Алгоритм обратного распространения ошибки

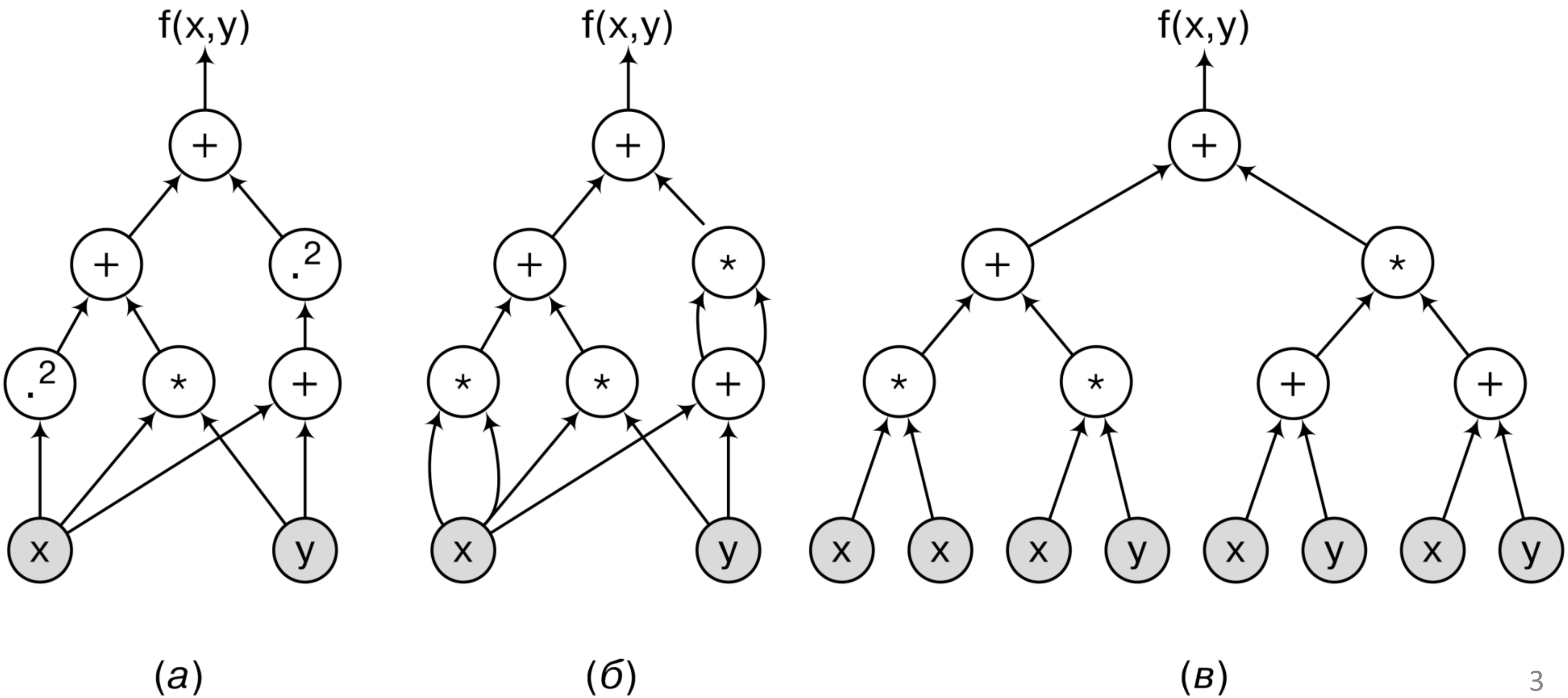
Лекция 2

Граф вычислений

- Нейронная сеть = сложная функция = композиция простых функций
- Градиентный спуск = дифференцирование сложной функции
- *Граф вычислений* – направленный граф, узлами которого являются функции (простые), а ребра связывают функции со своими аргументами
- Современные библиотеки для построения нейронных сетей включают модули автоматического дифференцирования

Граф вычислений

$$f(x, y) = x^2 + xy + (x + y)^2$$



Производная композиции функций

- Производная композиции функций (сложной функции):

$$\left(f\left(g(x)\right)\right)' = f'\left(g(x)\right)g'(x)$$

– цепное правило (chain rule)

$$\frac{df}{dx} = \frac{df}{dg}\frac{dg}{dx},$$

где f, g – скалярные функции, x – скалярная переменная

Производная композиции функций

- Если $\vec{x} = (x_1, \dots, x_d)$, f и g – скалярные функции, тогда градиент композиции функций:

$$\nabla_{\vec{x}} f\big(g(\vec{x})\big) = \begin{pmatrix} \dfrac{\partial f(g)}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(g)}{\partial x_d} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial f}{\partial g} \dfrac{\partial g}{\partial x_1} \\ \vdots \\ \dfrac{\partial f}{\partial g} \dfrac{\partial g}{\partial x_d} \end{pmatrix} = \frac{\partial f}{\partial g} \nabla_{\vec{x}} g$$

