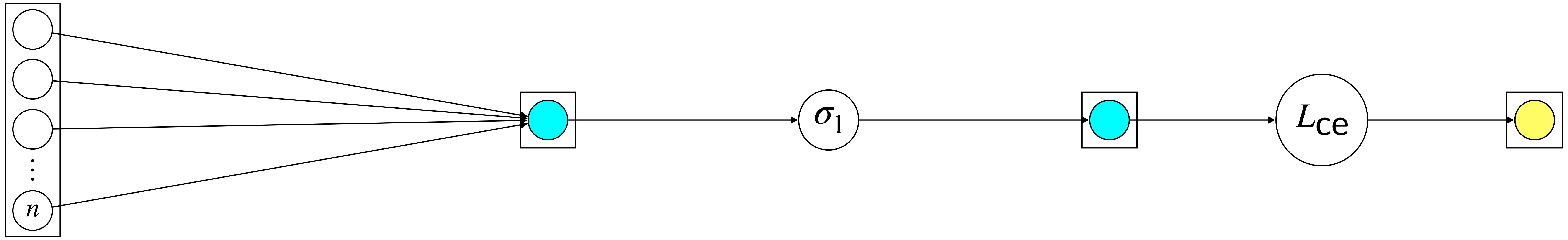
Backprop and the chain rule in a 1-dimensional NN



$$\mathbf{x} \longrightarrow z = \mathbf{w} \cdot \mathbf{x} \longrightarrow a = \sigma(z)$$

$$\mathbb{R}^n \qquad \mathbb{R} \qquad \mathbb{R}$$

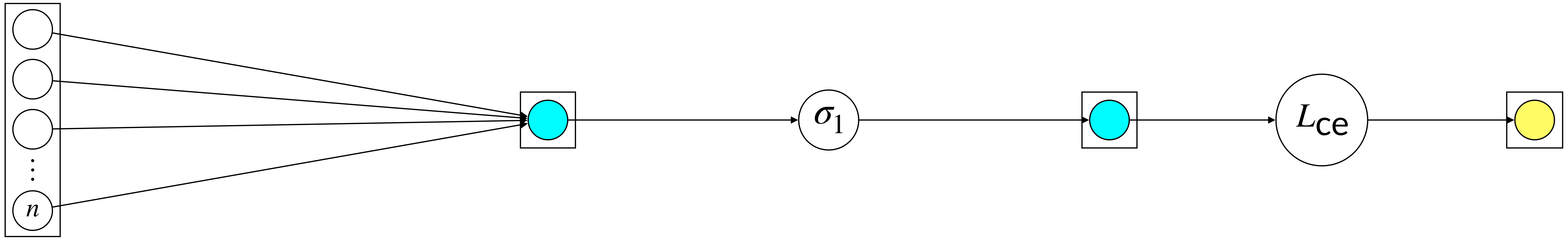
Backprop and the chain rule in a 1-dimensional NN



$\mathbf{x}$ → $z = \mathbf{w} \cdot \mathbf{x}$ → $a = \sigma(z)$ → $\ell = -\log(a)$

$\mathbb{R}^n$ $\mathbb{R}$ $\mathbb{R}$ $\mathbb{R}$

Backprop and the chain rule in a 1-dimensional NN



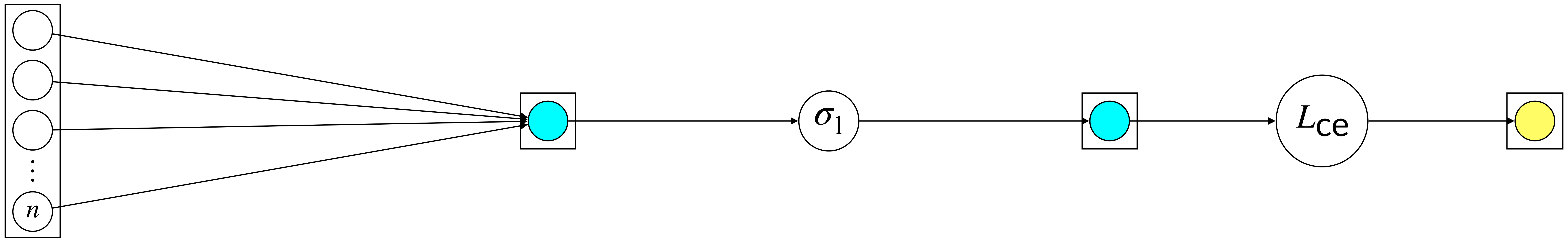
$$\frac{\partial \ell}{\partial \mathbf{w}_i} = ???$$

$\mathbf{x}$ → $z = \mathbf{w} \cdot \mathbf{x}$ → $a = \sigma(z)$ → $\ell = -\log(a)$

$\mathbb{R}^n$ $\mathbb{R}$ $\mathbb{R}$ $\mathbb{R}$

Backprop and the chain rule in a 1-dimensional NN

$$\frac{\partial \ell}{\partial \ell} = 1$$



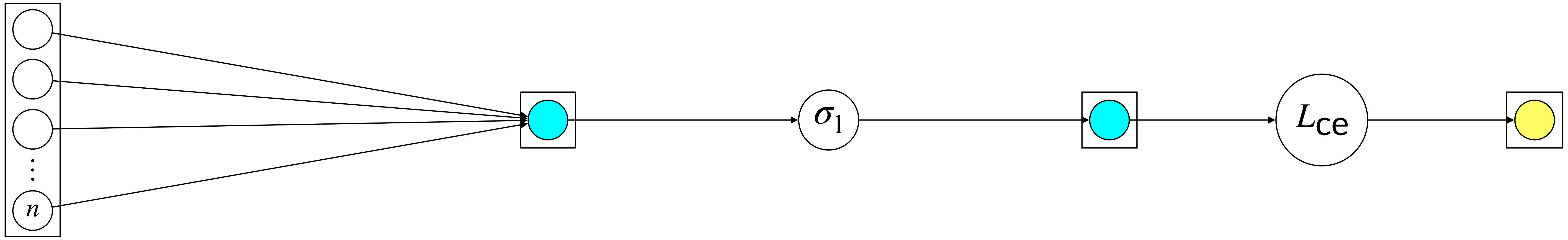
$$\boxed{\frac{\partial \ell}{\partial \mathbf{w}_i} = ???}$$

$\mathbf{x}$ $\rightarrow$ $z = \mathbf{w} \cdot \mathbf{x}$ $\rightarrow$ $a = \sigma(z)$ $\rightarrow$ $\ell = -\log(a)$

$\mathbb{R}^n$ $\qquad$ $\mathbb{R}$ $\qquad$ $\mathbb{R}$ $\qquad$ $\mathbb{R}$

Backprop and the chain rule in a 1-dimensional NN

$$\frac{\partial \ell}{\partial a} = \frac{\partial \ell}{\partial a} \cdot \frac{\partial \ell}{\partial \ell} \longleftarrow \frac{\partial \ell}{\partial \ell} = 1$$

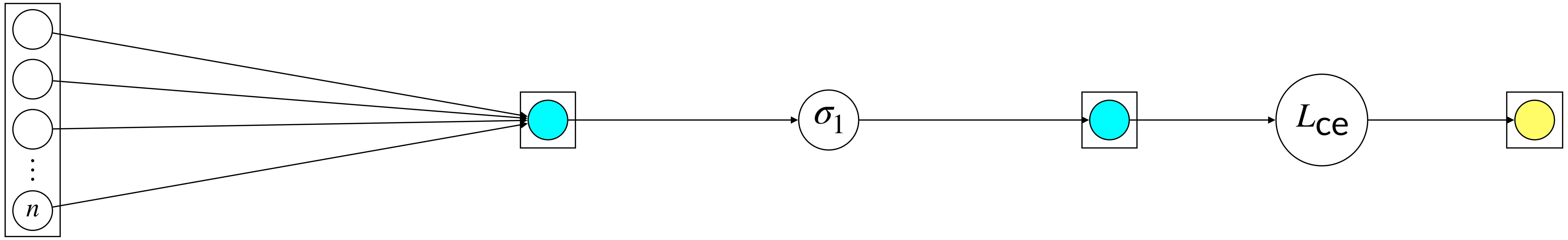


$$\frac{\partial \ell}{\partial \mathbf{w}_i} = ???$$

$$\mathbf{x} \longrightarrow z = \mathbf{w} \cdot \mathbf{x} \longrightarrow a = \sigma(z) \longrightarrow \ell = -\log(a)$$

$\mathbb{R}^n$ $\qquad$ $\mathbb{R}$ $\qquad$ $\mathbb{R}$ $\qquad$ $\mathbb{R}$

$$\frac{\partial \ell}{\partial z} = \frac{\partial a}{\partial z} \cdot \frac{\partial \ell}{\partial a} \longleftarrow \frac{\partial \ell}{\partial a} = \frac{\partial \ell}{\partial a} \cdot \frac{\partial \ell}{\partial \ell} \longleftarrow \frac{\partial \ell}{\partial \ell} = 1$$



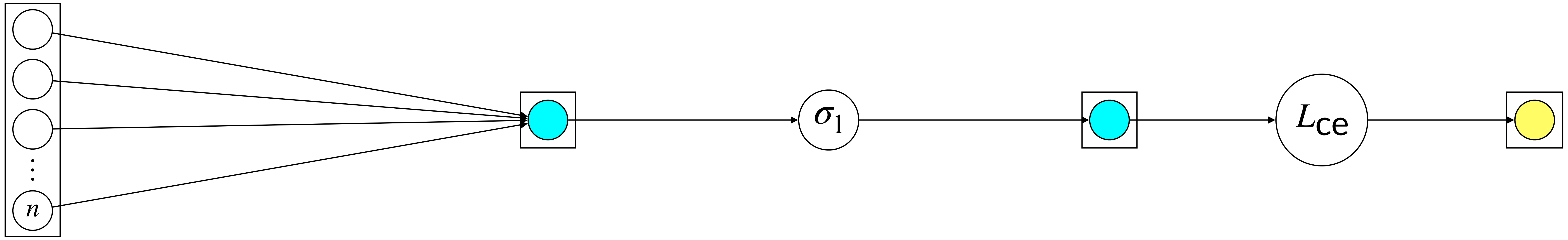
$$\boxed{\frac{\partial \ell}{\partial \mathbf{w}_i} = ???}$$

$$\mathbf{x} \longrightarrow z = \mathbf{w} \cdot \mathbf{x} \longrightarrow a = \sigma(z) \longrightarrow \ell = -\log(a)$$

$\mathbb{R}^n$ $\mathbb{R}$ $\mathbb{R}$ $\mathbb{R}$

Backprop and the chain rule in a 1-dimensional NN

$$\frac{\partial \ell}{\partial \mathbf{w}_i} = \frac{\partial z}{\partial w_i} \cdot \frac{\partial \ell}{\partial z} \quad \longleftarrow \quad \frac{\partial \ell}{\partial z} = \frac{\partial a}{\partial z} \cdot \frac{\partial \ell}{\partial a} \quad \longleftarrow \quad \frac{\partial \ell}{\partial a} = \frac{\partial \ell}{\partial a} \cdot \frac{\partial \ell}{\partial \ell} \quad \longleftarrow \quad \frac{\partial \ell}{\partial \ell} = 1$$



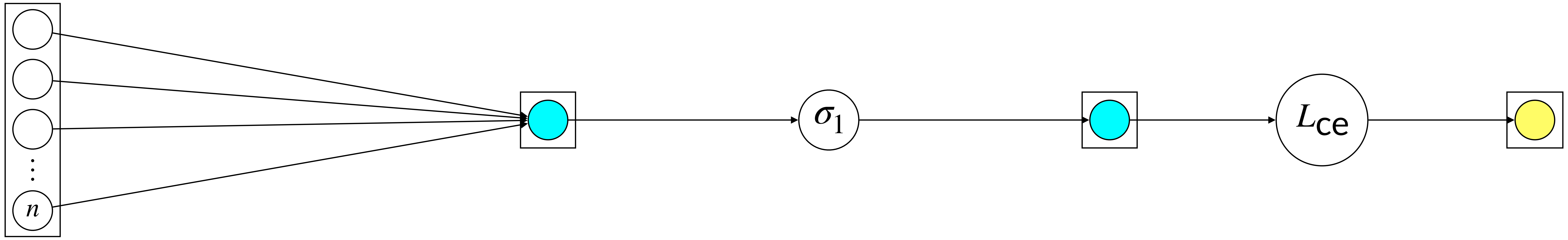
$$\boxed{\frac{\partial \ell}{\partial \mathbf{w}_i} = ???}$$

$$\mathbf{x} \quad \longrightarrow \quad z = \mathbf{w} \cdot \mathbf{x} \quad \longrightarrow \quad a = \sigma(z) \quad \longrightarrow \quad \ell = -\log(a)$$

$$\mathbb{R}^n \qquad \mathbb{R} \qquad \mathbb{R} \qquad \mathbb{R}$$

Backprop and the chain rule in a 1-dimensional NN

$$\frac{\partial \ell}{\partial \mathbf{w}_i} = \frac{\partial z}{\partial w_i} \cdot \boxed{\frac{\partial \ell}{\partial z}} \longleftarrow \boxed{\frac{\partial \ell}{\partial z}} = \frac{\partial a}{\partial z} \cdot \boxed{\frac{\partial \ell}{\partial a}} \longleftarrow \boxed{\frac{\partial \ell}{\partial a}} = \frac{\partial \ell}{\partial a} \cdot \boxed{\frac{\partial \ell}{\partial \ell}} \longleftarrow \boxed{\frac{\partial \ell}{\partial \ell}} = 1$$

