Neural Networks
and
Computation Graphs

THE UNIVERSITY OF UTAH

Based on slides and material from Geoffrey Hinton, Richard Socher, Yoav Goldberg, Chris Dyer, Graham Neubig and others.

Where are we?

- What is a neural network?
- Computation Graphs
- Algorithms over computation graphs
  - The forward pass
  - The backward pass

Three computational questions

1. Forward propagation
   - Given inputs to the graph, compute the value of the function expressed by the graph
   - Something to think about: Given a node, can we say which nodes are inputs? Which nodes are outputs?

2. Backpropagation
   - After computing the function value for an input, compute the gradient of the function at that input
   - Or equivalently: *How does the output change if I make a small change to the input?*

3. Constructing graphs
   - Need an easy-to-use framework to construct graphs
   - The size of the graph may be input dependent
     - A templating language that creates graphs on the fly
   - Tensorflow, PyTorch are the most popular frameworks today

Backpropagation with computation graphs

Three computational questions

1. Forward propagation
   - Given inputs to the graph, compute the value of the function expressed by the graph
   - Something to think about: Given a node, can we say which nodes are inputs? Which nodes are outputs?

2. Backpropagation
   - After computing the function value for an input, compute the gradient of the function at that input
   - Or equivalently: *How does the output change if I make a small change to the input?*

3. Constructing graphs
   - Need an easy-to-use framework to construct graphs
   - The size of the graph may be input dependent
     - A templating language that creates graphs on the fly
   - Tensorflow, PyTorch are the most popular frameworks today

Calculus refresher: The chain rule

Suppose we have two functions f and g

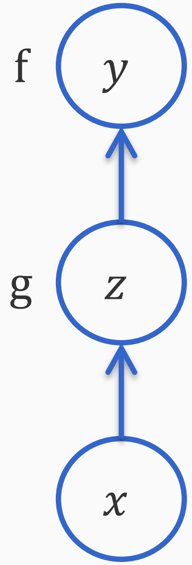
We wish to compute the gradient of $\mathrm{y} = \ f\big(g(x)\big)$.

We know that $\frac{dy}{dx} = f'\big(g(x)\big) \cdot g'(x)$

Or equivalently: if $z = g(x)$ and $y = f(z)$, then

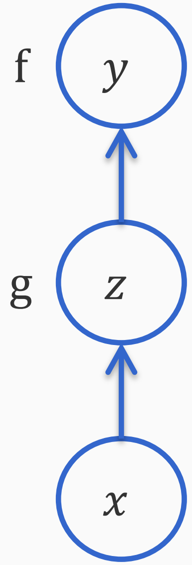
$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \frac{dz}{dx}$$

Or equivalently: In terms of computation graphs



The forward pass gives us z and y

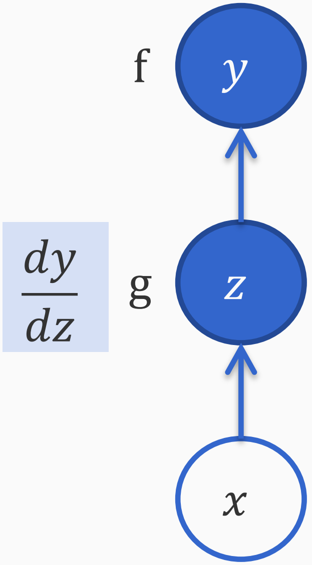
Or equivalently: In terms of computation graphs



The forward pass gives us z and y

Remember that each node knows not only how to compute its value given inputs, but also how to compute gradients

Or equivalently: In terms of computation graphs

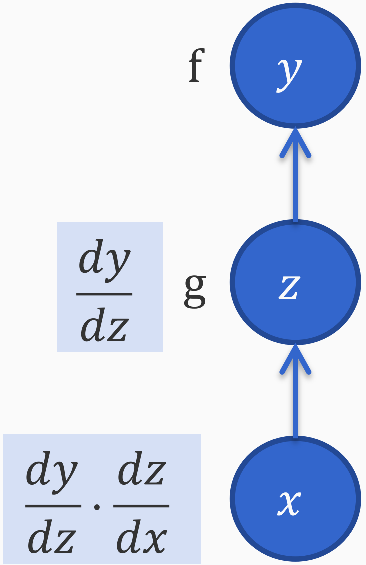


The forward pass gives us z and y

Remember that each node knows not only how to compute its value given inputs, but also how to compute gradients

Start from the root of the graph and work backwards.

Or equivalently: In terms of computation graphs



The forward pass gives us z and y

Remember that each node knows not only how to compute its value given inputs, but also how to compute gradients

Start from the root of the graph and work backwards.

When traversing an edge backwards to a new node:

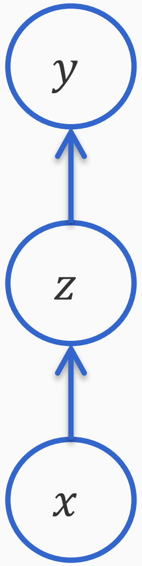
> the gradient of the root with respect to that node is the product of the gradient at the parent with the derivative along that edge

A concrete example

$$y = \frac{1}{x^2}$$

$$f(u) = \frac{1}{u}$$

$$\mathrm{g(u) = u^2}$$

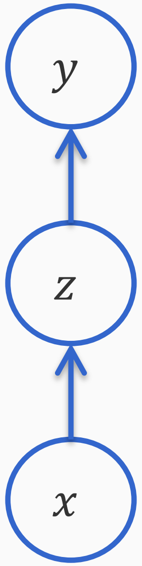


A concrete example

$$y = \frac{1}{x^2}$$

$$\frac{df}{du} = -\frac{1}{u^2} \qquad f(u) = \frac{1}{u}$$

$$\frac{dg}{du} = 2u \qquad \mathrm{g(u)} = \mathrm{u}^2$$



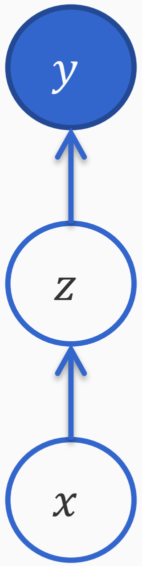
Let’s also explicitly write down the derivatives.

A concrete example

$$y = \frac{1}{x^2}$$

$$\frac{df}{du} = -\frac{1}{u^2} \qquad f(u) = \frac{1}{u}$$

$$\frac{dg}{du} = 2u \qquad \mathrm{g(u) = u^2}$$



$$\frac{dy}{dy} = 1$$

Now, we can proceed backwards from the output

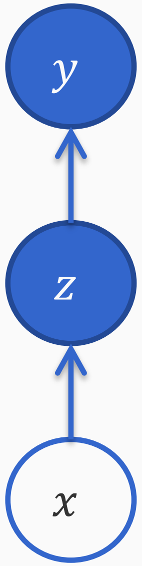
At each step, we compute the gradient of the function represented by the graph with respect to the node that we are at.

A concrete example

$$y = \frac{1}{x^2}$$

$$\frac{df}{du} = -\frac{1}{u^2} \qquad f(u) = \frac{1}{u}$$

$$\frac{dg}{du} = 2u \qquad \mathrm{g(u)} = \mathrm{u}^2$$



$$\frac{dy}{dy} = 1$$

$$\frac{dy}{dz} = \frac{dy}{dy} \cdot \left(\frac{df}{du}\right)_{u=z} = 1 \cdot \left(-\frac{1}{z^2}\right) = -\frac{1}{z^2}$$

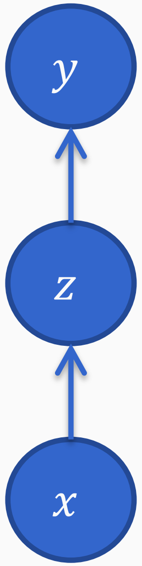
Product of the gradient so far and
the derivative computed at this step

A concrete example

$$y = \frac{1}{x^2}$$

$$\frac{df}{du} = -\frac{1}{u^2} \qquad f(u) = \frac{1}{u}$$

$$\frac{dg}{du} = 2u \qquad \mathrm{g(u) = u^2}$$



$$\frac{dy}{dy} = 1$$

$$\frac{dy}{dz} = -\frac{1}{z^2}$$

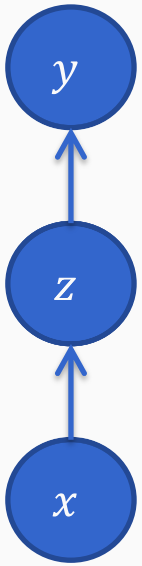
$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \left(\frac{dg}{du}\right)_{u=x} = -\frac{1}{z^2} \cdot 2x = -\frac{2\mathrm{x}}{\mathrm{z}^2}$$

A concrete example

$$y = \frac{1}{x^2}$$

$$\frac{df}{du} = -\frac{1}{u^2} \qquad f(u) = \frac{1}{u}$$

$$\frac{dg}{du} = 2u \qquad \mathrm{g(u) = u^2}$$



$$\frac{dy}{dy} = 1$$

$$\frac{dy}{dz} = -\frac{1}{z^2}$$

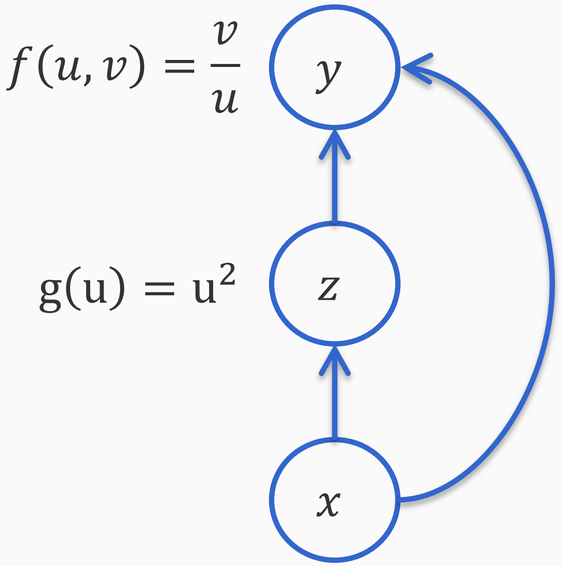
$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \left(\frac{dg}{du}\right)_{u=x} = -\frac{1}{z^2} \cdot 2x = -\frac{2\mathrm{x}}{\mathrm{z}^2}$$

We can simplify this to get $-\frac{2}{x^3}$

A concrete example

with multiple outgoing edges

$$y = \frac{1}{x}$$



A concrete example

with multiple outgoing edges

$$y = \frac{1}{x}$$

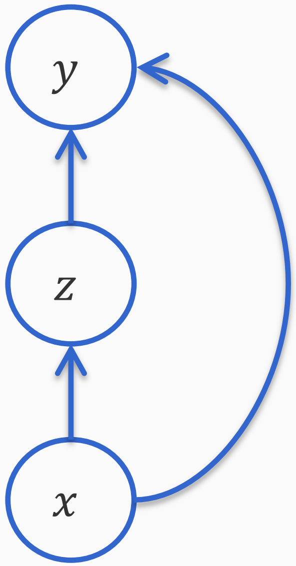
$$\frac{df}{du} = -\frac{v}{u^2}$$

$$\frac{df}{dv} = \frac{1}{u}$$

$$\frac{dg}{du} = 2u$$

$$f(u, v) = \frac{v}{u}$$

$$\mathrm{g(u) = u^2}$$



Let's also explicitly write down the derivatives. Note that f has two derivatives because it has two inputs.