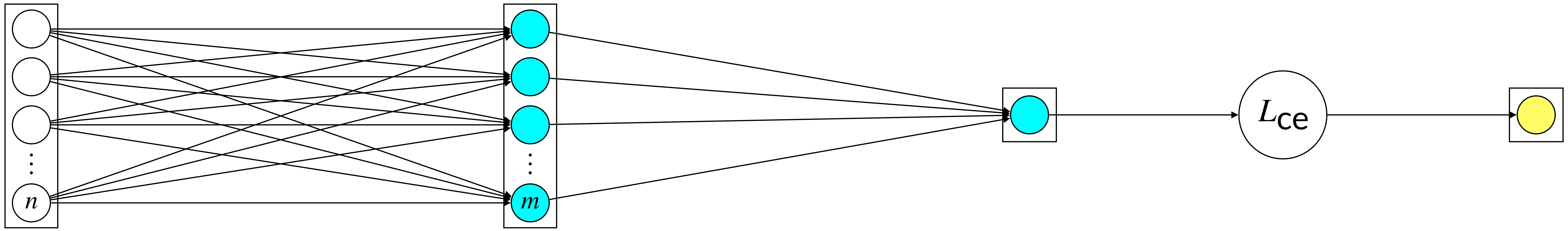


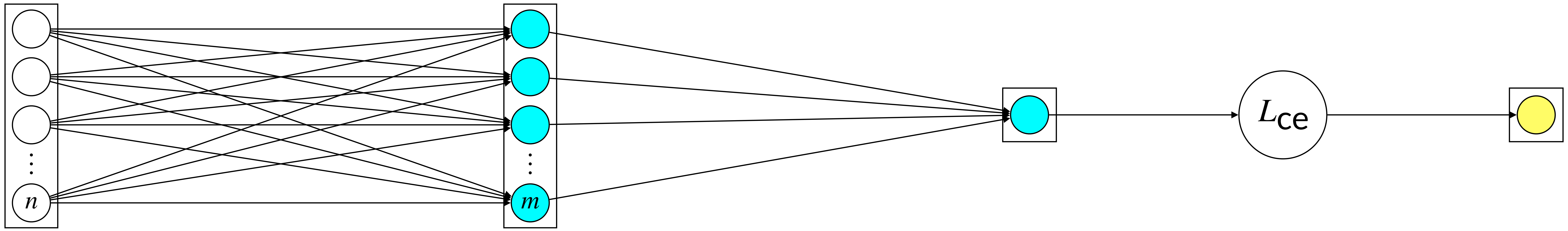
$$\frac{\partial \ell}{\partial \mathbf{w}_i} = ???$$

$$\mathbf{x} \rightarrow z = \mathbf{w} \cdot \mathbf{x} \rightarrow a = \sigma(z) \rightarrow \ell = -\log(a)$$

$$\mathbb{R}^n \qquad \mathbb{R} \qquad \mathbb{R} \qquad \mathbb{R}$$

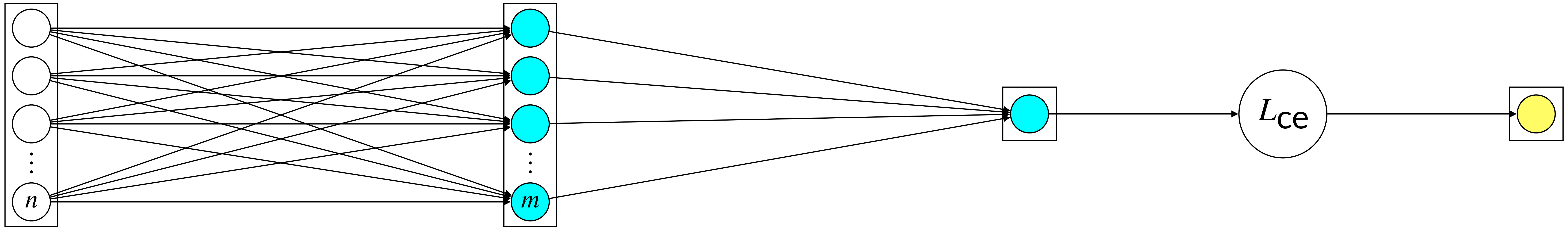
Backprop and the chain rule in multiple dimensions





$\mathbf{a}^{[0]}$

$\mathbb{R}^n$

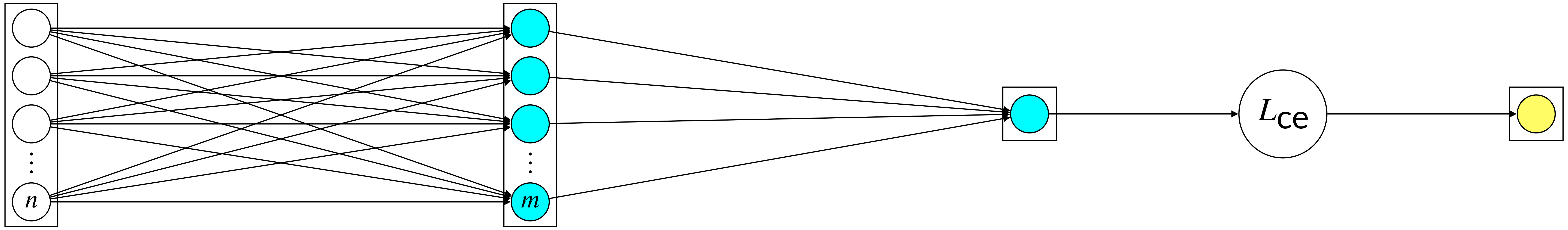


$\mathbf{a}^{[0]}$ → $\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$

$\mathbf{a}^{[1]} = \sigma\left(\mathbf{z}^{[1]}\right)$

$\mathbb{R}^n$ $\mathbb{R}^m$

Backprop and the chain rule in multiple dimensions

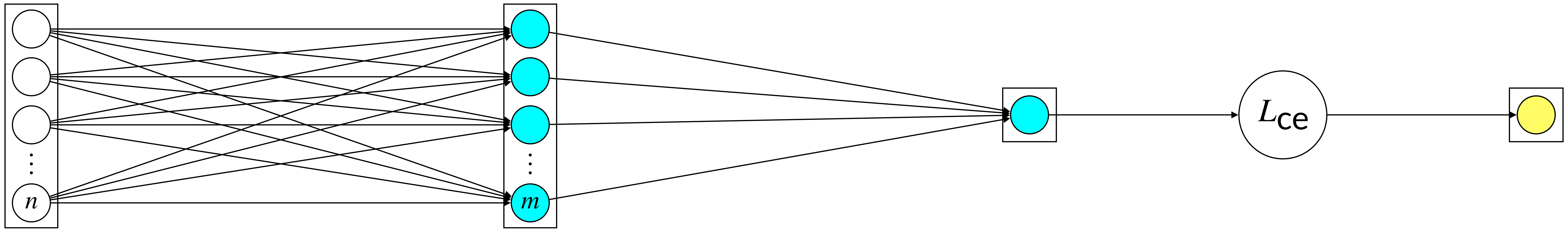


$\mathbf{a}^{[0]}$ → $\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$ → $z^{[2]} = \mathbf{w}^{[2]}\mathbf{a}^{[1]}$

$\mathbf{a}^{[1]} = \sigma\left(\mathbf{z}^{[1]}\right)$ $\quad$ $a^{[2]} = \sigma\left(z^{[2]}\right)$

$\mathbb{R}^n$ $\quad$ $\mathbb{R}^m$ $\quad$ $\mathbb{R}$

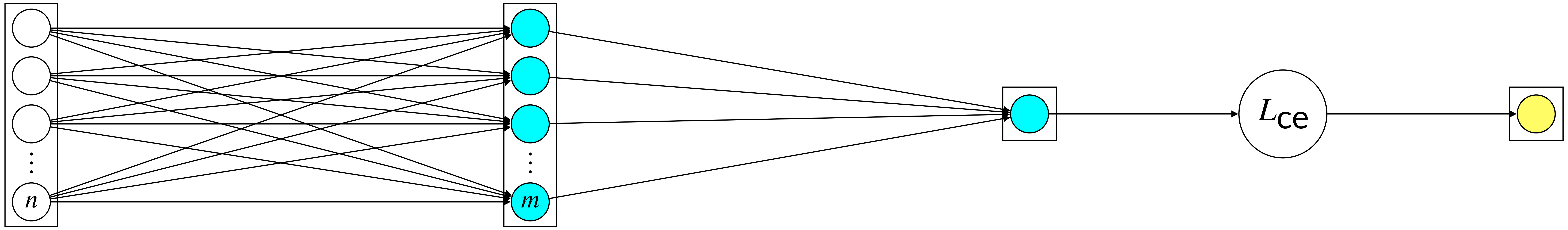
Backprop and the chain rule in multiple dimensions



$\mathbf{a}^{[0]}$ → $\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$, $\mathbf{a}^{[1]} = \sigma\left(\mathbf{z}^{[1]}\right)$ → $z^{[2]} = \mathbf{w}^{[2]}\mathbf{a}^{[1]}$, $a^{[2]} = \sigma\left(z^{[2]}\right)$ → $\ell = -\log\left(a^{[2]}\right)$

$\mathbb{R}^n$ $\quad$ $\mathbb{R}^m$ $\quad$ $\mathbb{R}$ $\quad$ $\mathbb{R}$

Backprop and the chain rule in multiple dimensions



$\mathbf{a}^{[0]}$ → $\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$, $\mathbf{a}^{[1]} = \sigma\left(\mathbf{z}^{[1]}\right)$ → $z^{[2]} = \mathbf{w}^{[2]}\mathbf{a}^{[1]}$, $a^{[2]} = \sigma\left(z^{[2]}\right)$ → $\ell = -\log\left(a^{[2]}\right)$

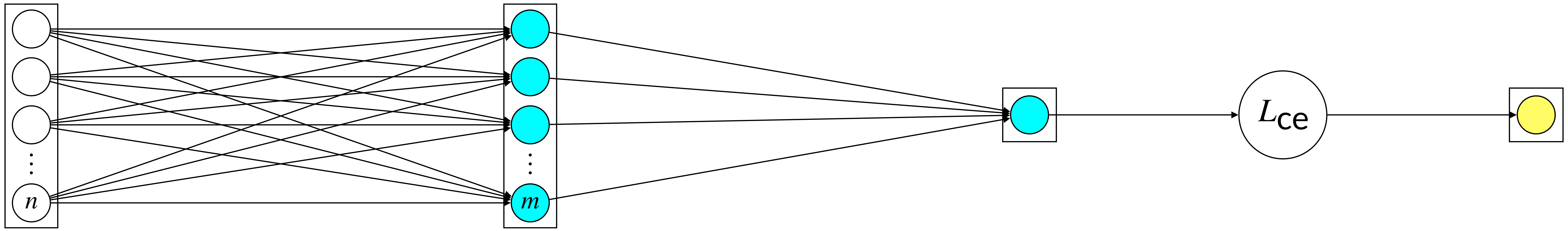
$\mathbb{R}^n$ $\quad \mathbb{R}^m \quad \mathbb{R} \quad \mathbb{R}$

$$\frac{\partial \ell}{\partial \mathbf{w}_i^{[2]}} = ???$$

$$\frac{\partial \ell}{\partial \mathbf{W}_{ij}^{[1]}} = ???$$

Backprop and the chain rule in multiple dimensions

$$\frac{\partial \ell}{\partial a^{[2]}} = \frac{\partial \ell}{\partial a^{[2]}} \cdot \frac{\partial \ell}{\partial \ell} \quad \longleftarrow \quad \frac{\partial \ell}{\partial \ell} = 1$$



$\mathbf{a}^{[0]}$ → $\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$, $\mathbf{a}^{[1]} = \sigma\left(\mathbf{z}^{[1]}\right)$ → $z^{[2]} = \mathbf{w}^{[2]}\mathbf{a}^{[1]}$, $a^{[2]} = \sigma\left(z^{[2]}\right)$ → $\ell = -\log\left(a^{[2]}\right)$

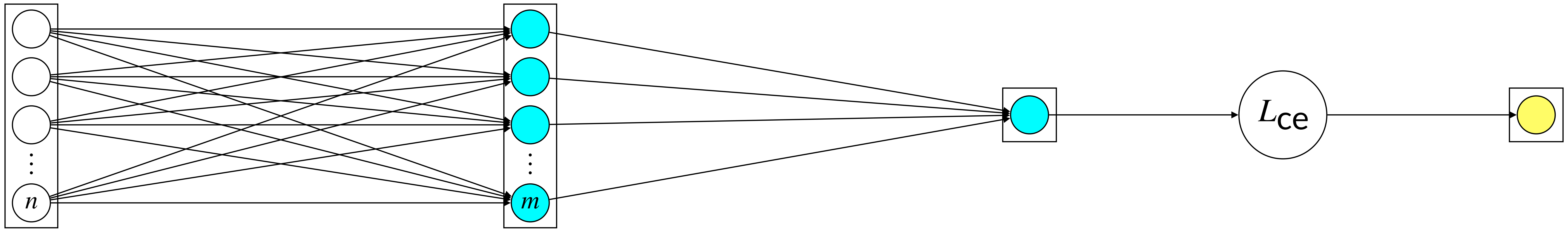
$\mathbb{R}^n$ | $\mathbb{R}^m$ | $\mathbb{R}$ | $\mathbb{R}$

$$\frac{\partial \ell}{\partial \mathbf{w}_i^{[2]}} = ???$$

$$\frac{\partial \ell}{\partial \mathbf{W}_{ij}^{[1]}} = ???$$

Backprop and the chain rule in multiple dimensions

$$\frac{\partial \ell}{\partial z^{[2]}} = \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot \frac{\partial \ell}{\partial a^{[2]}} \longleftarrow \frac{\partial \ell}{\partial a^{[2]}} = \frac{\partial \ell}{\partial a^{[2]}} \cdot \frac{\partial \ell}{\partial \ell} \longleftarrow \frac{\partial \ell}{\partial \ell} = 1$$



$$\mathbf{a}^{[0]} \longrightarrow \mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]} \longrightarrow z^{[2]} = \mathbf{w}^{[2]}\mathbf{a}^{[1]} \longrightarrow \ell = -\log\left(a^{[2]}\right)$$

$$\mathbf{a}^{[1]} = \sigma\left(\mathbf{z}^{[1]}\right) \qquad a^{[2]} = \sigma\left(z^{[2]}\right)$$

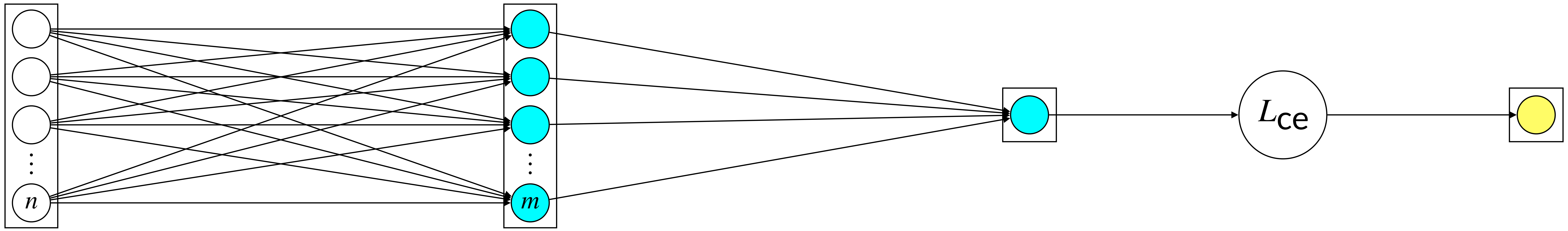
$\mathbb{R}^n$ $\qquad \mathbb{R}^m$ $\qquad \mathbb{R}$ $\qquad \mathbb{R}$

$$\frac{\partial \ell}{\partial \mathbf{w}_i^{[2]}} = ???$$

$$\frac{\partial \ell}{\partial \mathbf{W}_{ij}^{[1]}} = ???$$

Backprop and the chain rule in multiple dimensions

$$\frac{\partial \ell}{\partial \mathbf{w}_i^{[2]}} = \frac{\partial z^{[2]}}{\partial \mathbf{w}_i^{[2]}} \cdot \frac{\partial \ell}{\partial z^{[2]}} \longleftarrow \frac{\partial \ell}{\partial z^{[2]}} = \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot \frac{\partial \ell}{\partial a^{[2]}} \longleftarrow \frac{\partial \ell}{\partial a^{[2]}} = \frac{\partial \ell}{\partial a^{[2]}} \cdot \frac{\partial \ell}{\partial \ell} \longleftarrow \frac{\partial \ell}{\partial \ell} = 1$$



$$\mathbf{a}^{[0]} \longrightarrow \mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]} \longrightarrow z^{[2]} = \mathbf{w}^{[2]}\mathbf{a}^{[1]} \longrightarrow \ell = -\log\left(a^{[2]}\right)$$

$$\mathbf{a}^{[1]} = \sigma\left(\mathbf{z}^{[1]}\right) \qquad a^{[2]} = \sigma\left(z^{[2]}\right)$$

$\mathbb{R}^n$ $\qquad$ $\mathbb{R}^m$ $\qquad$ $\mathbb{R}$ $\qquad$ $\mathbb{R}$

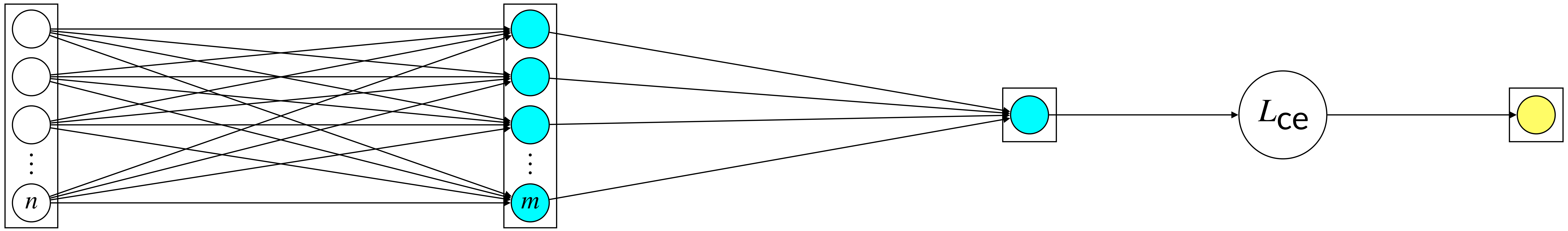
$$\frac{\partial \ell}{\partial \mathbf{w}_i^{[2]}} = ???$$

$$\frac{\partial \ell}{\partial \mathbf{W}_{ij}^{[1]}} = ???$$

Backprop and the chain rule in multiple dimensions

$$\frac{\partial \ell}{\partial \mathbf{W}^{[1]}_{ij}} = \frac{\partial \ell}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial \mathbf{a}^{[1]}_i} \cdot \frac{\partial \mathbf{a}^{[1]}_i}{\partial \mathbf{z}^{[1]}_i} \cdot \frac{\partial \mathbf{z}^{[1]}_i}{\partial \mathbf{W}^{[1]}_{ij}}$$

$$\frac{\partial \ell}{\partial \mathbf{w}^{[2]}_i} = \frac{\partial z^{[2]}}{\partial \mathbf{w}^{[2]}_i} \cdot \frac{\partial \ell}{\partial z^{[2]}} \leftarrow \frac{\partial \ell}{\partial z^{[2]}} = \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot \frac{\partial \ell}{\partial a^{[2]}} \leftarrow \frac{\partial \ell}{\partial a^{[2]}} = \frac{\partial \ell}{\partial a^{[2]}} \cdot \frac{\partial \ell}{\partial \ell} \leftarrow \frac{\partial \ell}{\partial \ell} = 1$$



$$\mathbf{a}^{[0]} \rightarrow \mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]} \rightarrow z^{[2]} = \mathbf{w}^{[2]}\mathbf{a}^{[1]} \rightarrow \ell = -\log\left(a^{[2]}\right)$$

$$\mathbf{a}^{[1]} = \sigma\left(\mathbf{z}^{[1]}\right) \qquad a^{[2]} = \sigma\left(z^{[2]}\right)$$

$\mathbb{R}^n$ $\mathbb{R}^m$ $\mathbb{R}$ $\mathbb{R}$

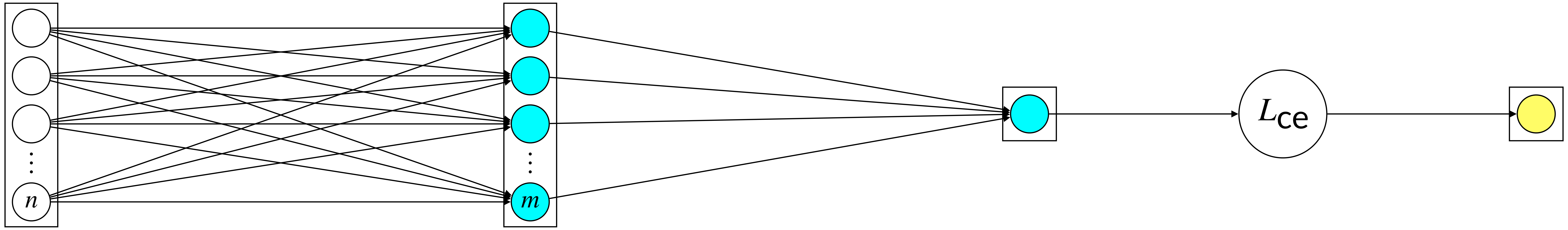
$$\frac{\partial \ell}{\partial \mathbf{w}^{[2]}_i} = ???$$

$$\frac{\partial \ell}{\partial \mathbf{W}^{[1]}_{ij}} = ???$$

Backprop and the chain rule in multiple dimensions

$$\frac{\partial \ell}{\partial \mathbf{W}_{ij}^{[1]}} = \boxed{\frac{\partial \ell}{\partial a^{[2]}}} \cdot \boxed{\frac{\partial a^{[2]}}{\partial z^{[2]}}} \cdot \frac{\partial z^{[2]}}{\partial \mathbf{a}_i^{[1]}} \cdot \frac{\partial \mathbf{a}_i^{[1]}}{\partial \mathbf{z}_i^{[1]}} \cdot \frac{\partial \mathbf{z}_i^{[1]}}{\partial \mathbf{W}_{ij}^{[1]}}$$

$$\frac{\partial \ell}{\partial \mathbf{w}_i^{[2]}} = \frac{\partial z^{[2]}}{\partial \mathbf{w}_i^{[2]}} \cdot \frac{\partial \ell}{\partial z^{[2]}} \quad \leftarrow \quad \frac{\partial \ell}{\partial z^{[2]}} = \boxed{\frac{\partial a^{[2]}}{\partial z^{[2]}}} \cdot \frac{\partial \ell}{\partial a^{[2]}} \quad \leftarrow \quad \boxed{\frac{\partial \ell}{\partial a^{[2]}}} = \frac{\partial \ell}{\partial a^{[2]}} \cdot \frac{\partial \ell}{\partial \ell} \quad \leftarrow \quad \frac{\partial \ell}{\partial \ell} = 1$$



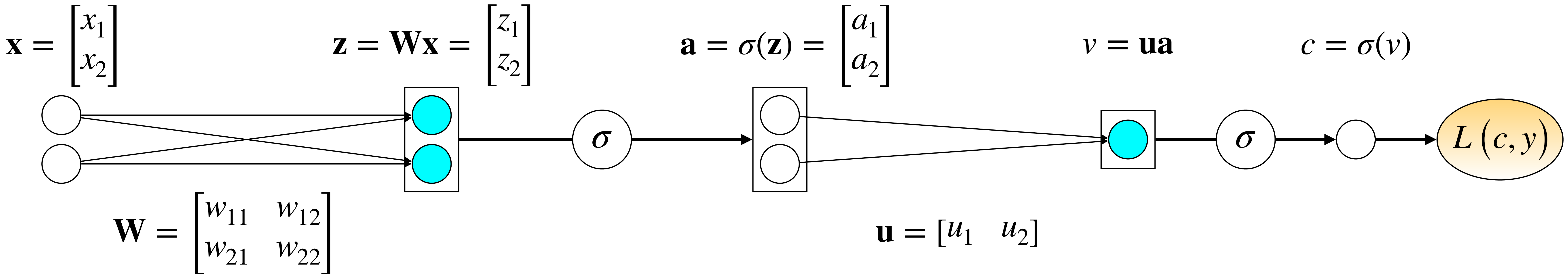
$\mathbf{a}^{[0]}$ → $\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{a}^{[0]}$, $\mathbf{a}^{[1]} = \sigma(\mathbf{z}^{[1]})$ → $z^{[2]} = \mathbf{w}^{[2]}\mathbf{a}^{[1]}$, $a^{[2]} = \sigma(z^{[2]})$ → $\ell = -\log(a^{[2]})$