A concrete example

with multiple outgoing edges

$$y = \frac{1}{x}$$

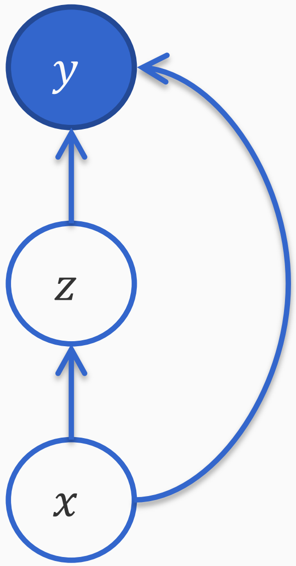
$$\frac{df}{du} = -\frac{v}{u^2}$$

$$\frac{df}{dv} = \frac{1}{u}$$

$$f(u, v) = \frac{v}{u}$$

$$\mathrm{g(u) = u^2}$$

$$\frac{dg}{du} = 2u$$



$$\frac{dy}{dy} = 1$$

A concrete example

with multiple outgoing edges

$$y = \frac{1}{x}$$

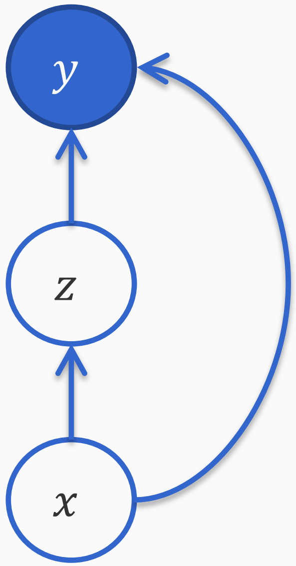
$$\frac{df}{du} = -\frac{v}{u^2}$$

$$\frac{df}{dv} = \frac{1}{u}$$

$$\frac{dg}{du} = 2u$$

$$f(u, v) = \frac{v}{u}$$

$$\mathrm{g(u)} = \mathrm{u}^2$$



$$\frac{dy}{dy} = 1$$

At this point, we can compute the gradient of y with respect to z by following the edge from y to z.

But we can not follow the edge from y to x because all of x's descendants are not marked as done.

A concrete example

with multiple outgoing edges

$$y = \frac{1}{x}$$

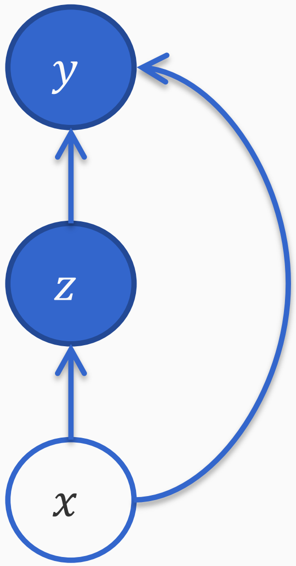
$$\frac{df}{du} = -\frac{v}{u^2}$$

$$\frac{df}{dv} = \frac{1}{u}$$

$$f(u, v) = \frac{v}{u}$$

$$\frac{dg}{du} = 2u$$

$$\mathrm{g(u) = u^2}$$



$$\frac{dy}{dy} = 1$$

$$\frac{dy}{dz} = \frac{dy}{dy} \cdot \left(\frac{df}{du}\right)_{u=z} = 1 \cdot \left(-\frac{x}{z^2}\right) = -\frac{x}{z^2}$$

Product of the gradient so far and
the derivative computed at this step

A concrete example

with multiple outgoing edges

$$y = \frac{1}{x}$$

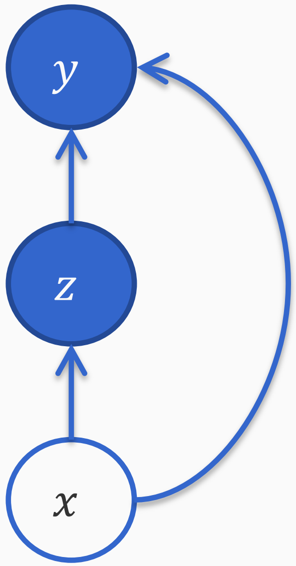
$$\frac{df}{du} = -\frac{v}{u^2}$$

$$\frac{df}{dv} = \frac{1}{u}$$

$$\frac{dg}{du} = 2u$$

$$f(u, v) = \frac{v}{u}$$

$$\mathrm{g(u) = u^2}$$



$$\frac{dy}{dy} = 1$$

$$\frac{dy}{dz} = \frac{dy}{dy} \cdot \left(\frac{df}{du}\right)_{u=z} = 1 \cdot \left(-\frac{x}{z^2}\right) = -\frac{x}{z^2}$$

Now we can get to x

There are multiple backward paths into x.
The general rule: Add the gradients along all the paths.

A concrete example

with multiple outgoing edges

$$y = \frac{1}{x}$$

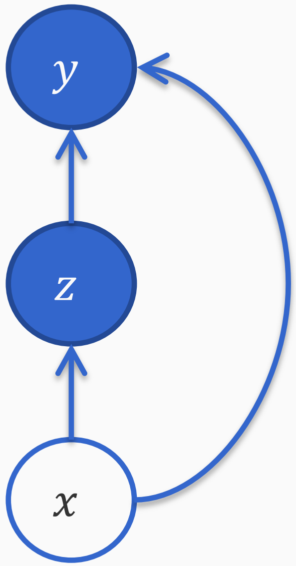
$$\frac{df}{du} = -\frac{v}{u^2}$$

$$\frac{df}{dv} = \frac{1}{u}$$

$$f(u, v) = \frac{v}{u}$$

$$\frac{dg}{du} = 2u$$

$$\mathrm{g(u) = u^2}$$



$$\frac{dy}{dy} = 1$$

$$\frac{dy}{dz} = -\frac{x}{z^2}$$

$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \left(\frac{dg}{du}\right)_{u=x} + \frac{dy}{dy} \cdot \left(\frac{df}{dv}\right)_{v=x}$$

There are multiple backward paths into x.
The general rule: Add the gradients along all the paths.

A concrete example

with multiple outgoing edges

$$y = \frac{1}{x}$$

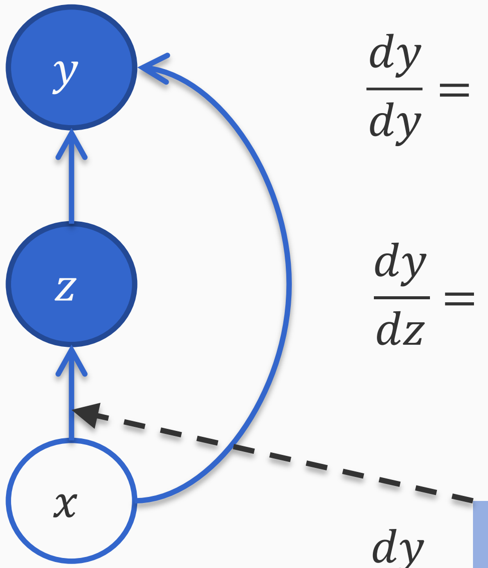
$$\frac{df}{du} = -\frac{v}{u^2}$$

$$\frac{df}{dv} = \frac{1}{u}$$

$$f(u,v) = \frac{v}{u}$$

$$\frac{dg}{du} = 2u$$

$$\mathrm{g(u)} = \mathrm{u}^2$$



$$\frac{dy}{dy} = 1$$

$$\frac{dy}{dz} = -\frac{x}{z^2}$$

$$\frac{dy}{dx} = \boxed{\frac{dy}{dz} \cdot \left(\frac{dg}{du}\right)_{u=x}} + \frac{dy}{dy} \cdot \left(\frac{df}{dv}\right)_{v=x}$$

There are multiple backward paths into x.
The general rule: Add the gradients along all the paths.

A concrete example

with multiple outgoing edges

$$y = \frac{1}{x}$$

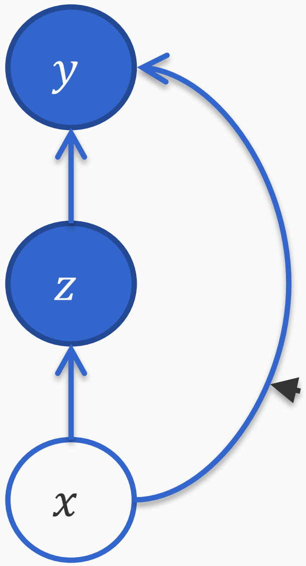
$$\frac{df}{du} = -\frac{v}{u^2}$$

$$\frac{df}{dv} = \frac{1}{u}$$

$$f(u, v) = \frac{v}{u}$$

$$\frac{dg}{du} = 2u$$

$$\mathrm{g(u)} = \mathrm{u}^2$$



$$\frac{dy}{dy} = 1$$

$$\frac{dy}{dz} = -\frac{x}{z^2}$$

$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \left(\frac{dg}{du}\right)_{u=x} + \frac{dy}{dy} \cdot \left(\frac{df}{dv}\right)_{v=x}$$

There are multiple backward paths into x.
The general rule: Add the gradients along all the paths.

A concrete example

with multiple outgoing edges

$$y = \frac{1}{x}$$

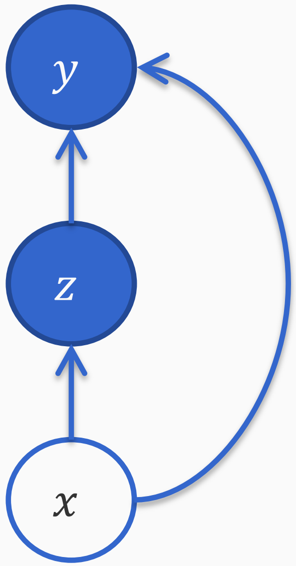
$$\frac{df}{du} = -\frac{v}{u^2}$$

$$\frac{df}{dv} = \frac{1}{u}$$

$$\frac{dg}{du} = 2u$$

$$f(u, v) = \frac{v}{u}$$

$$\mathrm{g(u)} = \mathrm{u}^2$$



$$\frac{dy}{dy} = 1$$

$$\frac{dy}{dz} = -\frac{x}{z^2}$$

$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \left(\frac{dg}{du}\right)_{u=x} + \frac{dy}{dy} \cdot \left(\frac{df}{dv}\right)_{v=x}$$

$$\frac{dy}{dx} = -\frac{x}{z^2} \cdot 2x + 1 \cdot \frac{1}{z} = -\frac{2x^2}{z^2} + \frac{1}{z} = -\frac{1}{x^2}$$