$\mathbb{R}^n$ | $\mathbb{R}^m$ | $\mathbb{R}$ | $\mathbb{R}$

$$\frac{\partial \ell}{\partial \mathbf{w}_i^{[2]}} = ???$$

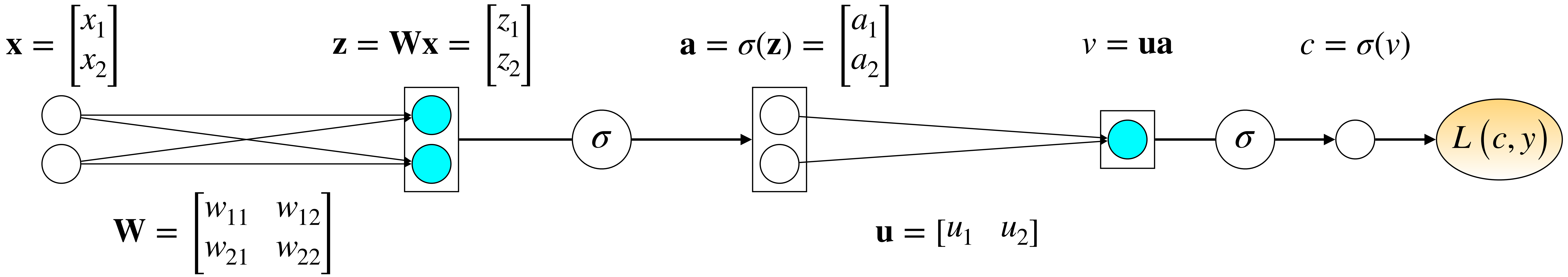
$$\frac{\partial \ell}{\partial \mathbf{W}_{ij}^{[1]}} = ???$$

Backprop (*toy*) example (1)



- Feedforward NN with 2 layers with 2 and 1 units
- Just 2 inputs and 1 output (*binary classification*)
- No bias terms and different letters used for different variables to simplify notation in upcoming slides
- Sigmoid (*logistic*) activation function everywhere

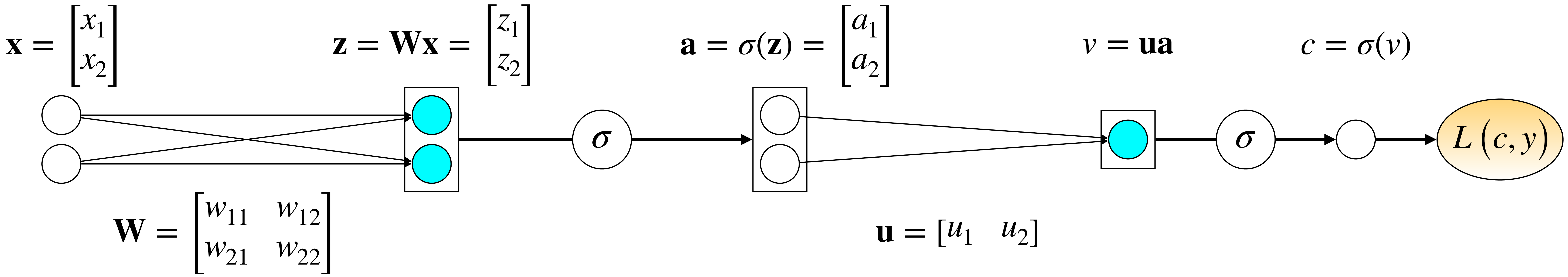
Backprop (*toy*) example (1)



- Feedforward NN with 2 layers with 2 and 1 units
- Just 2 inputs and 1 output (*binary classification*)
- No bias terms and different letters used for different variables to simplify notation in upcoming slides
- Sigmoid (*logistic*) activation function everywhere

We want to update $\mathbf{W}$ and $\mathbf{u}$ with respect to the loss $L(c, y)$

Backprop (*toy*) example (2)



Cross-entropy loss
for binary classification

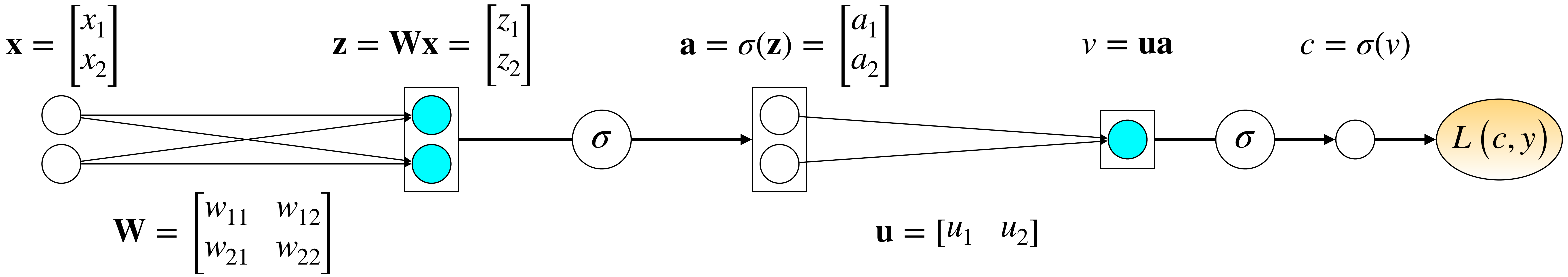
$$L(c, y) = -\left[y \ln(c) + \left(1 - y\right) \ln(1 - c)\right]$$

Understanding how a matrix operation looks like is key in getting the derivatives right

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \mathbf{W} \cdot \mathbf{x} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$z_i = \sum_{j=1}^{2} w_{ij} \cdot x_j$$

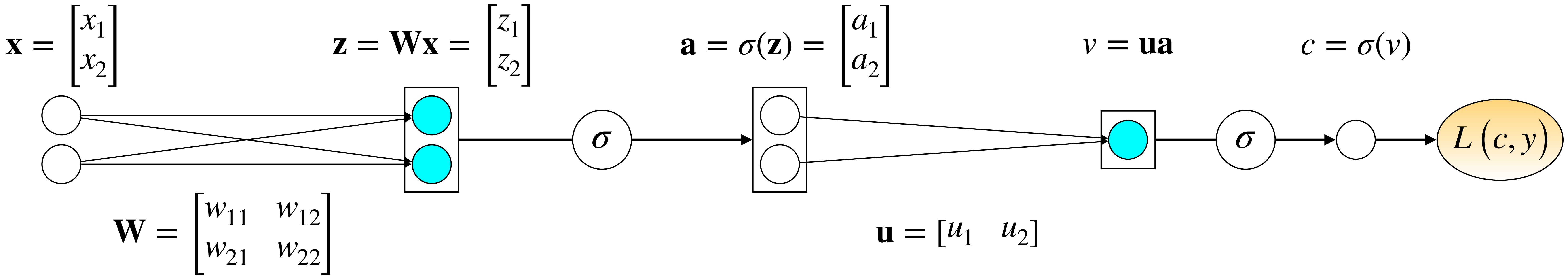
Backprop (*toy*) example (3)



The derivative of the sigmoid activation is *neat*

$$\sigma(x) = \frac{1}{1 - \exp(-x)}$$

$$\frac{d\sigma}{dx} = \frac{-\exp(-x)}{\left(1 - \exp(-x)\right)^2}$$

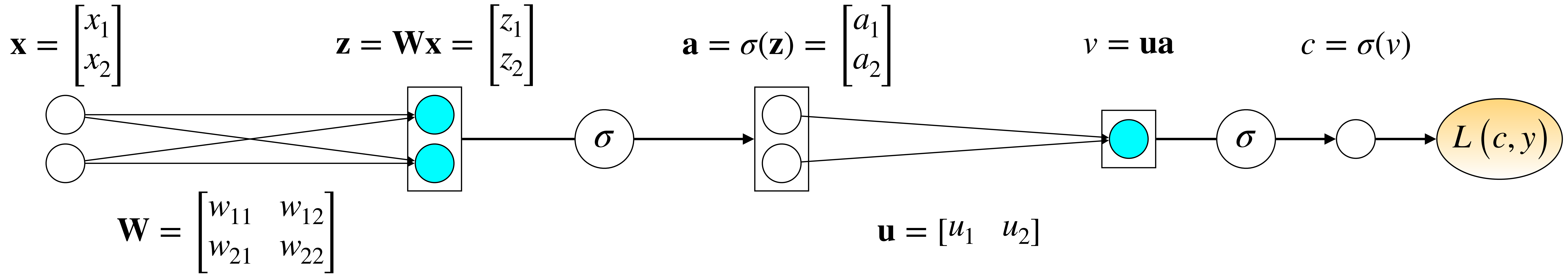


The derivative of the sigmoid activation is *neat*

$$\sigma(x) = \frac{1}{1 - \exp(-x)}$$

$$\frac{d\sigma}{dx} = \frac{-\exp(-x)}{\left(1 - \exp(-x)\right)^2} \quad = \frac{1}{1 - \exp(-x)} \cdot \frac{-\exp(-x)}{1 - \exp(-x)}$$

Backprop (*toy*) example (3)



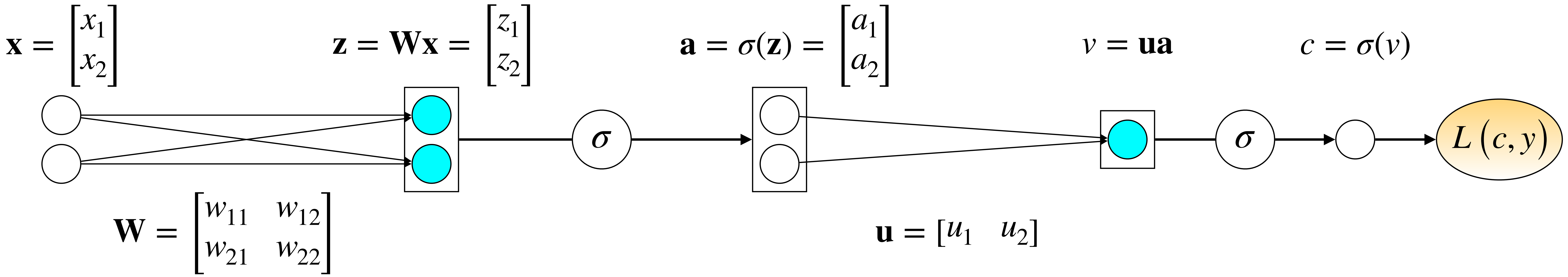
The derivative of the sigmoid activation is *neat*

$$\sigma(x) = \frac{1}{1 - \exp(-x)}$$

add and subtract 1

$$\frac{d\sigma}{dx} = \frac{-\exp(-x)}{\left(1 - \exp(-x)\right)^2} = \frac{1}{1 - \exp(-x)} \cdot \frac{-\exp(-x)}{1 - \exp(-x)} = \frac{1}{1 - \exp(-x)} \cdot \frac{\mathbf{1} - \exp(-\mathbf{x}) - \mathbf{1}}{1 - \exp(-x)}$$

Backprop (*toy*) example (3)



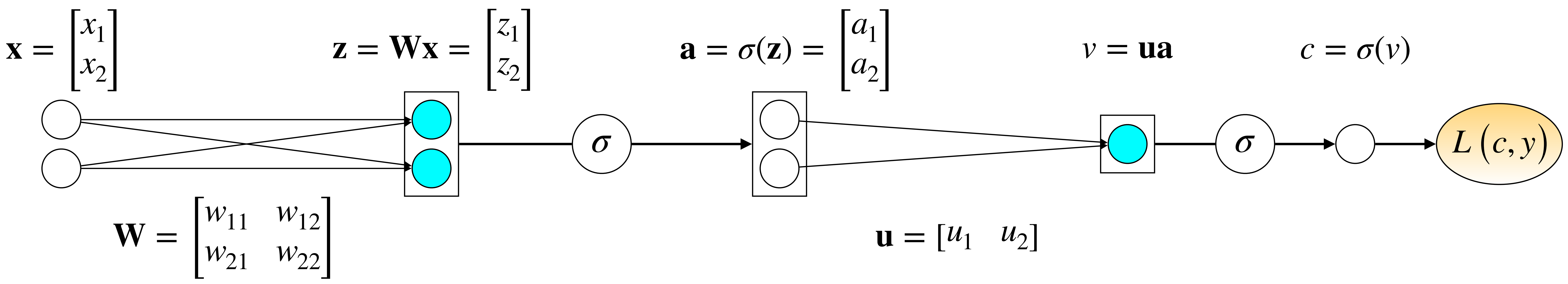
The derivative of the sigmoid activation is *neat*

$$\sigma(x) = \frac{1}{1 - \exp(-x)}$$

add and subtract 1

$$\frac{d\sigma}{dx} = \frac{-\exp(-x)}{\left(1 - \exp(-x)\right)^2} = \frac{1}{1 - \exp(-x)} \cdot \frac{-\exp(-x)}{1 - \exp(-x)} = \frac{1}{1 - \exp(-x)} \cdot \frac{\mathbf{1} - \exp(-\mathbf{x}) - \mathbf{1}}{1 - \exp(-x)}$$

$$= \frac{1}{1 - \exp(-x)} \cdot \left(\frac{1 - \exp(-x)}{1 - \exp(-x)} - \frac{1}{1 - \exp(-x)} \right)$$



The derivative of the sigmoid activation is *neat*

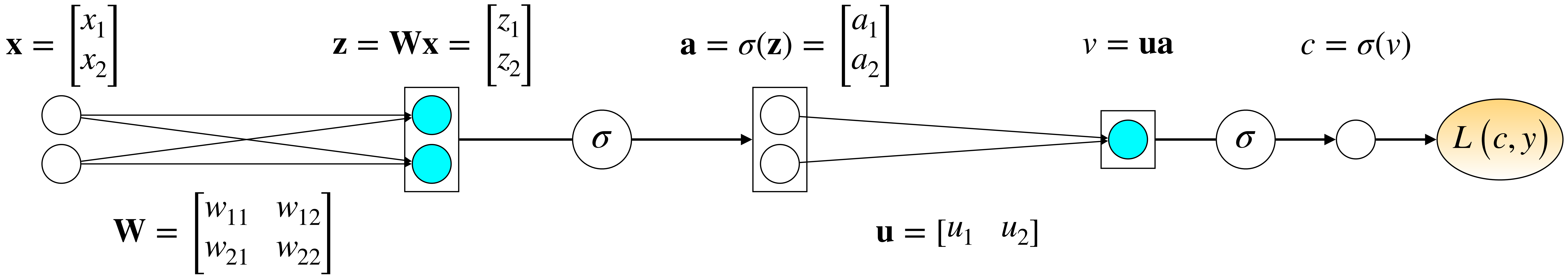
$$\sigma(x) = \frac{1}{1 - \exp(-x)}$$

add and subtract 1

$$\frac{d\sigma}{dx} = \frac{-\exp(-x)}{\left(1 - \exp(-x)\right)^2} = \frac{1}{1 - \exp(-x)} \cdot \frac{-\exp(-x)}{1 - \exp(-x)} = \frac{1}{1 - \exp(-x)} \cdot \frac{\mathbf{1} - \exp(-\mathbf{x}) - \mathbf{1}}{1 - \exp(-x)}$$

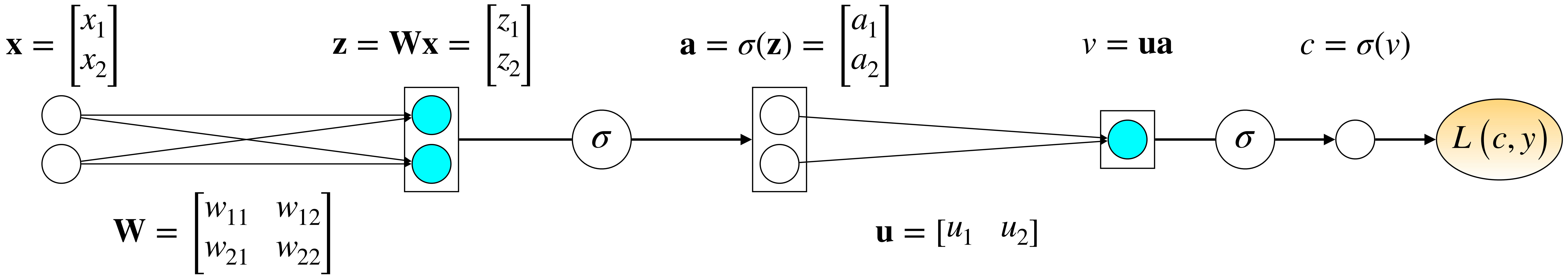
$$= \underbrace{\frac{1}{1 - \exp(-x)}}_{\sigma(x)} \cdot \left(\frac{1 - \exp(-x)}{1 - \exp(-x)} - \underbrace{\frac{1}{1 - \exp(-x)}}_{\sigma(x)} \right) = \boxed{\sigma(x) \cdot \left(1 - \sigma(x)\right)}$$

Backprop (*toy*) example (4)



We first want to obtain this $$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial u_i}$$

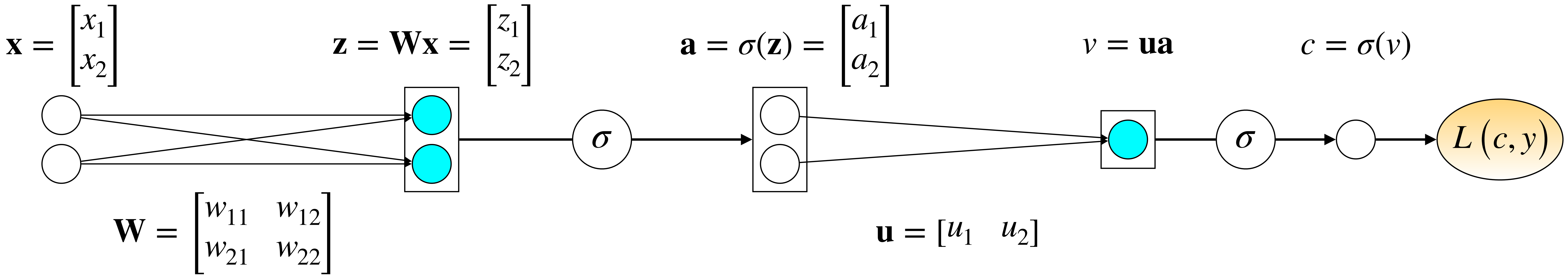
Backprop (*toy*) example (4)



We first want to obtain this $$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial u_i}$$

Chain rule $$= \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial v} \cdot \frac{\partial v}{\partial u_i}$$

Backprop (*toy*) example (5)

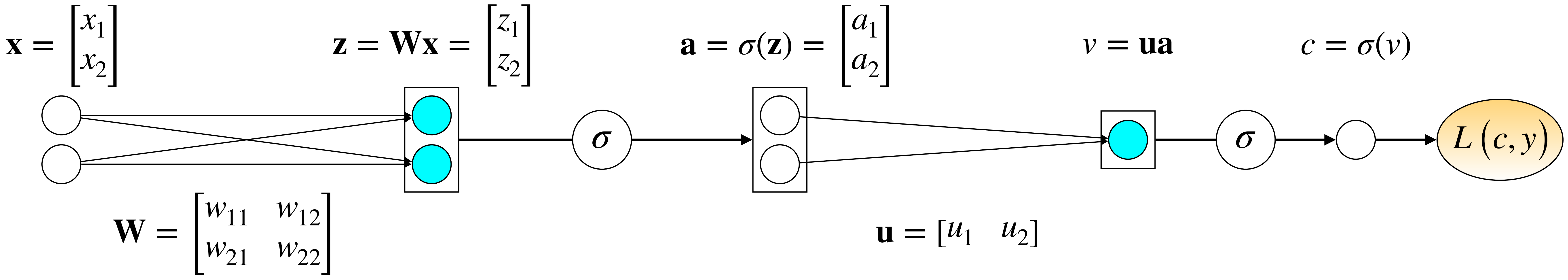


$$\frac{\partial L}{\partial u_i} = \boxed{\frac{\partial L}{\partial c}} \cdot \frac{\partial c}{\partial v} \cdot \frac{\partial v}{\partial u_i}$$

$$L(c, y) = -y \ln(c) - (1 - y) \ln(1 - c)$$

$$\frac{\partial L}{\partial c} =$$

Backprop (*toy*) example (5)

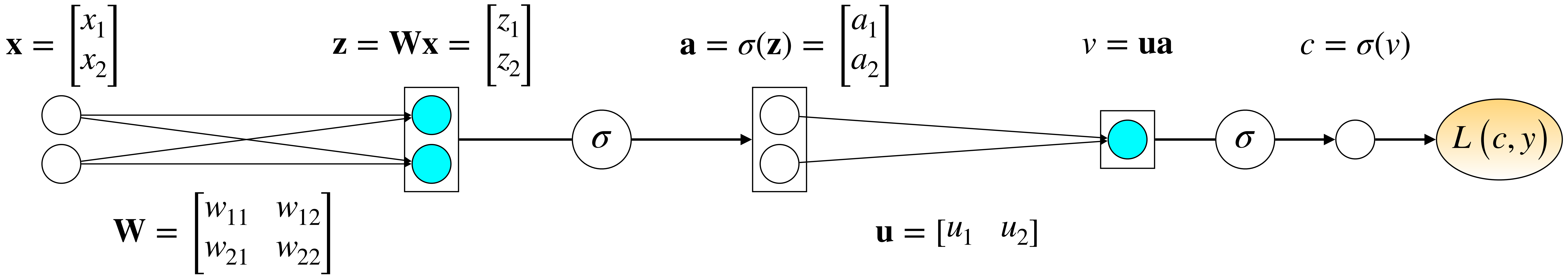


$$\frac{\partial L}{\partial u_i} = \boxed{\frac{\partial L}{\partial c}} \cdot \frac{\partial c}{\partial v} \cdot \frac{\partial v}{\partial u_i}$$

$$L(c, y) = -\, y \ln(c) - \left(1 - y\right) \ln(1 - c)$$

$$\frac{\partial L}{\partial c} = -y \cdot \frac{1}{c}$$

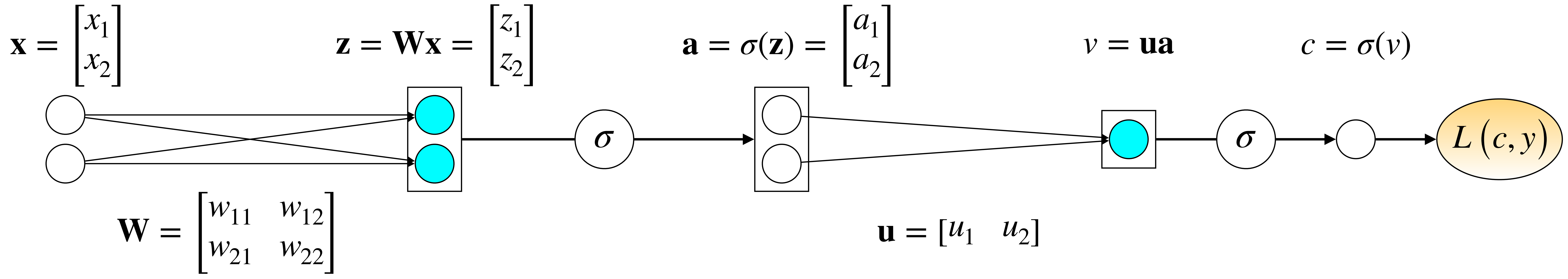
Backprop (*toy*) example (5)



$$\frac{\partial L}{\partial u_i} = \boxed{\frac{\partial L}{\partial c}} \cdot \frac{\partial c}{\partial v} \cdot \frac{\partial v}{\partial u_i}$$

$$L(c, y) = -\,y \ln(c) - \left(1 - y\right) \ln(1 - c)$$

$$\frac{\partial L}{\partial c} = -y \cdot \frac{1}{c} - \left(1 - y\right) \cdot \frac{1}{1 - c} \cdot (-1)$$