A neural network example

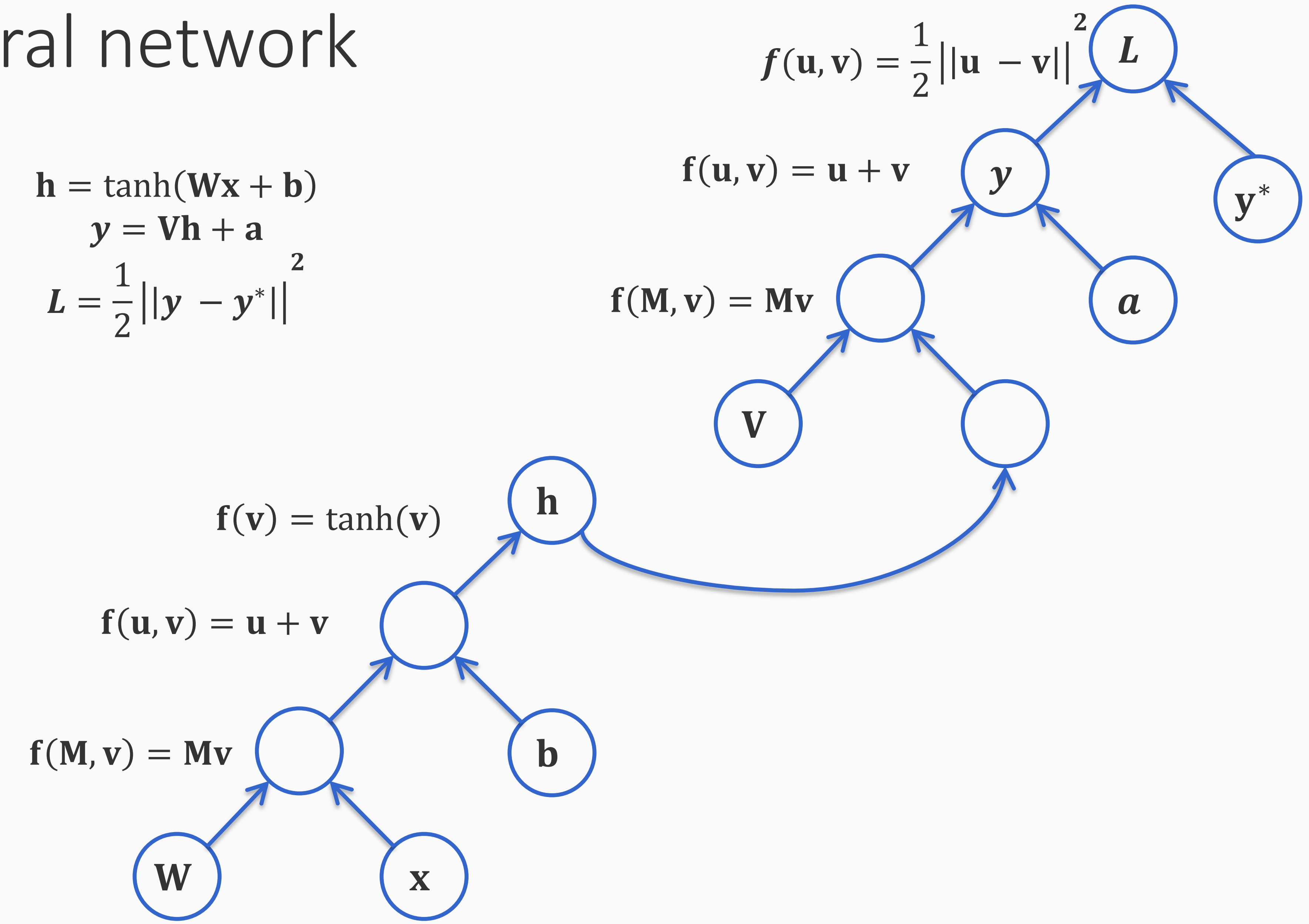
$$\mathbf{h} = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{V}\mathbf{h} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}\left|\left|\boldsymbol{y} - \boldsymbol{y}^*\right|\right|^{\mathbf{2}}$$

This is the same two-layer network we saw before. But this time we have added a new loss term at the end.

Suppose our goal is to compute the derivative of the loss with respect to $\mathbf{W}, \mathbf{V}, \mathbf{a}, \mathbf{b}$

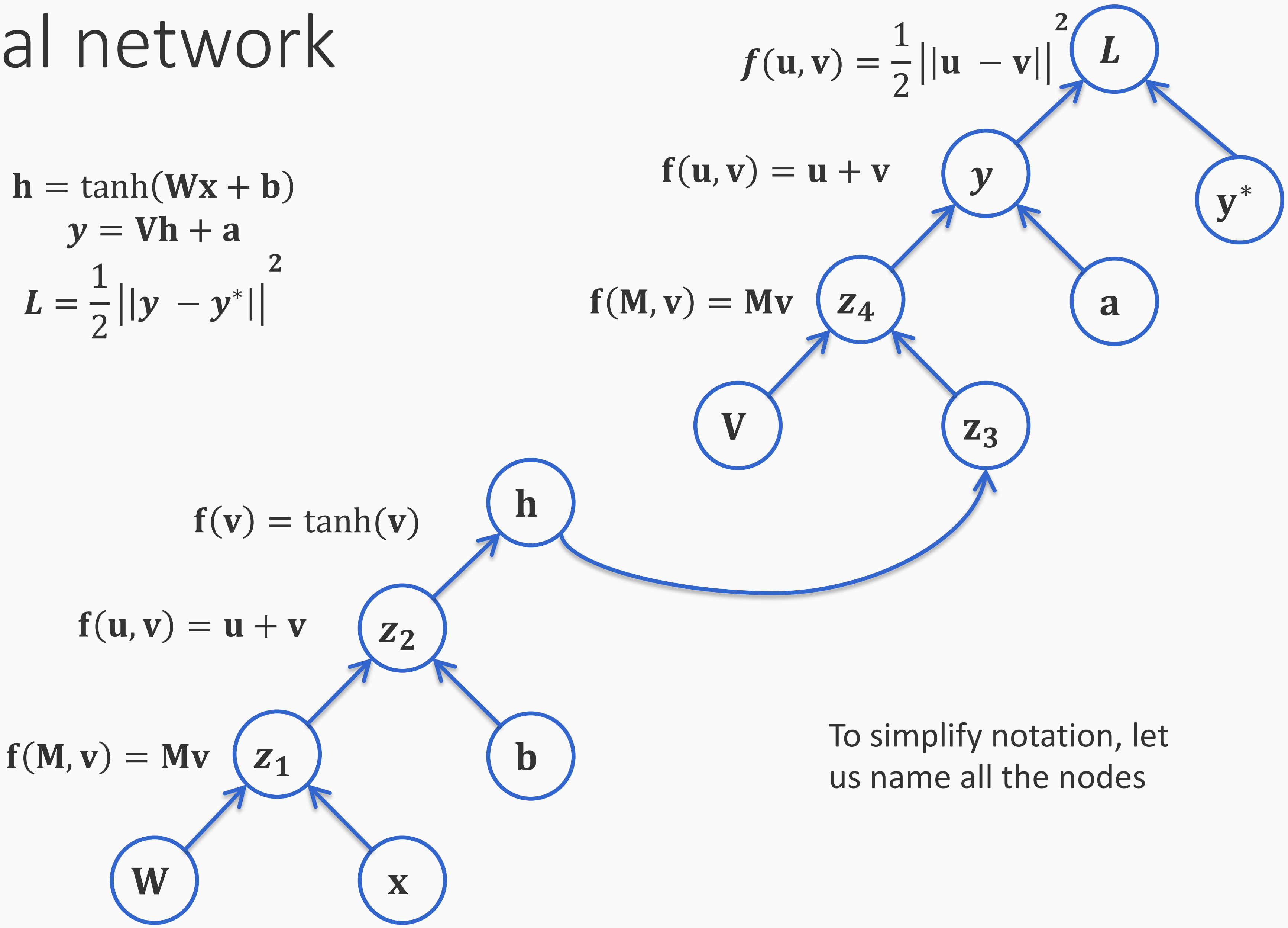
A neural network

$$\mathbf{h} = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{V}\mathbf{h} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{y}^*||^2$$



A neural network

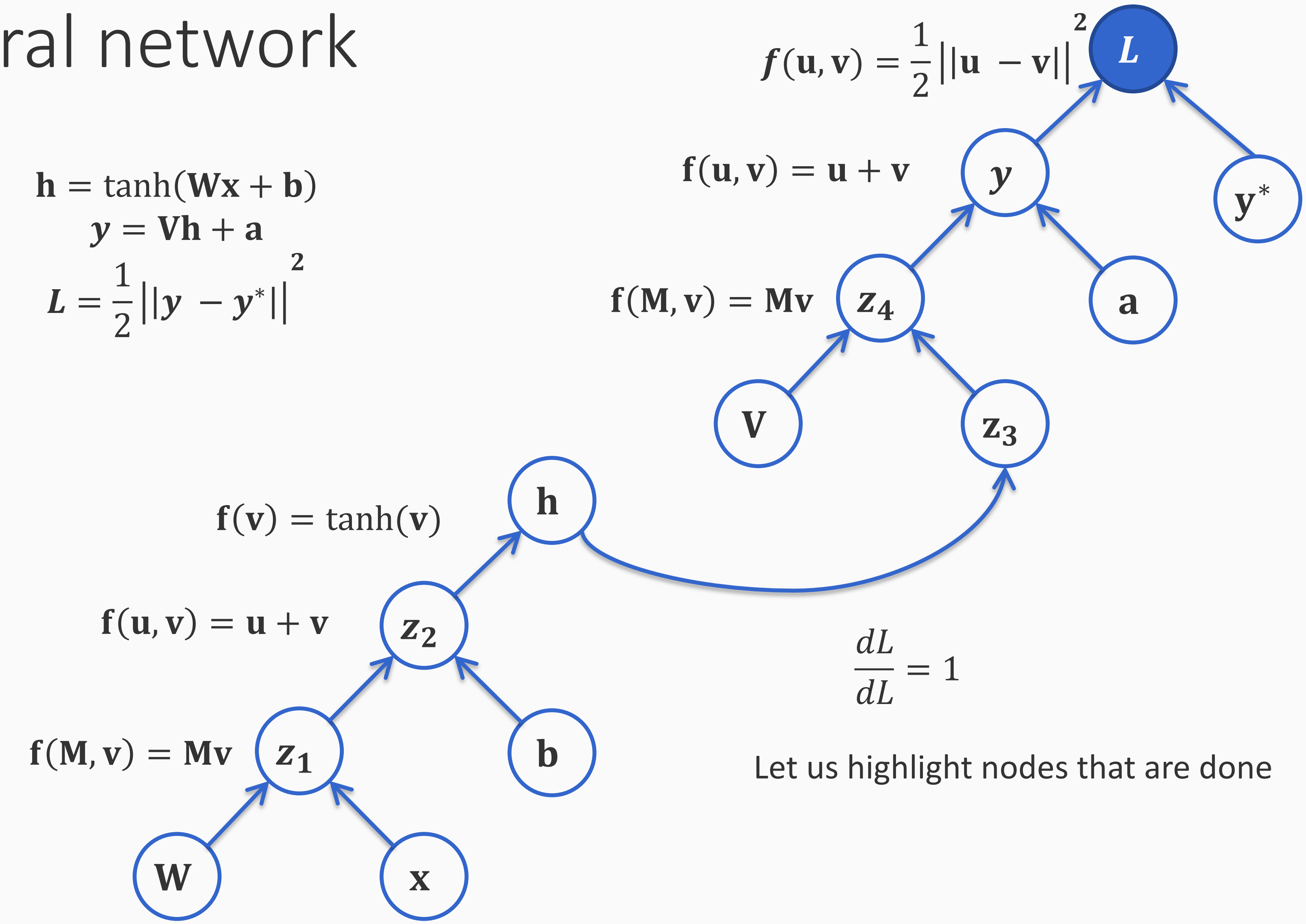
$$\mathbf{h} = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{V}\mathbf{h} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{y}^*\right\|^2$$