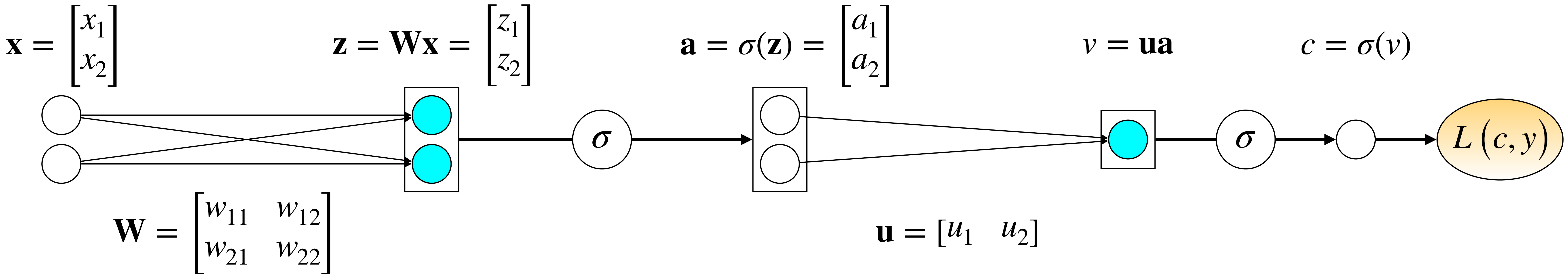


$$\frac{\partial L}{\partial u_i} = \boxed{\frac{\partial L}{\partial c}} \cdot \frac{\partial c}{\partial v} \cdot \frac{\partial v}{\partial u_i}$$

$$L(c, y) = -\, y \ln(c) - \left(1 - y\right) \ln(1 - c)$$

$$\frac{\partial L}{\partial c} = -y \cdot \frac{1}{c} - \left(1 - y\right) \cdot \frac{1}{1 - c} \cdot (-1) = \frac{c - y}{c\,(1 - c)}$$

Backprop (*toy*) example (6)

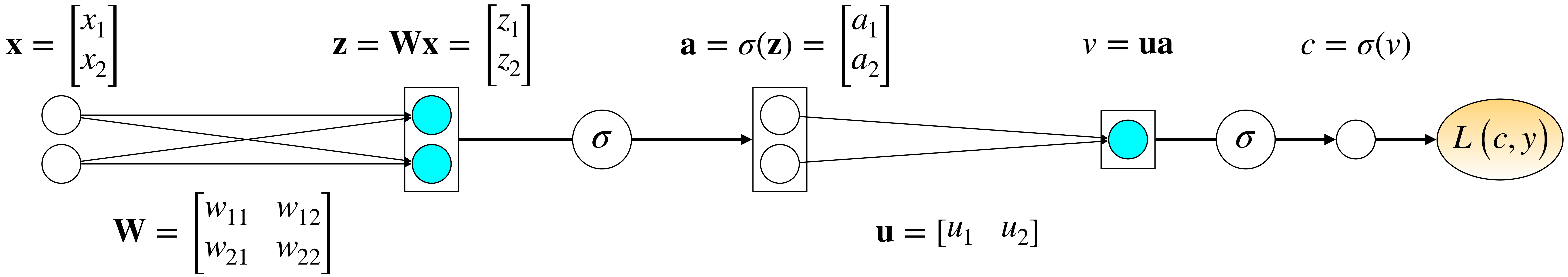


$$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \boxed{\frac{\partial c}{\partial v}} \cdot \frac{\partial v}{\partial u_i}$$

reminder $$\frac{d\sigma}{dx} = \sigma(x) \cdot \left(1 - \sigma(x)\right)$$

$$\frac{\partial c}{\partial v} =$$

Backprop (*toy*) example (6)

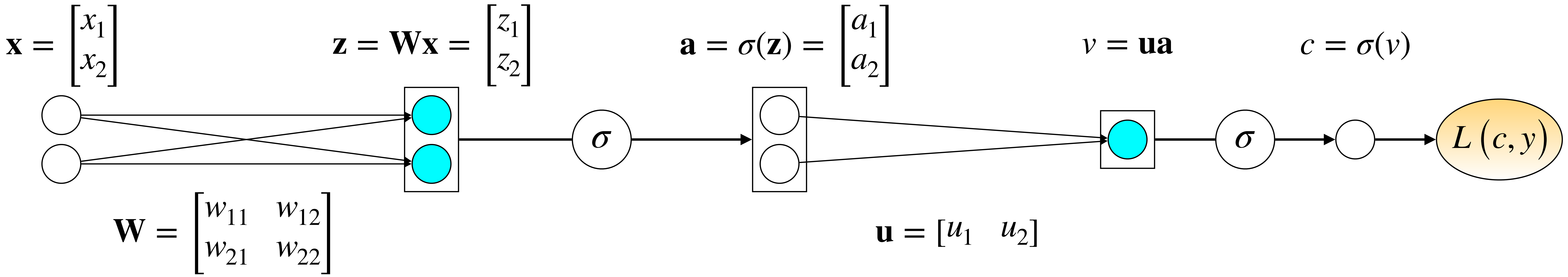


$$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \boxed{\frac{\partial c}{\partial v}} \cdot \frac{\partial v}{\partial u_i}$$

reminder $$\frac{d\sigma}{dx} = \sigma(x) \cdot \left(1 - \sigma(x)\right)$$

$$\frac{\partial c}{\partial v} = \frac{\partial \sigma(v)}{\partial v}$$

Backprop (*toy*) example (6)

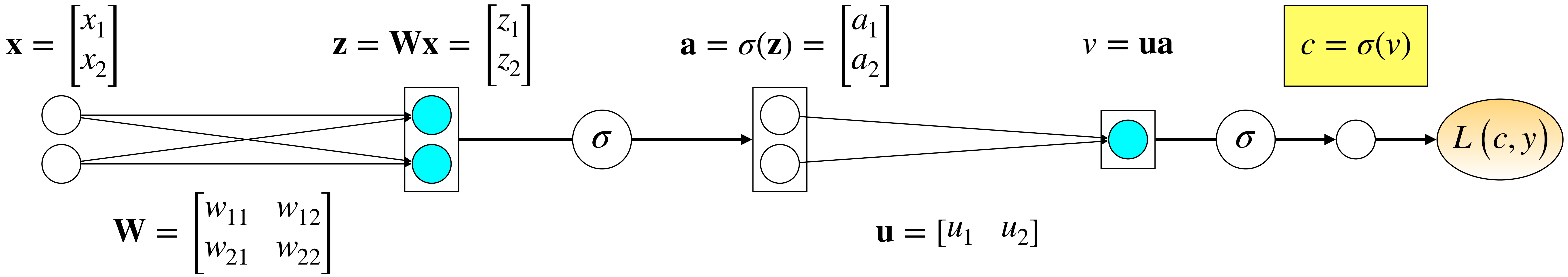


$$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \boxed{\frac{\partial c}{\partial v}} \cdot \frac{\partial v}{\partial u_i}$$

reminder $\quad \dfrac{d\sigma}{dx} = \sigma(x) \cdot \left(1 - \sigma(x)\right)$

$$\frac{\partial c}{\partial v} = \frac{\partial \sigma(v)}{\partial v} = \sigma(v) \cdot \left(1 - \sigma(v)\right)$$

Backprop (*toy*) example (6)

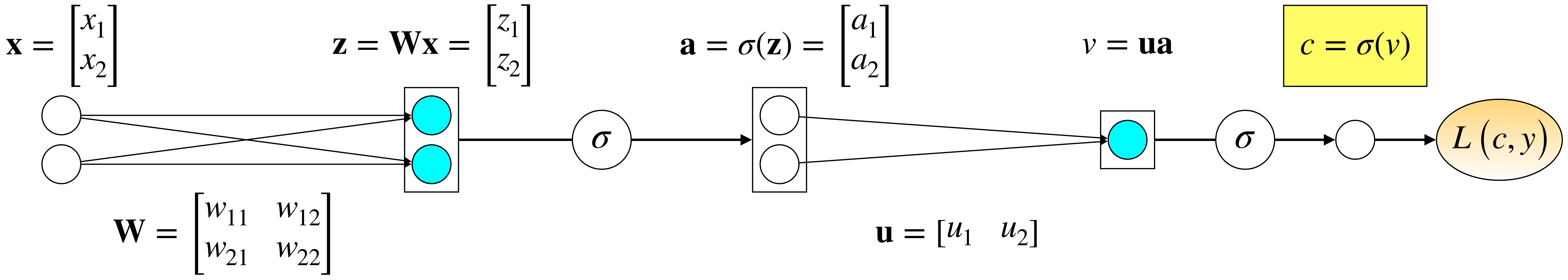


$$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \boxed{\frac{\partial c}{\partial v}} \cdot \frac{\partial v}{\partial u_i}$$

reminder $$\frac{d\sigma}{dx} = \sigma(x) \cdot \left(1 - \sigma(x)\right)$$

$$\frac{\partial c}{\partial v} = \frac{\partial \sigma(v)}{\partial v} = \sigma(v) \cdot \left(1 - \sigma(v)\right)$$

Backprop (*toy*) example (6)

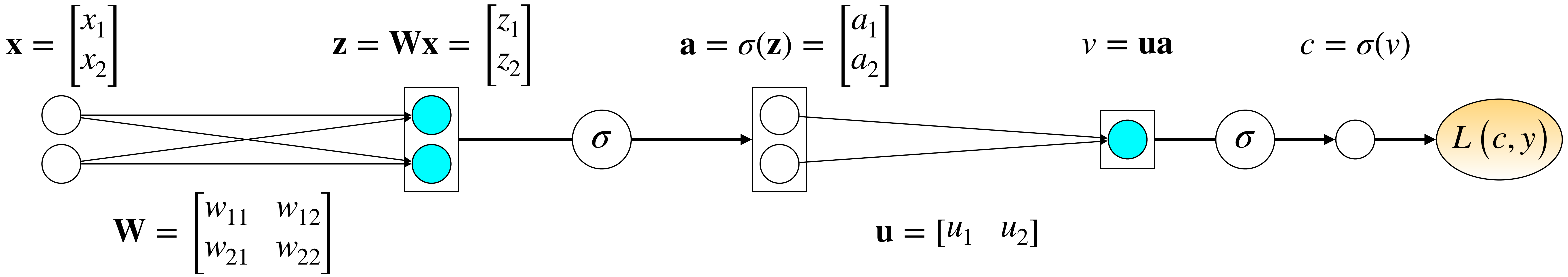


$$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \boxed{\frac{\partial c}{\partial v}} \cdot \frac{\partial v}{\partial u_i}$$

reminder $$\frac{d\sigma}{dx} = \sigma(x) \cdot \left(1 - \sigma(x)\right)$$

$$\frac{\partial c}{\partial v} = \frac{\partial \sigma(v)}{\partial v} = \sigma(v) \cdot \left(1 - \sigma(v)\right) = c \cdot (1 - c)$$

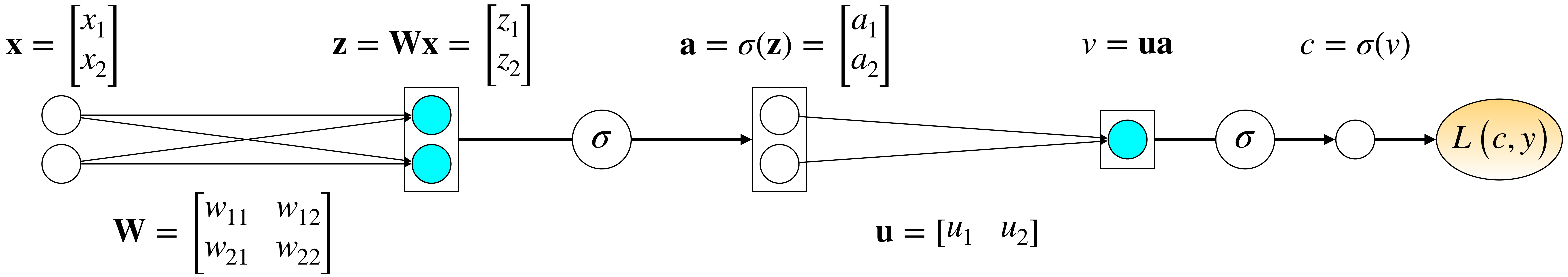
Backprop (*toy*) example (7)



$$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial v} \cdot \boxed{\frac{\partial v}{\partial u_i}}$$

$$\frac{\partial v}{\partial u_i} = \frac{\partial(\mathbf{u}\mathbf{a})}{\partial u_i}$$

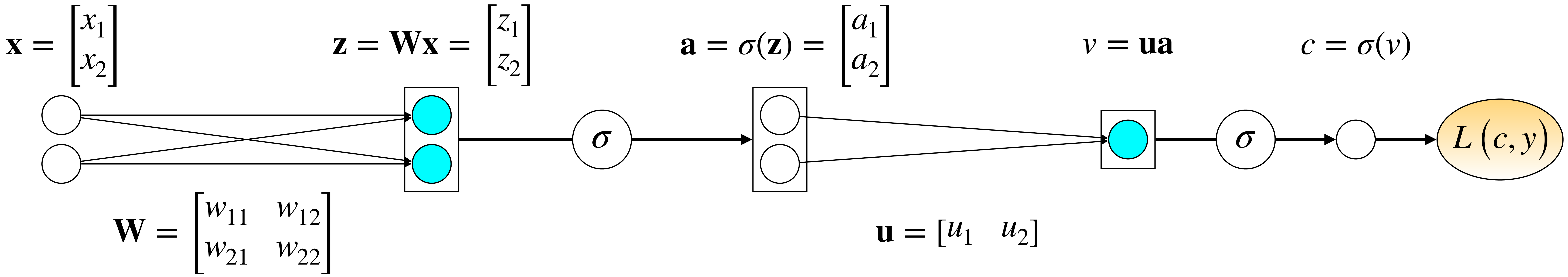
Backprop (*toy*) example (7)



$$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial v} \cdot \boxed{\frac{\partial v}{\partial u_i}}$$

$$\frac{\partial v}{\partial u_i} = \frac{\partial(\mathbf{ua})}{\partial u_i} = \frac{\partial\left(\sum_{i=1}^{2} u_i \cdot a_i\right)}{\partial u_i}$$

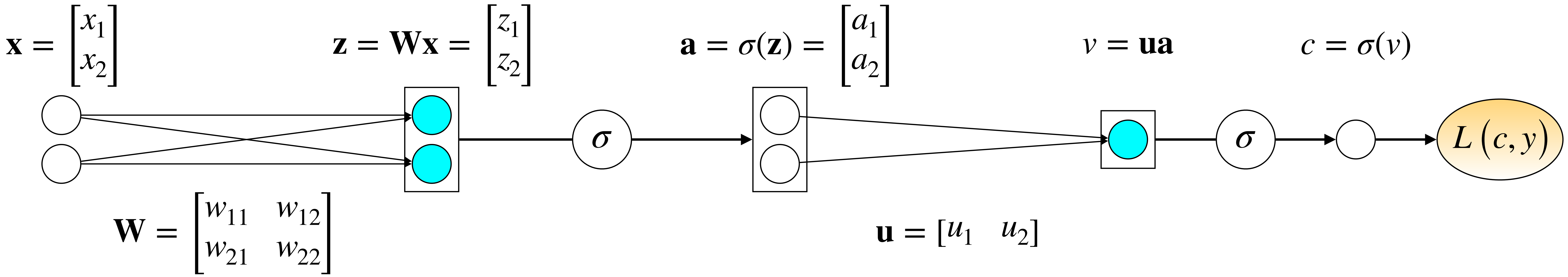
Backprop (*toy*) example (7)



$$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial v} \cdot \boxed{\frac{\partial v}{\partial u_i}}$$

$$\frac{\partial v}{\partial u_i} = \frac{\partial(\mathbf{ua})}{\partial u_i} = \frac{\partial\left(\sum_{i=1}^{2} u_i \cdot a_i\right)}{\partial u_i} = a_i$$

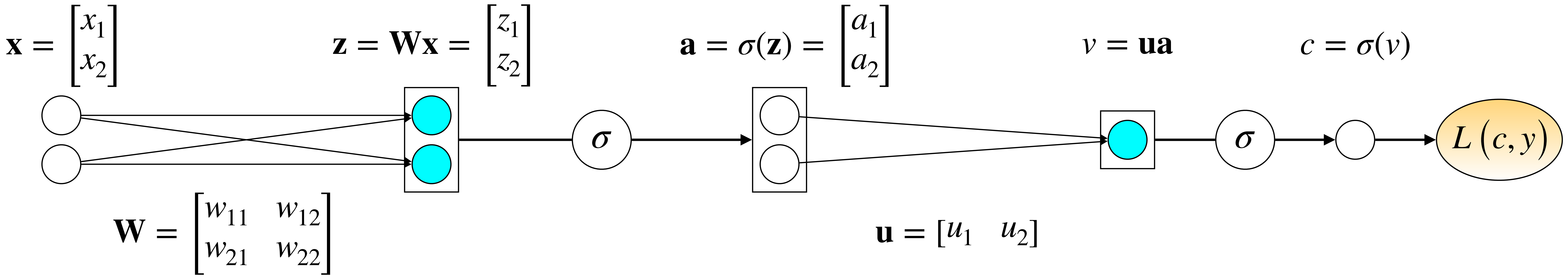
Backprop (*toy*) example (8)



$$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial v} \cdot \frac{\partial v}{\partial u_i}$$

$$\frac{\partial L}{\partial c} = \frac{c - y}{c \cdot (1 - c)} \qquad \frac{\partial c}{\partial v} = c \cdot (1 - c) \qquad \frac{\partial v}{\partial u_i} = a_i$$

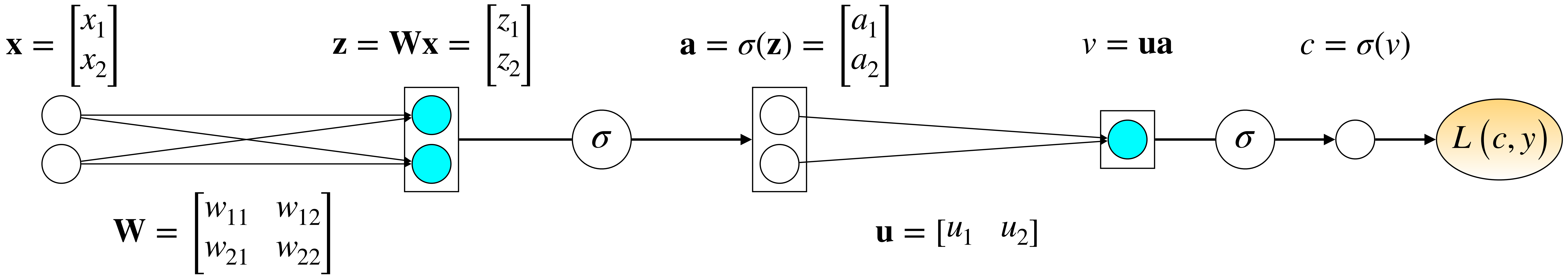
Backprop (*toy*) example (8)



$$\frac{\partial L}{\partial u_i} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial v} \cdot \frac{\partial v}{\partial u_i} \boxed{= (c - y) \cdot a_i}$$

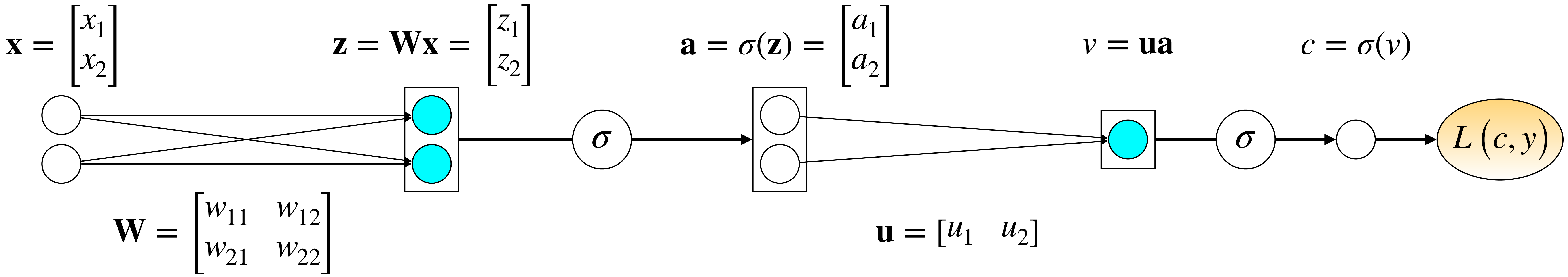
$$\frac{\partial L}{\partial c} = \frac{c - y}{c \cdot (1 - c)} \qquad \frac{\partial c}{\partial v} = c \cdot (1 - c) \qquad \frac{\partial v}{\partial u_i} = a_i$$

Backprop (*toy*) example (9)

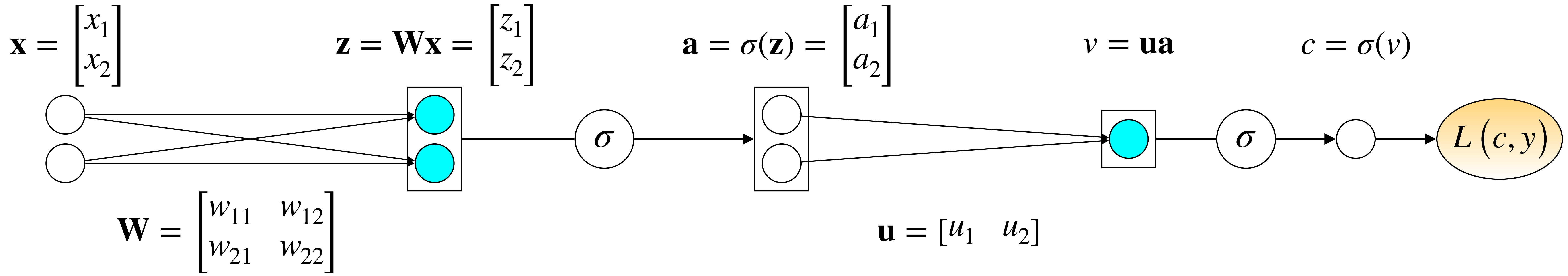


$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial v} \cdot \frac{\partial v}{\partial a_i} \cdot \frac{\partial a_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_{ij}}$$

Backprop (*toy*) example (9)

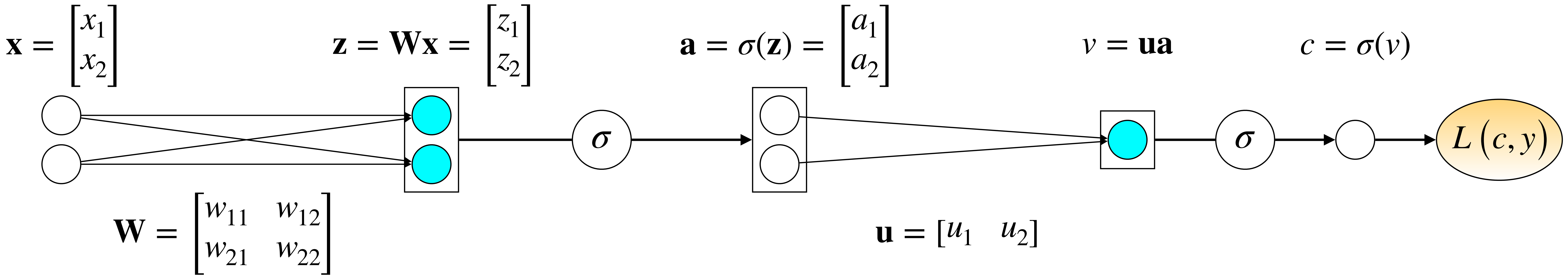


$$\frac{\partial L}{\partial w_{ij}} = \boxed{\frac{\partial L}{\partial c}} \cdot \boxed{\frac{\partial c}{\partial v}} \cdot \frac{\partial v}{\partial a_i} \cdot \frac{\partial a_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_{ij}}$$



$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial v} \cdot \frac{\partial v}{\partial a_i} \cdot \frac{\partial a_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_{ij}}$$

$$\frac{\partial v}{\partial a_i} = \frac{\partial \left(\sum_{i=1}^{2} u_i \cdot a_i \right)}{\partial a_i} = u_i$$

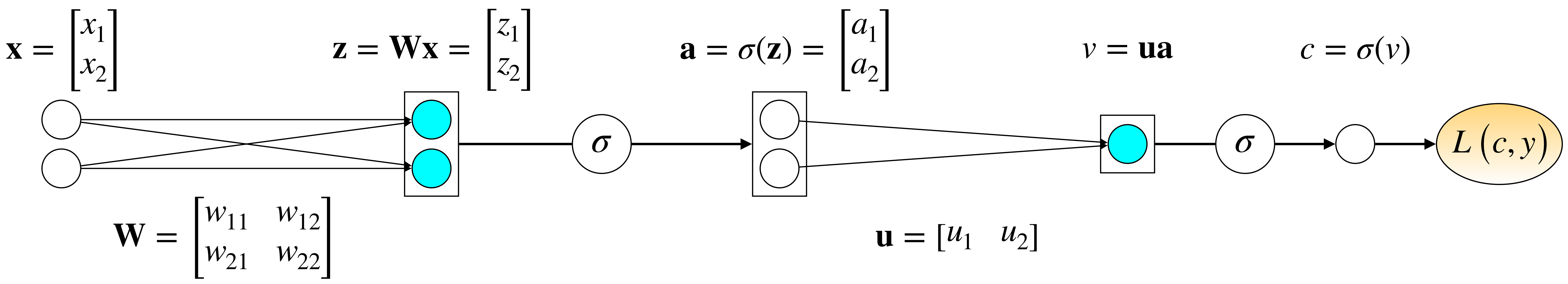


$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial v} \cdot \frac{\partial v}{\partial a_i} \cdot \frac{\partial a_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_{ij}}$$

$$\frac{\partial v}{\partial a_i} = \frac{\partial \left(\sum_{i=1}^{2} u_i \cdot a_i \right)}{\partial a_i} = u_i$$

$$\frac{\partial a_i}{\partial z_i} = \sigma(z_i) \cdot \left(1 - \sigma(z_i)\right)$$
$$= a_i \cdot \left(1 - a_i\right)$$