To simplify notation, let us name all the nodes

A neural network

$$\mathbf{h} = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{V}\mathbf{h} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{y}^*||^2$$

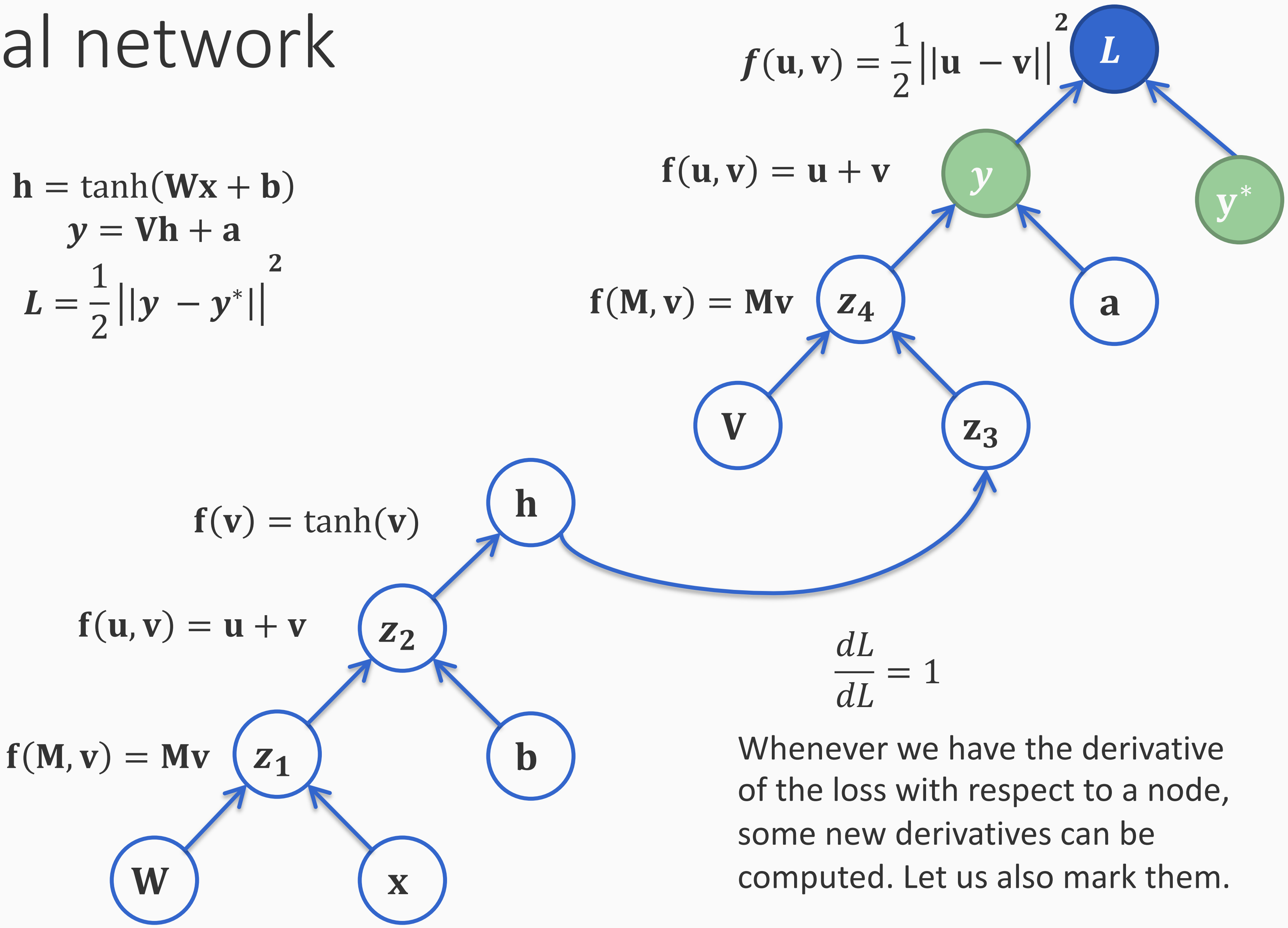


$$\frac{dL}{dL} = 1$$

Let us highlight nodes that are done

A neural network

$$\mathbf{h} = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{V}\mathbf{h} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{y}^*||^2$$

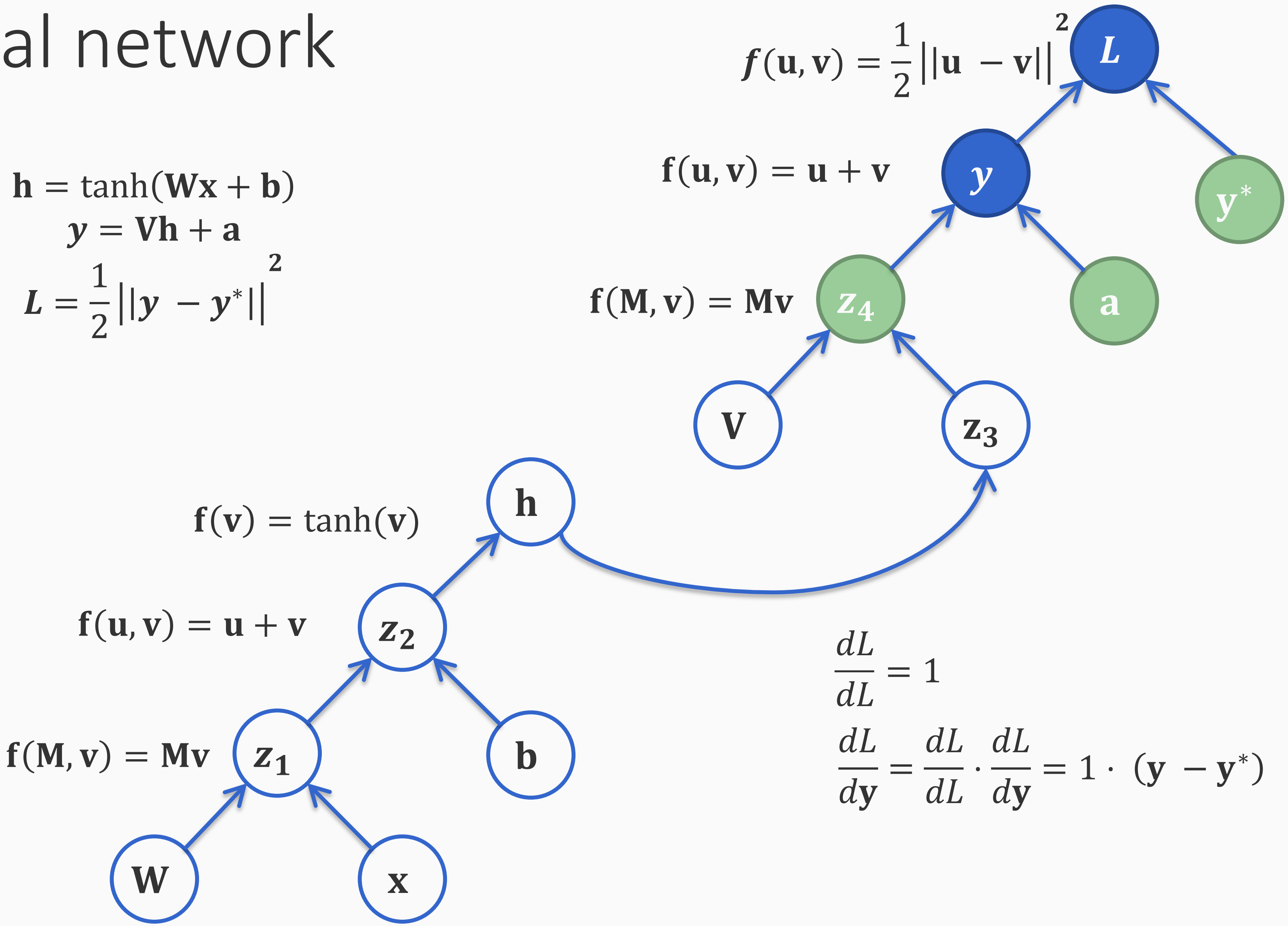


$$\frac{dL}{dL} = 1$$

Whenever we have the derivative of the loss with respect to a node, some new derivatives can be computed. Let us also mark them.

A neural network

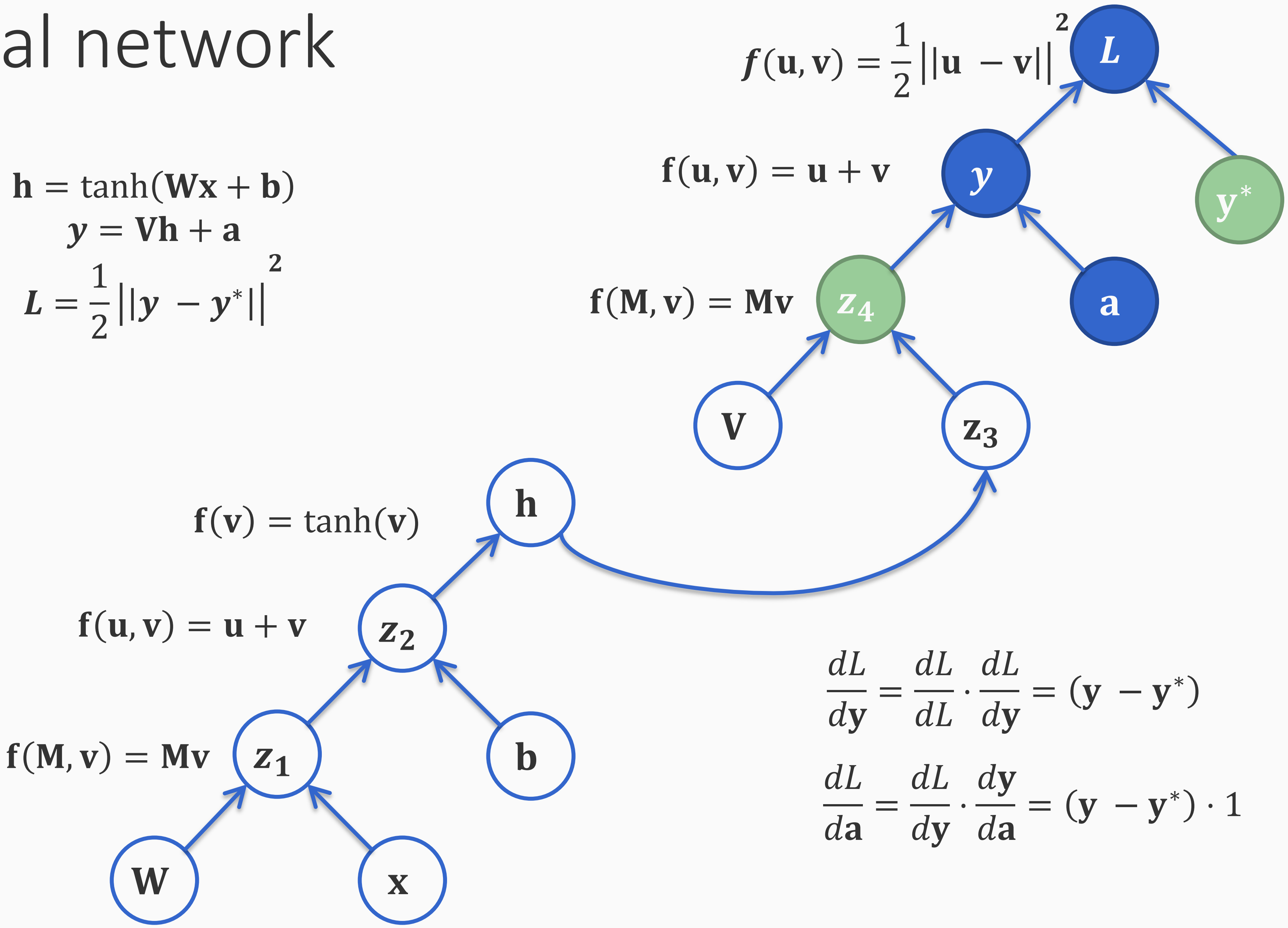
$$\mathbf{h} = \tanh(\mathbf{Wx} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{Vh} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}\left|\left|\boldsymbol{y} - \boldsymbol{y}^*\right|\right|^2$$



$$\frac{dL}{dL} = 1$$
$$\frac{dL}{d\mathbf{y}} = \frac{dL}{dL} \cdot \frac{dL}{d\mathbf{y}} = 1 \cdot (\mathbf{y} - \mathbf{y}^*)$$

A neural network

$$\mathbf{h} = \tanh(\mathbf{Wx} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{Vh} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{y}^*||^2$$

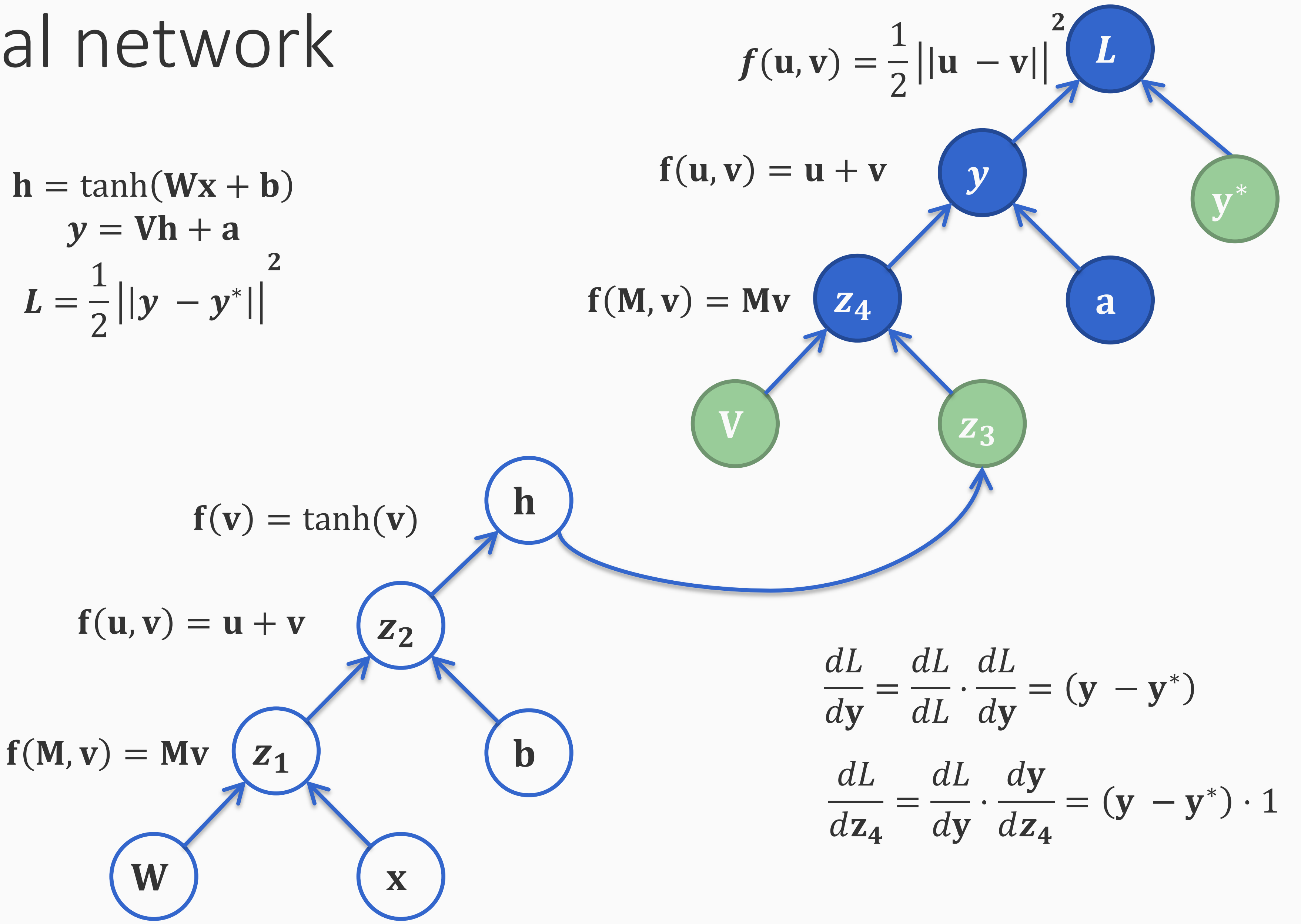


$$\frac{dL}{d\mathbf{y}} = \frac{dL}{dL} \cdot \frac{dL}{d\mathbf{y}} = (\mathbf{y} - \mathbf{y}^*)$$

$$\frac{dL}{d\mathbf{a}} = \frac{dL}{d\mathbf{y}} \cdot \frac{d\mathbf{y}}{d\mathbf{a}} = (\mathbf{y} - \mathbf{y}^*) \cdot 1$$

A neural network

$$\mathbf{h} = \tanh(\mathbf{Wx} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{Vh} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}\left|\left|\boldsymbol{y} - \boldsymbol{y}^*\right|\right|^2$$

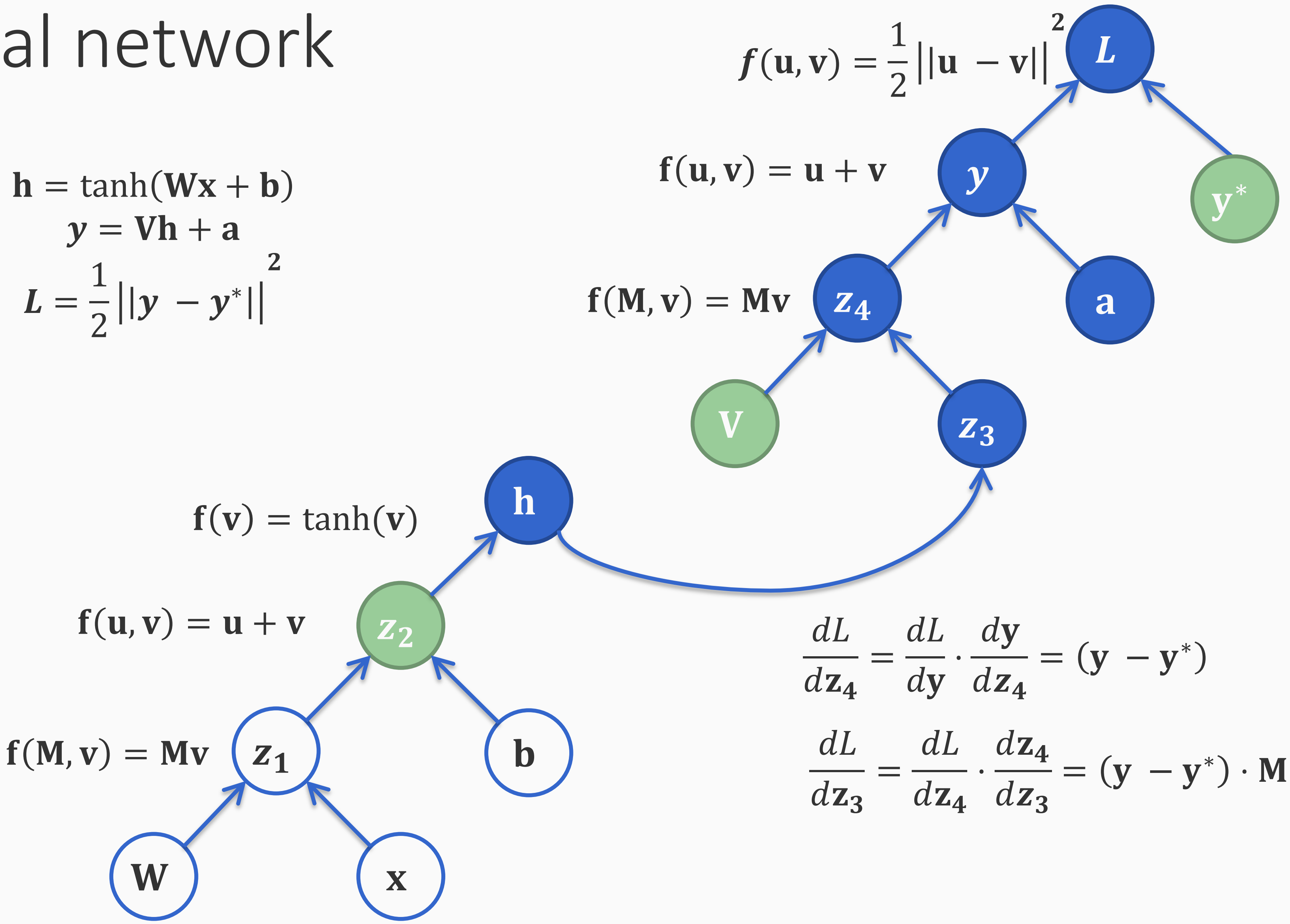


$$\frac{dL}{d\mathbf{y}} = \frac{dL}{dL}\cdot\frac{dL}{d\mathbf{y}} = (\mathbf{y} - \mathbf{y}^*)$$

$$\frac{dL}{d\mathbf{z}_4} = \frac{dL}{d\mathbf{y}}\cdot\frac{d\mathbf{y}}{d\mathbf{z}_4} = (\mathbf{y} - \mathbf{y}^*)\cdot 1$$

A neural network

$$\mathbf{h} = \tanh(\mathbf{Wx} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{Vh} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}\left|\left|\boldsymbol{y} - \boldsymbol{y}^*\right|\right|^2$$