Backprop (*toy*) example (9)



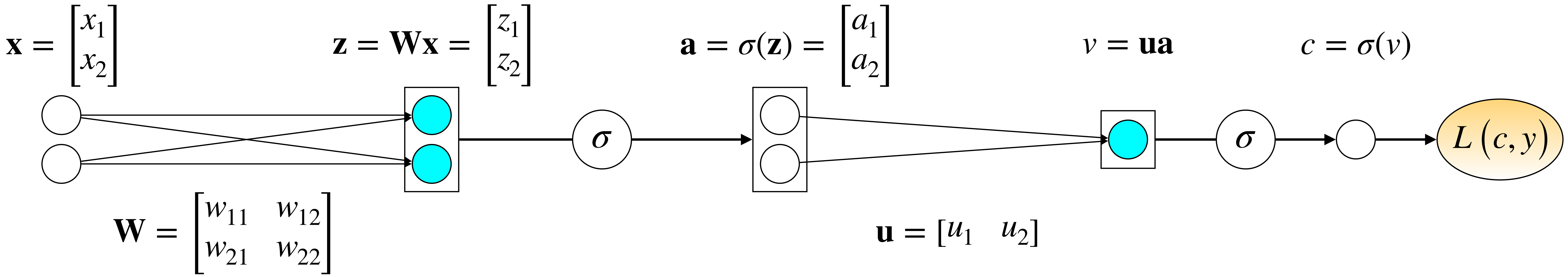
$$\frac{\partial L}{\partial w_{ij}} = \boxed{\frac{\partial L}{\partial c}} \cdot \boxed{\frac{\partial c}{\partial v}} \cdot \boxed{\frac{\partial v}{\partial a_i}} \cdot \boxed{\frac{\partial a_i}{\partial z_i}} \cdot \boxed{\frac{\partial z_i}{\partial w_{ij}}}$$

Given that $z_i = \sum_{j=1}^{2} w_{ij} \cdot x_j$

$$\frac{\partial v}{\partial a_i} = \frac{\partial \left(\sum_{i=1}^{2} u_i \cdot a_i \right)}{\partial a_i} = u_i$$

$$\begin{aligned} \frac{\partial a_i}{\partial z_i} &= \sigma(z_i) \cdot \left(1 - \sigma(z_i)\right) \\ &= a_i \cdot \left(1 - a_i\right) \end{aligned}$$

Backprop (*toy*) example (9)



$$\frac{\partial L}{\partial w_{ij}} = \boxed{\frac{\partial L}{\partial c}} \cdot \boxed{\frac{\partial c}{\partial v}} \cdot \boxed{\frac{\partial v}{\partial a_i}} \cdot \boxed{\frac{\partial a_i}{\partial z_i}} \cdot \boxed{\frac{\partial z_i}{\partial w_{ij}}}$$

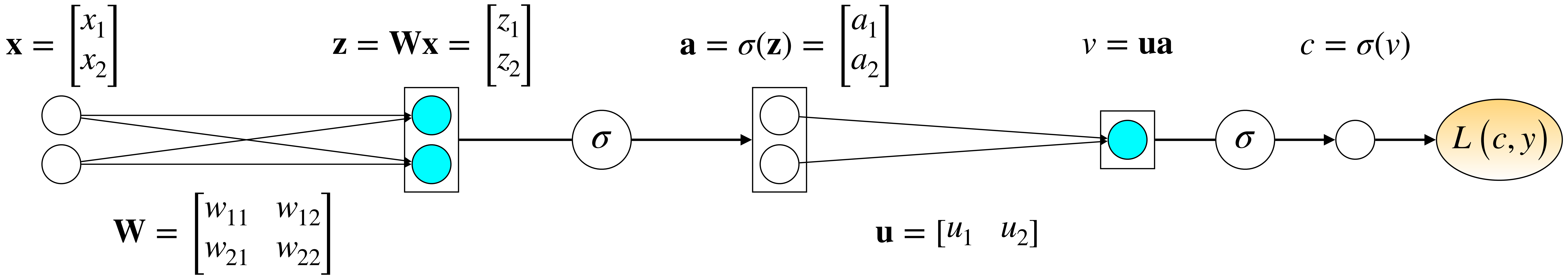
Given that $z_i = \sum_{j=1}^{2} w_{ij} \cdot x_j$

$$\frac{\partial v}{\partial a_i} = \frac{\partial \left(\sum_{i=1}^{2} u_i \cdot a_i \right)}{\partial a_i} = u_i$$

$$\frac{\partial a_i}{\partial z_i} = \sigma(z_i) \cdot \left(1 - \sigma(z_i)\right)$$
$$= a_i \cdot \left(1 - a_i\right)$$

$$\frac{\partial z_i}{\partial w_{ij}} = x_j$$

Backprop (*toy*) example (9)



$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial v} \cdot \frac{\partial v}{\partial a_i} \cdot \frac{\partial a_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_{ij}} \boxed{= (c - y) \cdot u_i \cdot a_i \cdot \left(1 - a_i\right) \cdot x_j}$$

Given that $z_i = \sum_{j=1}^{2} w_{ij} \cdot x_j$

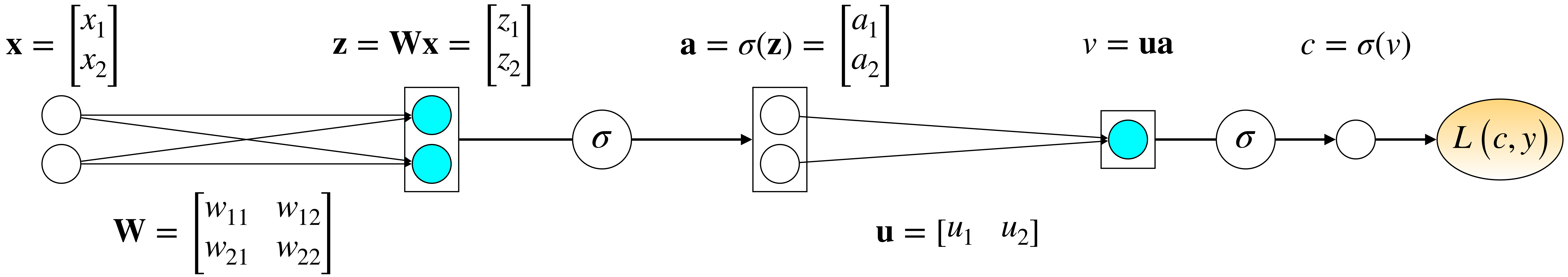
$$\frac{\partial v}{\partial a_i} = \frac{\partial \left(\sum_{i=1}^{2} u_i \cdot a_i \right)}{\partial a_i} = u_i$$

$$\frac{\partial a_i}{\partial z_i} = \sigma(z_i) \cdot \left(1 - \sigma(z_i)\right)$$

$$= a_i \cdot \left(1 - a_i\right)$$

$$\frac{\partial z_i}{\partial w_{ij}} = x_j$$

Updating the parameters of the NN



$$u_i^{\text{new}} = u_i^{\text{old}} - \eta \frac{\partial L}{\partial u_i}$$

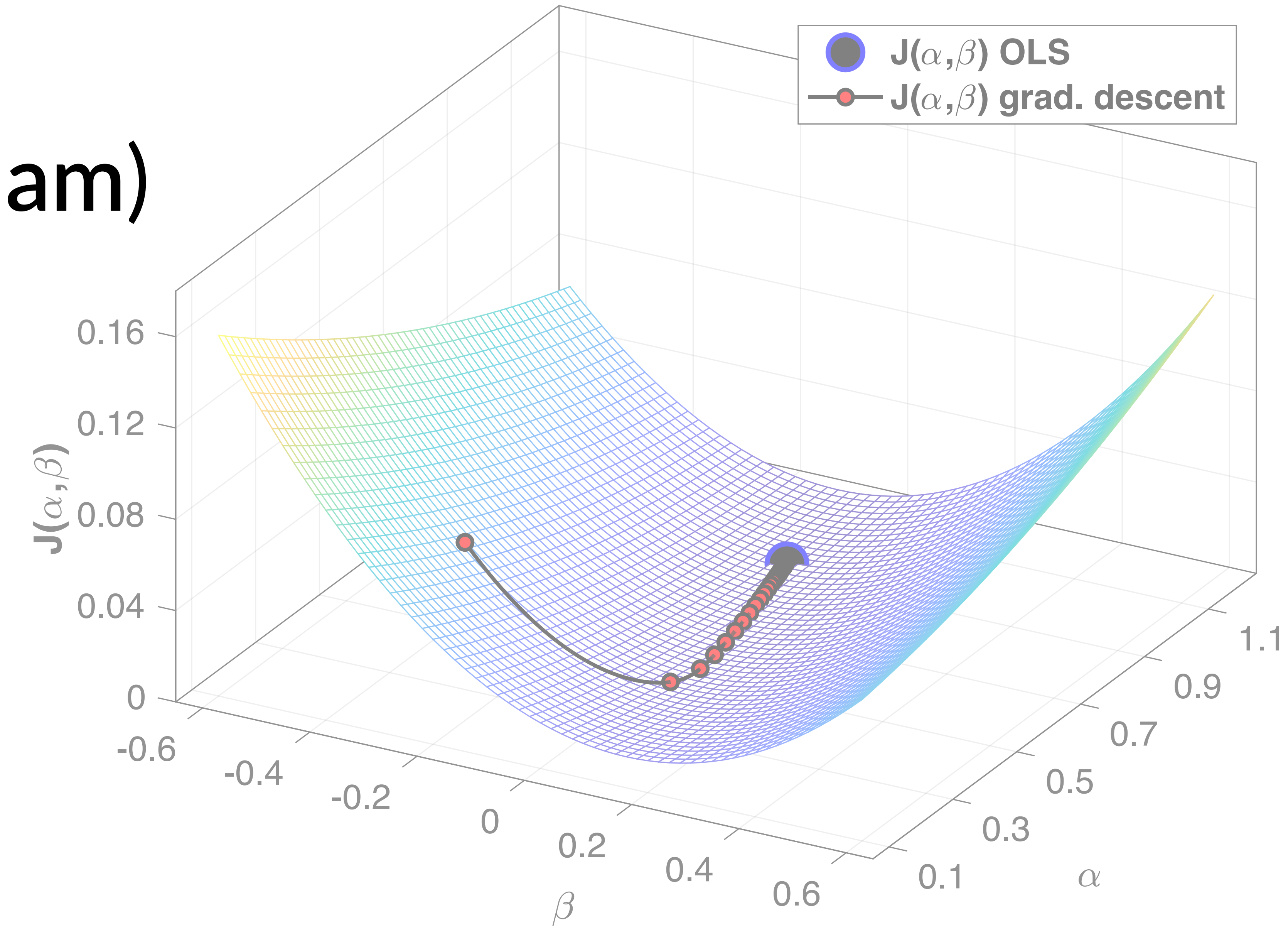
using a learning rate η

$$w_{ij}^{\text{new}} = w_{ij}^{\text{old}} - \eta \frac{\partial L}{\partial w_{ij}}$$

Optimisation (training)

- Stochastic gradient descent (SGD) *works* most of the time with some effort
  - ➡ if we know our data / task well and can handle the learning rate (η)

- Adaptive (more sophisticated) optimisers perform generally better; *keep track how much gradients change and dynamically decide how much to update the weights*
  - ➡ RMSProp
  - ➡ Adaptive Moment Estimation Method (**Adam**)
  - ➡ Adagrad
  - ➡ AdaDelta
  - ➡ SparseAdam
  - ➡ many other variants



Learning rate (η)

- We want the learning rate to be just right (not too large or small)
- Too large $\implies$ learning too fast: the model may diverge and not converge
- Too small $\implies$ learning too slow: the model will not diverge, but may take ages to converge
- ≈ 0.001 is a common starting point value for a learning rate
  tune it by orders of magnitude e.g. $\left[0.01, 0.001, 0.0001\right]$
- In SGD, you might want to decrease the learning rate as the training epochs increase
- In fancier optimisers (e.g. Adam) we set the initial learning rate, but then the optimiser takes care of dynamically tuning it

Next lecture

- Friday, January 26
- Word embeddings (`word2vec` mainly)