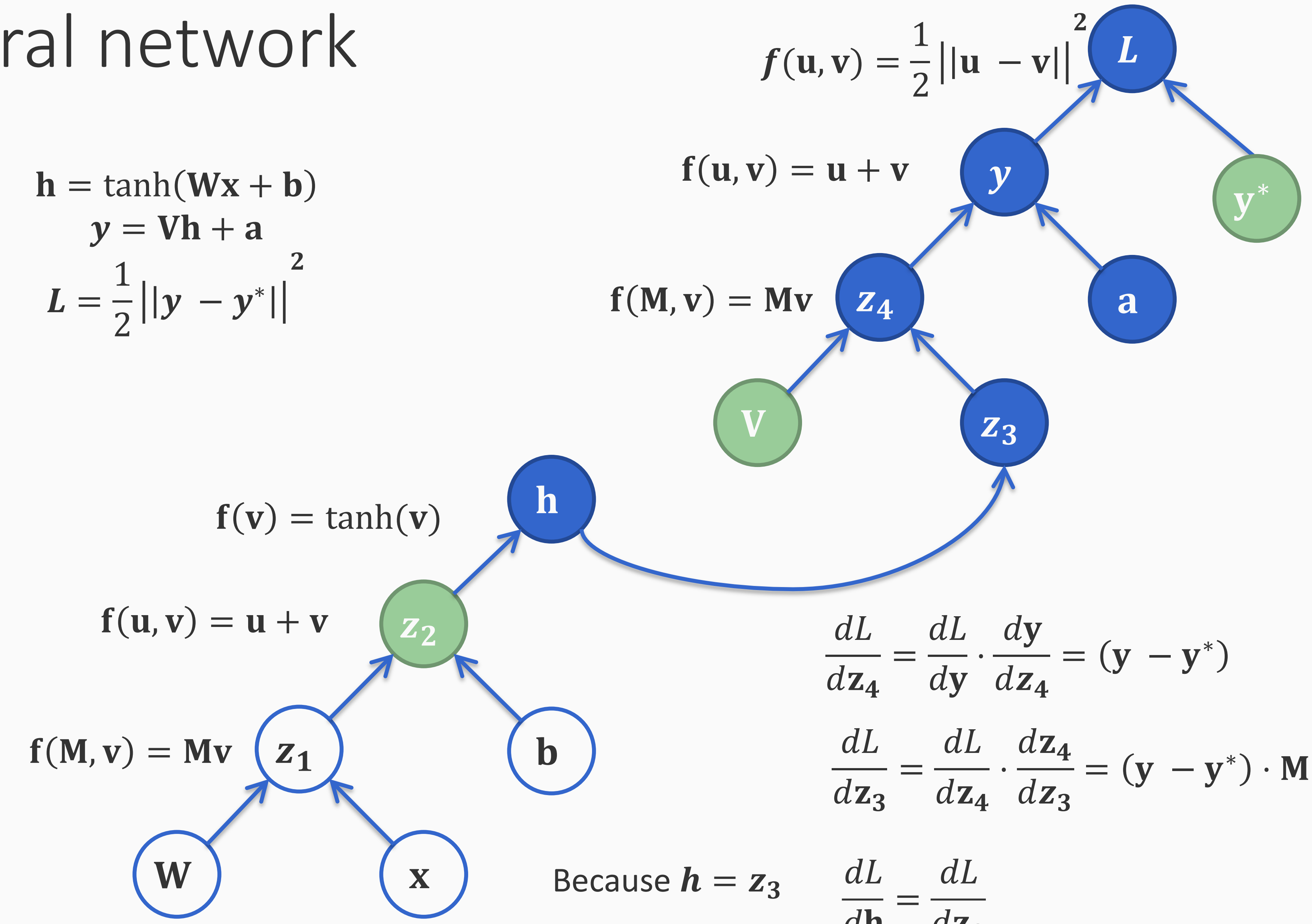


$$\frac{dL}{d\mathbf{z_4}} = \frac{dL}{d\mathbf{y}} \cdot \frac{d\mathbf{y}}{d\mathbf{z_4}} = (\mathbf{y} - \mathbf{y}^*)$$

$$\frac{dL}{d\mathbf{z_3}} = \frac{dL}{d\mathbf{z_4}} \cdot \frac{d\mathbf{z_4}}{d\mathbf{z_3}} = (\mathbf{y} - \mathbf{y}^*) \cdot \mathbf{M}$$

A neural network

$$\mathbf{h} = \tanh(\mathbf{Wx} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{Vh} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}\left|\left|\boldsymbol{y} - \boldsymbol{y}^*\right|\right|^2$$



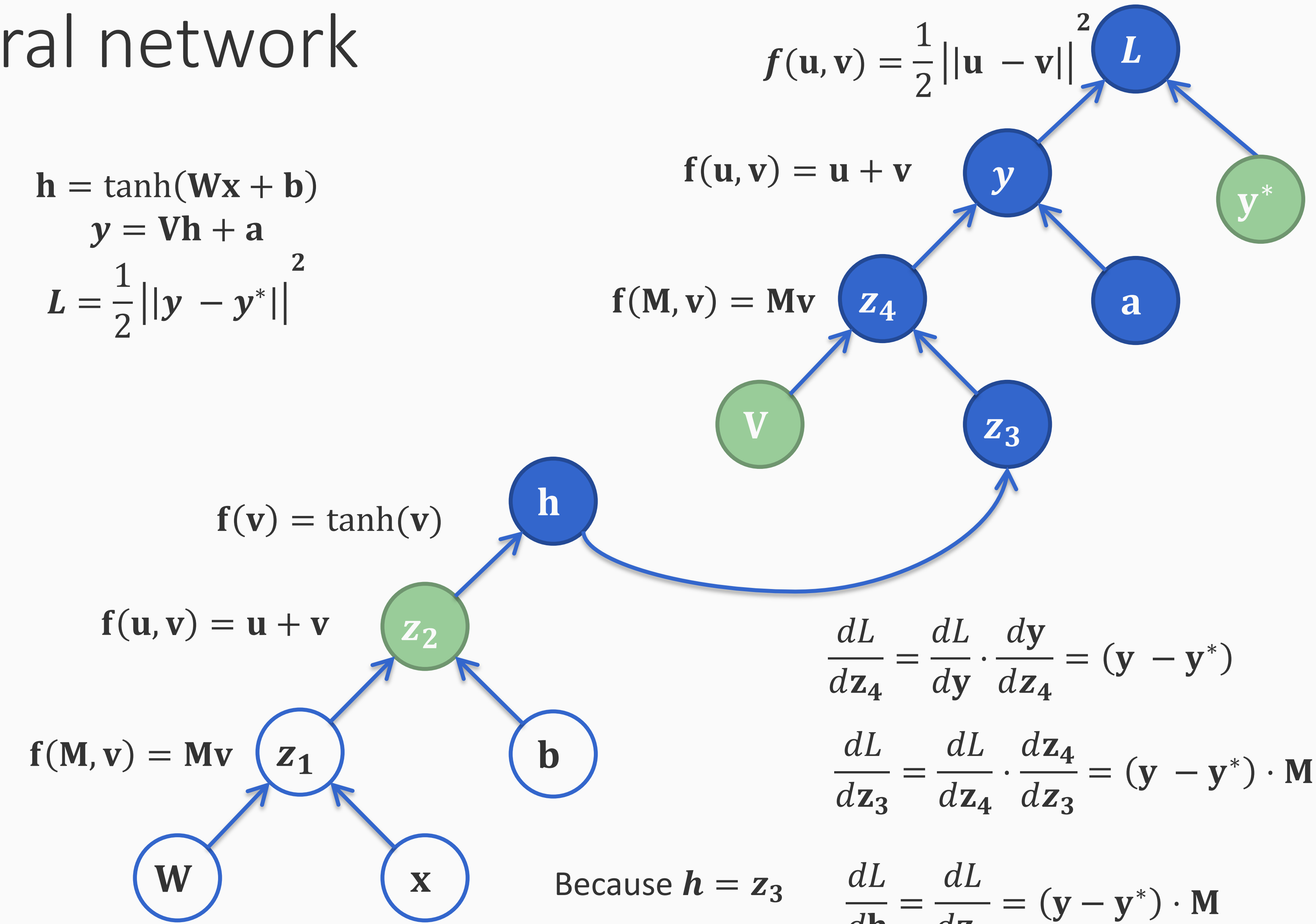
$$\frac{dL}{d\mathbf{z}_4} = \frac{dL}{d\mathbf{y}} \cdot \frac{d\mathbf{y}}{d\mathbf{z}_4} = (\mathbf{y} - \mathbf{y}^*)$$

$$\frac{dL}{d\mathbf{z}_3} = \frac{dL}{d\mathbf{z}_4} \cdot \frac{d\mathbf{z}_4}{d\mathbf{z}_3} = (\mathbf{y} - \mathbf{y}^*) \cdot \mathbf{M}$$

Because $\boldsymbol{h} = \boldsymbol{z}_3$ $\quad \dfrac{dL}{d\mathbf{h}} = \dfrac{dL}{d\mathbf{z}_3}$

A neural network

$$\mathbf{h} = \tanh(\mathbf{Wx} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{Vh} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}\left|\left|\boldsymbol{y} - \boldsymbol{y}^*\right|\right|^2$$



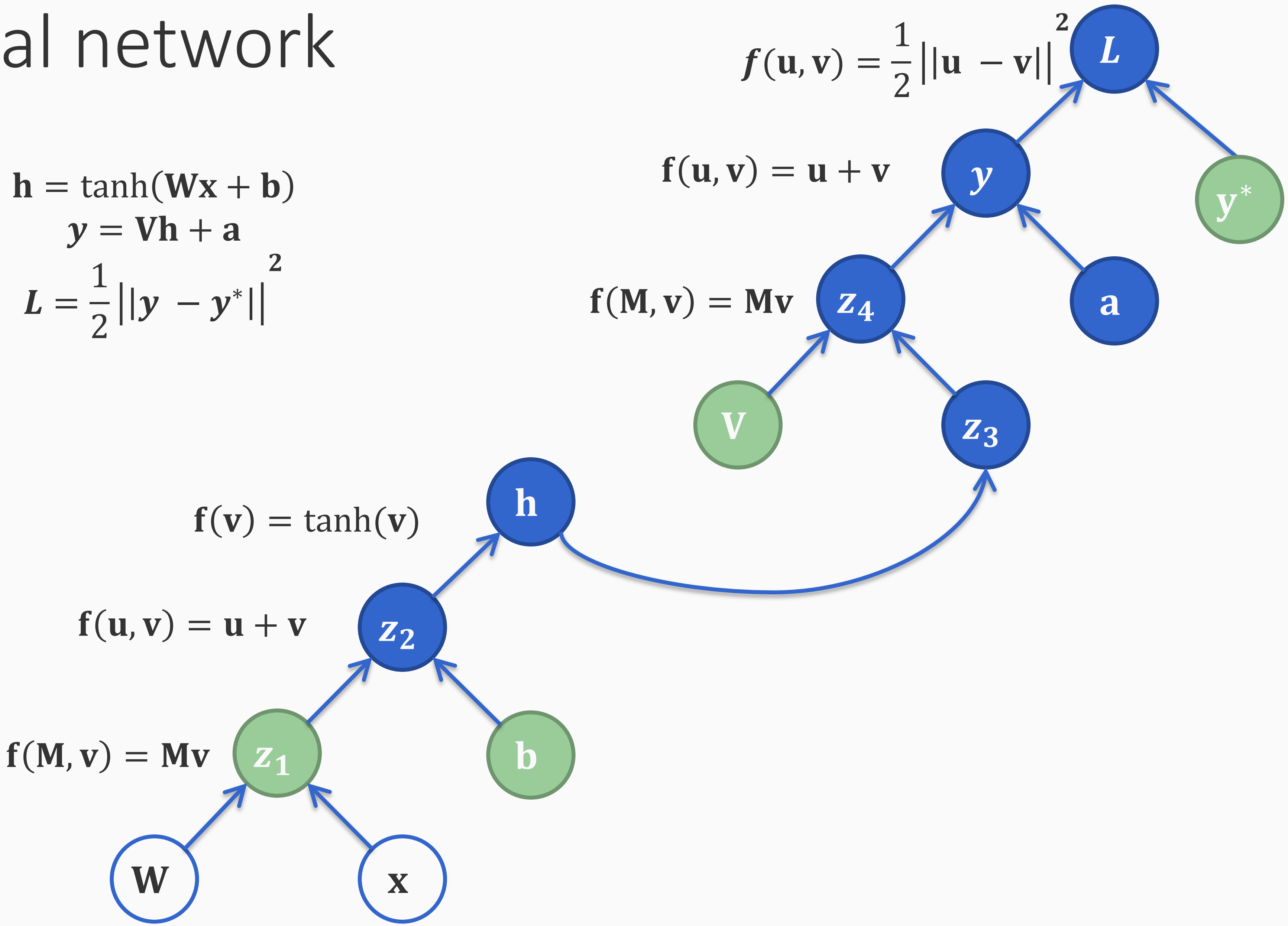
$$\frac{dL}{d\mathbf{z}_4} = \frac{dL}{d\mathbf{y}} \cdot \frac{d\mathbf{y}}{d\mathbf{z}_4} = (\mathbf{y} - \mathbf{y}^*)$$

$$\frac{dL}{d\mathbf{z}_3} = \frac{dL}{d\mathbf{z}_4} \cdot \frac{d\mathbf{z}_4}{d\mathbf{z}_3} = (\mathbf{y} - \mathbf{y}^*) \cdot \mathbf{M}$$

Because $\boldsymbol{h} = \boldsymbol{z}_3$ $$\frac{dL}{d\mathbf{h}} = \frac{dL}{d\mathbf{z}_3} = (\mathbf{y} - \mathbf{y}^*) \cdot \mathbf{M}$$

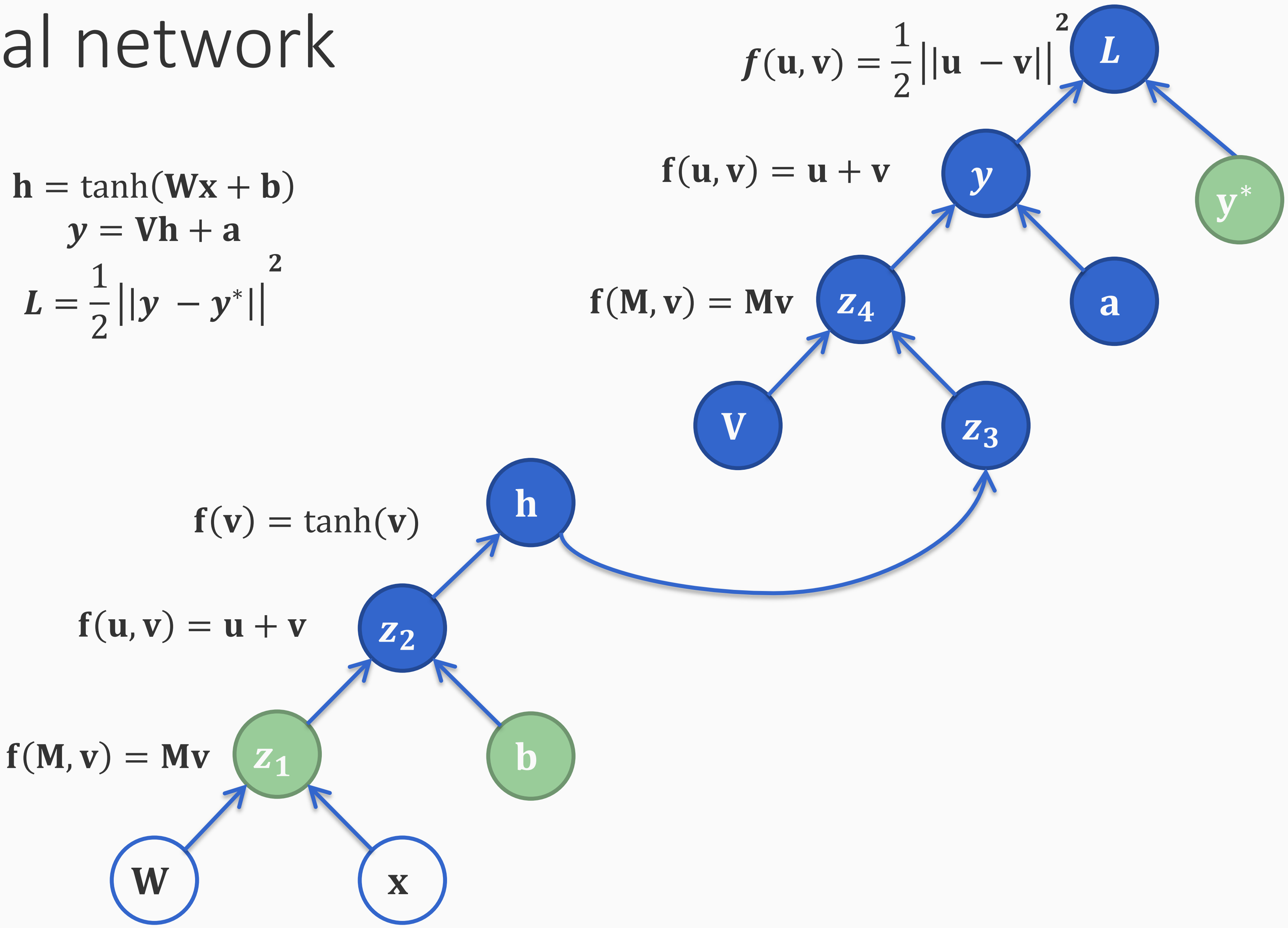
A neural network

$$\mathbf{h} = \tanh(\mathbf{Wx} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{Vh} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}\left|\left|\boldsymbol{y} - \boldsymbol{y}^*\right|\right|^2$$



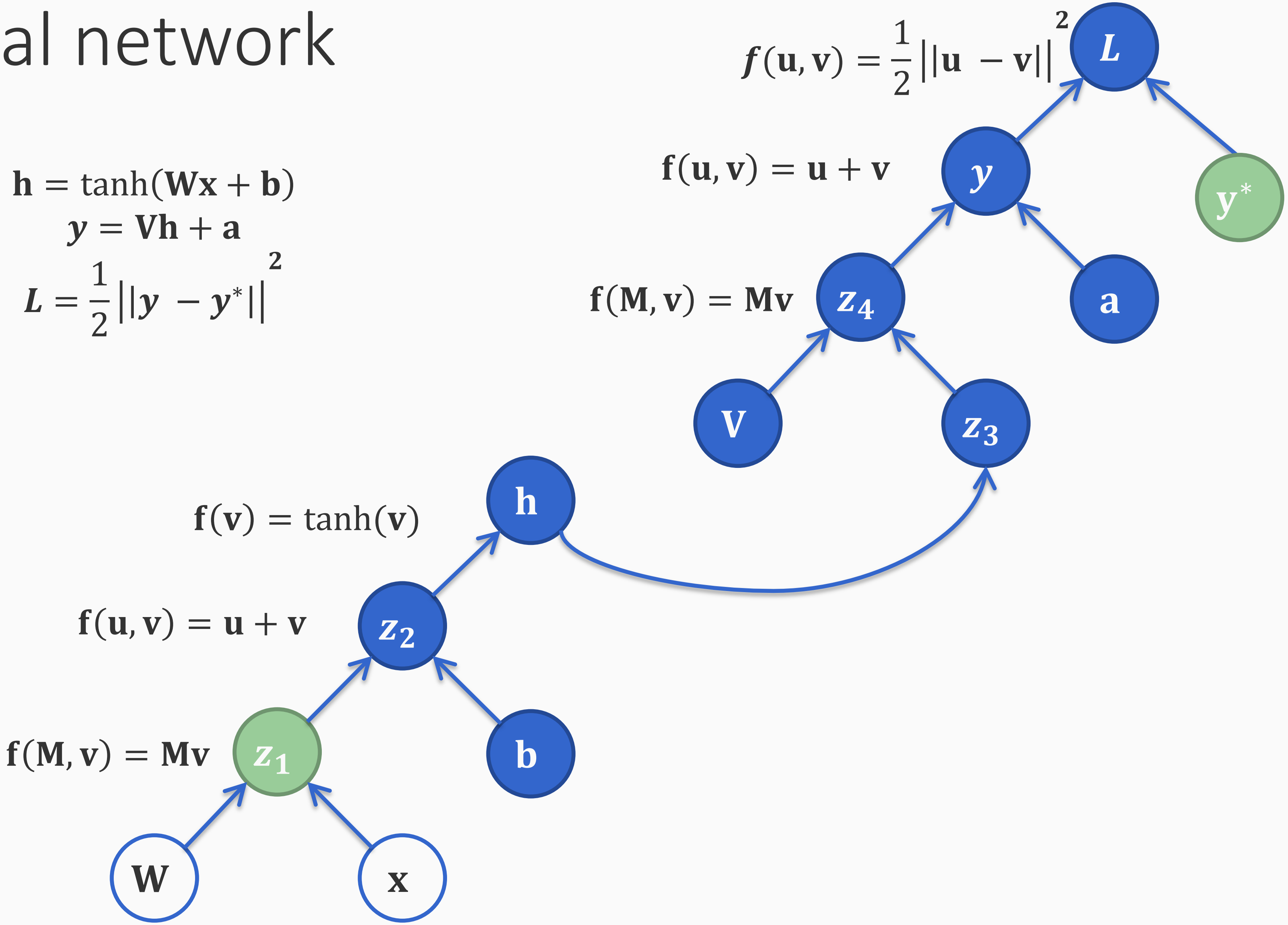
A neural network

$$\mathbf{h} = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{V}\mathbf{h} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}\left|\left|\boldsymbol{y} - \boldsymbol{y}^{*}\right|\right|^{2}$$



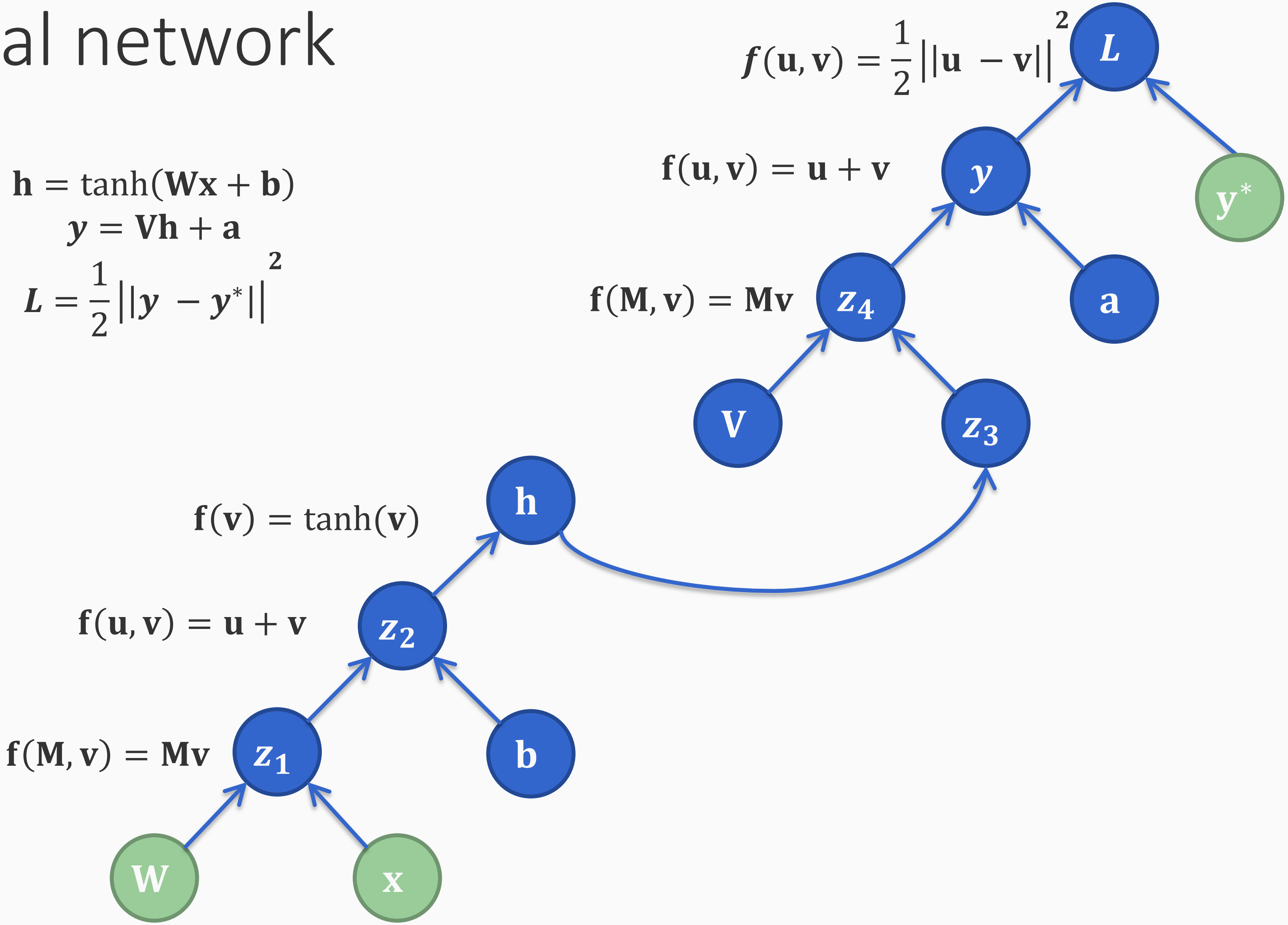
A neural network

$$\mathbf{h} = \tanh(\mathbf{Wx} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{Vh} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{y}^*||^2$$



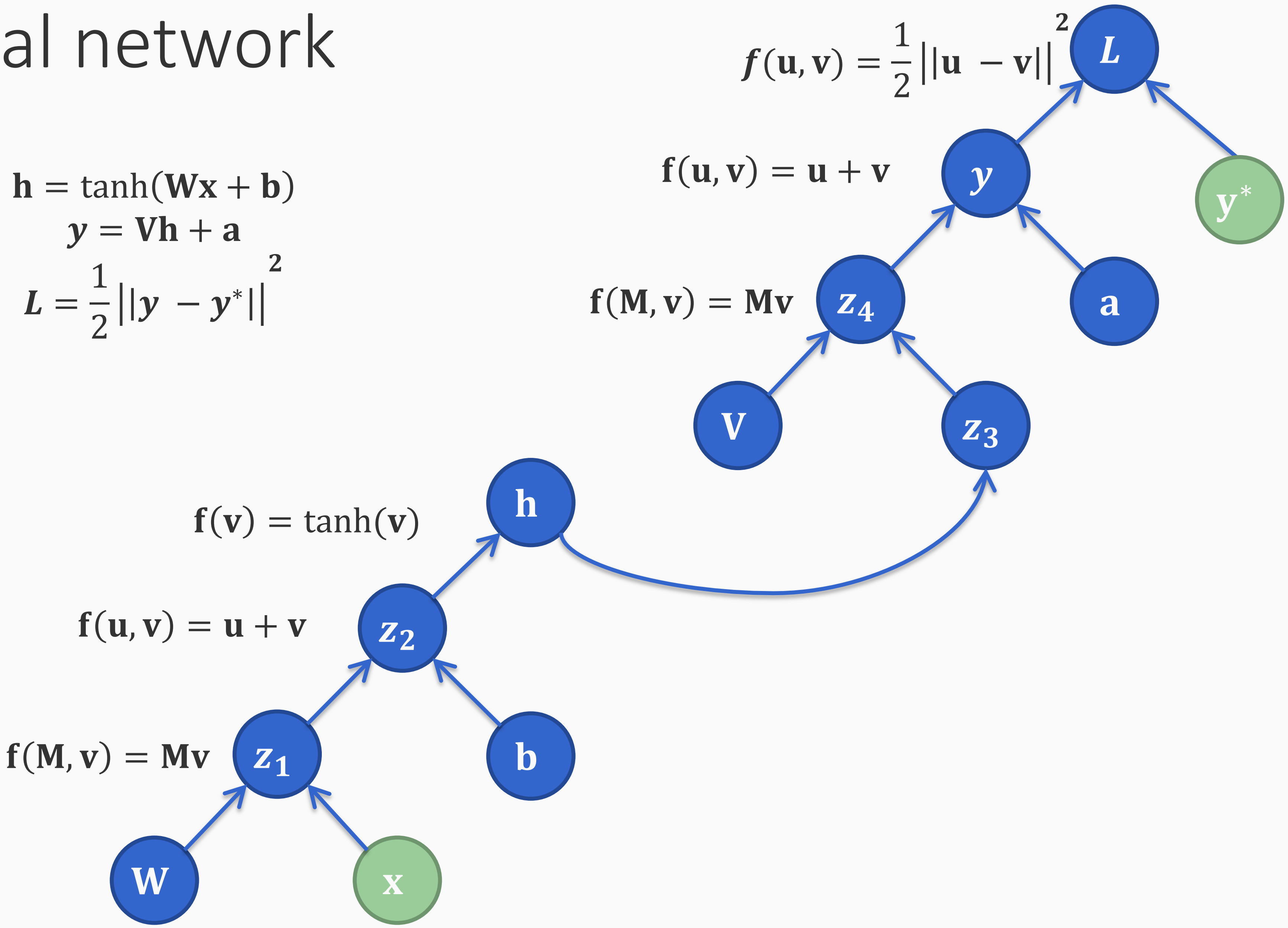
A neural network

$$\mathbf{h} = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{V}\mathbf{h} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{y}^*||^2$$



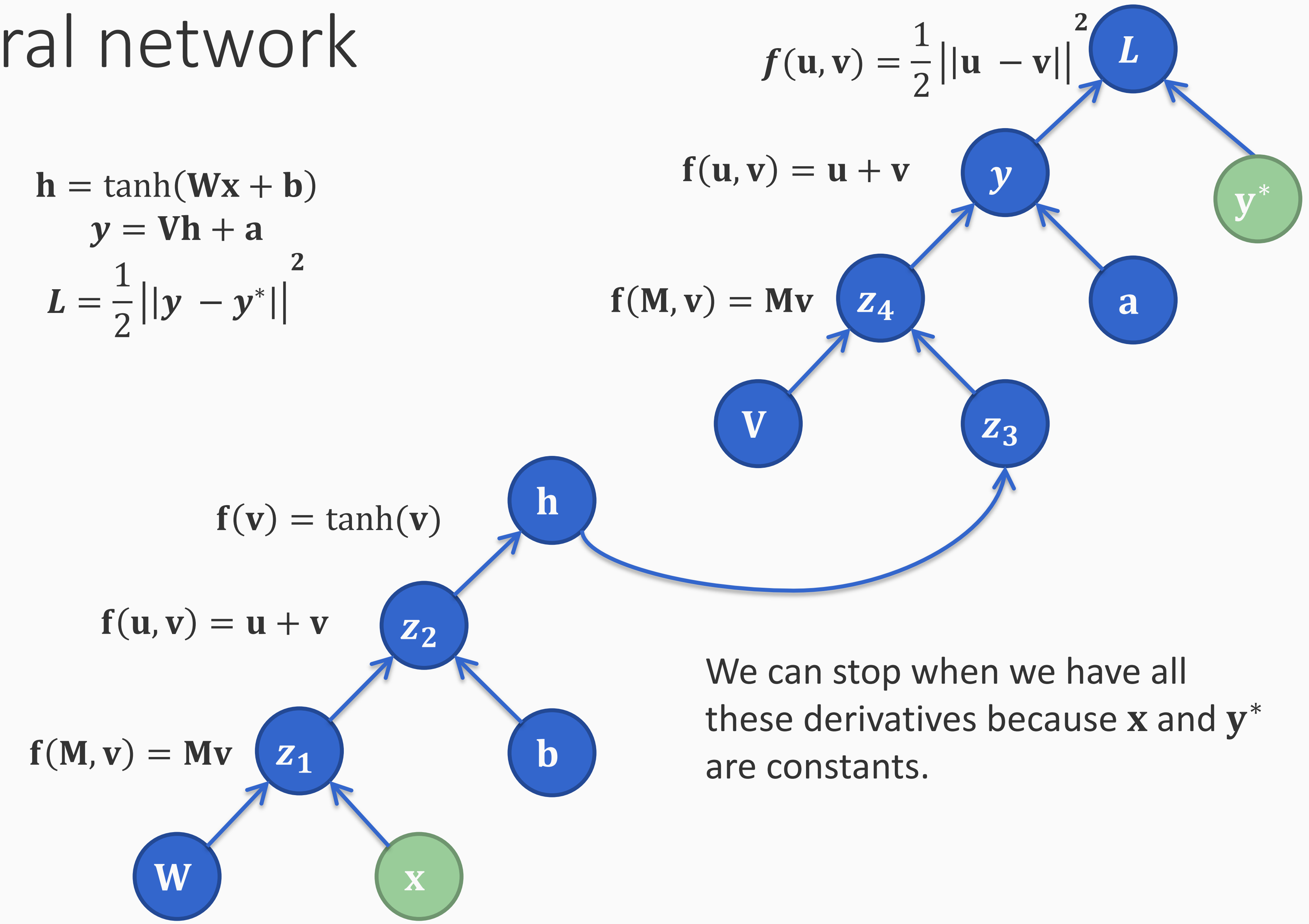
A neural network

$$\mathbf{h} = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b})$$
$$y = \mathbf{V}\mathbf{h} + \mathbf{a}$$
$$L = \frac{1}{2}||y - y^*||^2$$



A neural network

$$\mathbf{h} = \tanh(\mathbf{Wx} + \mathbf{b})$$
$$\boldsymbol{y} = \mathbf{Vh} + \mathbf{a}$$
$$\boldsymbol{L} = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{y}^*||^{\mathbf{2}}$$



We can stop when we have all these derivatives because $\mathbf{x}$ and $\mathbf{y}^*$ are constants.

Backpropagation, in general

After we have done the forward propagation,

Loop over the nodes in reverse topological order starting with a final goal node

- Compute derivatives of final goal node value with respect to each edge's tail node
  - If there are multiple outgoing edges from a node, sum up all the derivatives for the edges