Производная композиции функций

- Если x – скалярная переменная,

  $\vec{g} = (g_1, \dots g_m)$ – вектор-функция,

  $f = f\big(g_1(x), \dots, g_m(x)\big)$ – скалярная функция,

  тогда производная:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g_1}\frac{\partial g_1}{\partial x} + \cdots + \frac{\partial f}{\partial g_m}\frac{\partial g_m}{\partial x} = \sum_{i=1}^{m} \frac{\partial f}{\partial g_i}\frac{\partial g_i}{\partial x}$$

Производная композиции функций

- Если $\vec{x} = (x_1, \dots, x_d)$ – вектор,

  $\vec{g} = (g_1, \dots g_m)$ – вектор-функция,

  $f = f\big(g_1(\vec{x}), \dots, g_m(\vec{x})\big)$ – скалярная функция,

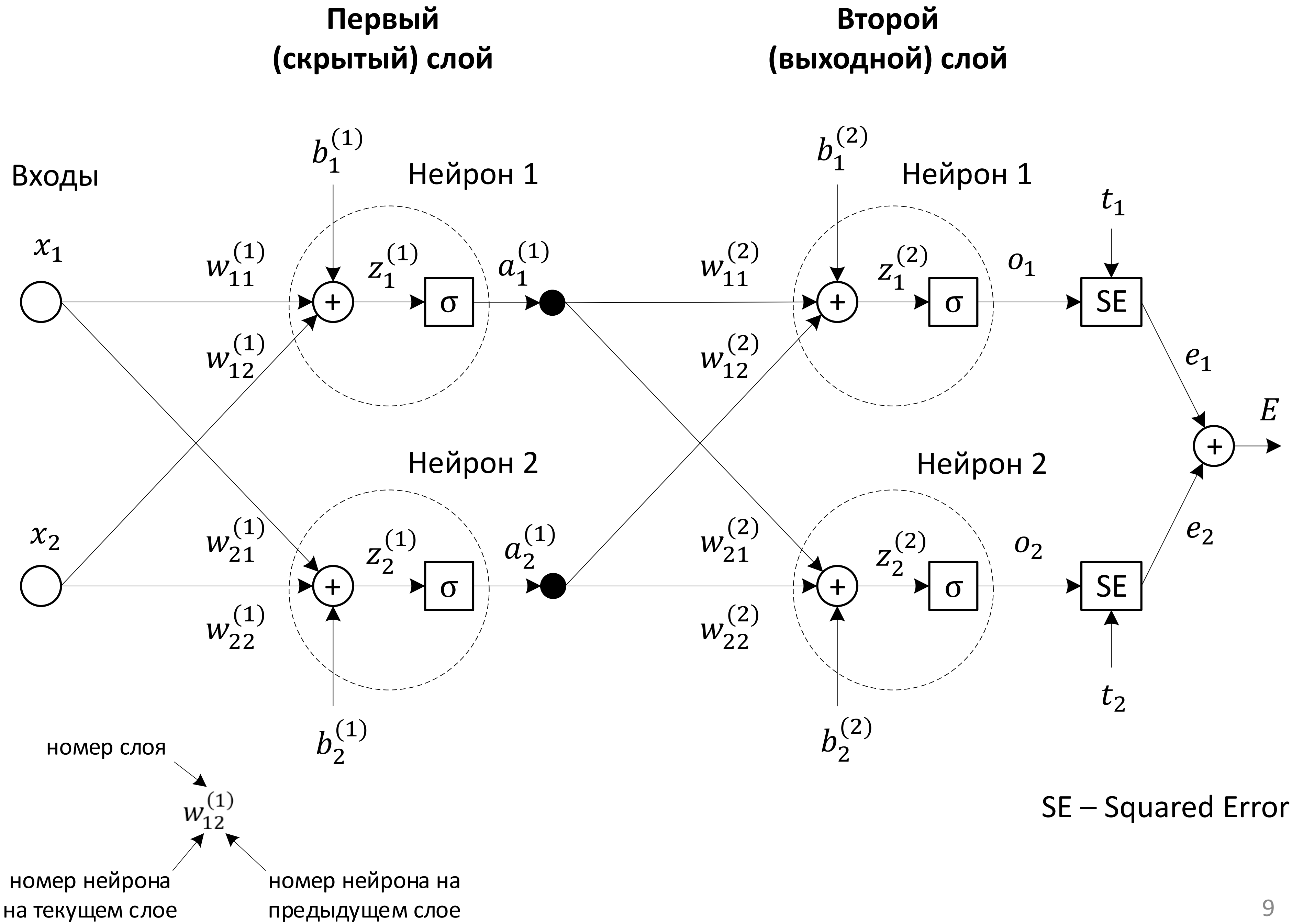
  тогда градиент композиции функций:

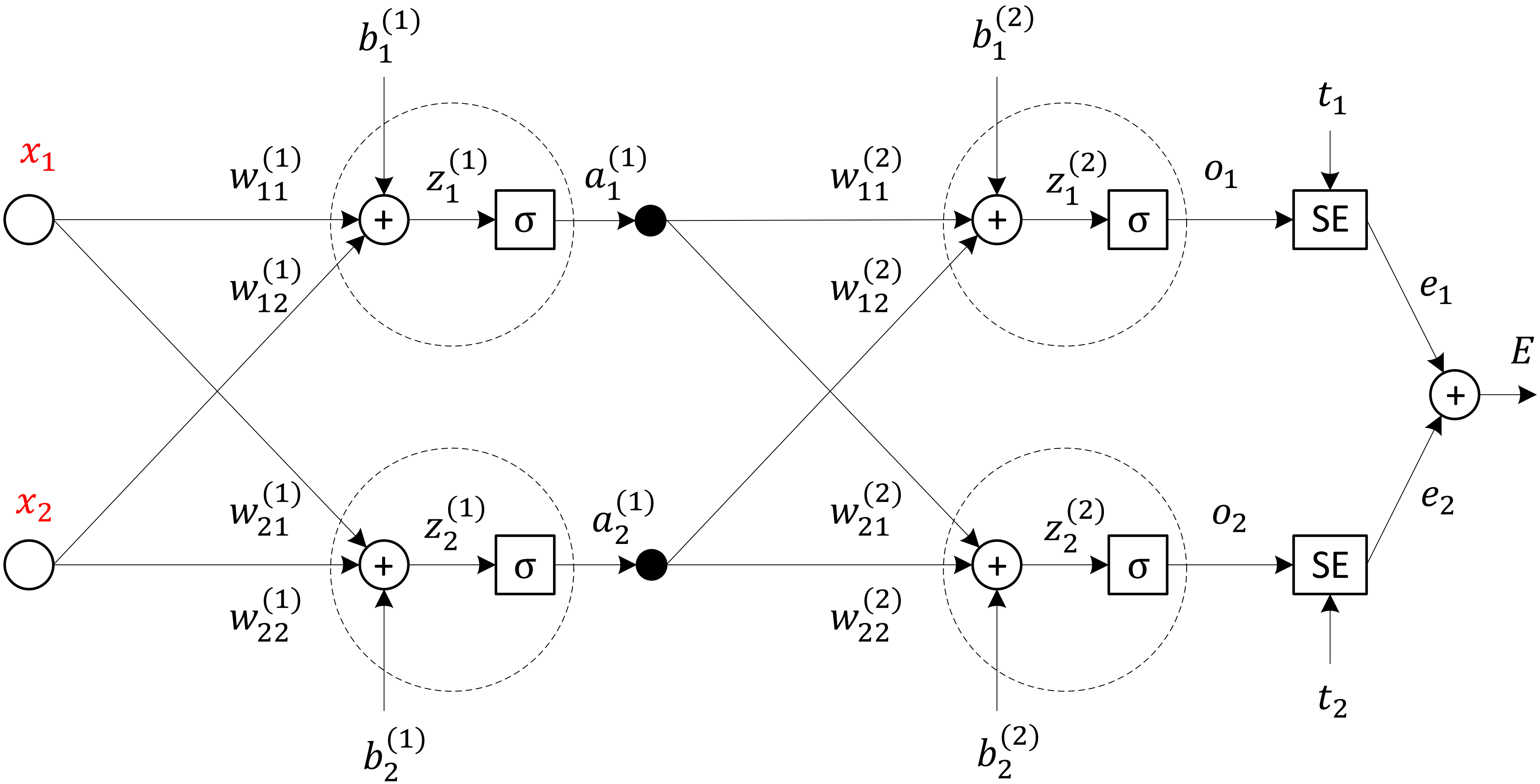
$$\nabla_{\vec{x}} f = \frac{\partial f}{\partial g_1} \nabla_{\vec{x}} g_1 + \cdots + \frac{\partial f}{\partial g_m} \nabla_{\vec{x}} g_m = \sum_{i=1}^{m} \frac{\partial f}{\partial g_i} \nabla_{\vec{x}} g_i$$

$$\nabla_{\vec{x}} f = \nabla_{\vec{x}} \vec{g} \nabla_{\vec{g}} f = \underbrace{\begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_d} & \cdots & \frac{\partial g_m}{\partial x_d} \end{pmatrix}}_{\text{матрица Якоби}} \begin{pmatrix} \frac{\partial f}{\partial g_1} \\ \vdots \\ \frac{\partial f}{\partial g_m} \end{pmatrix}$$

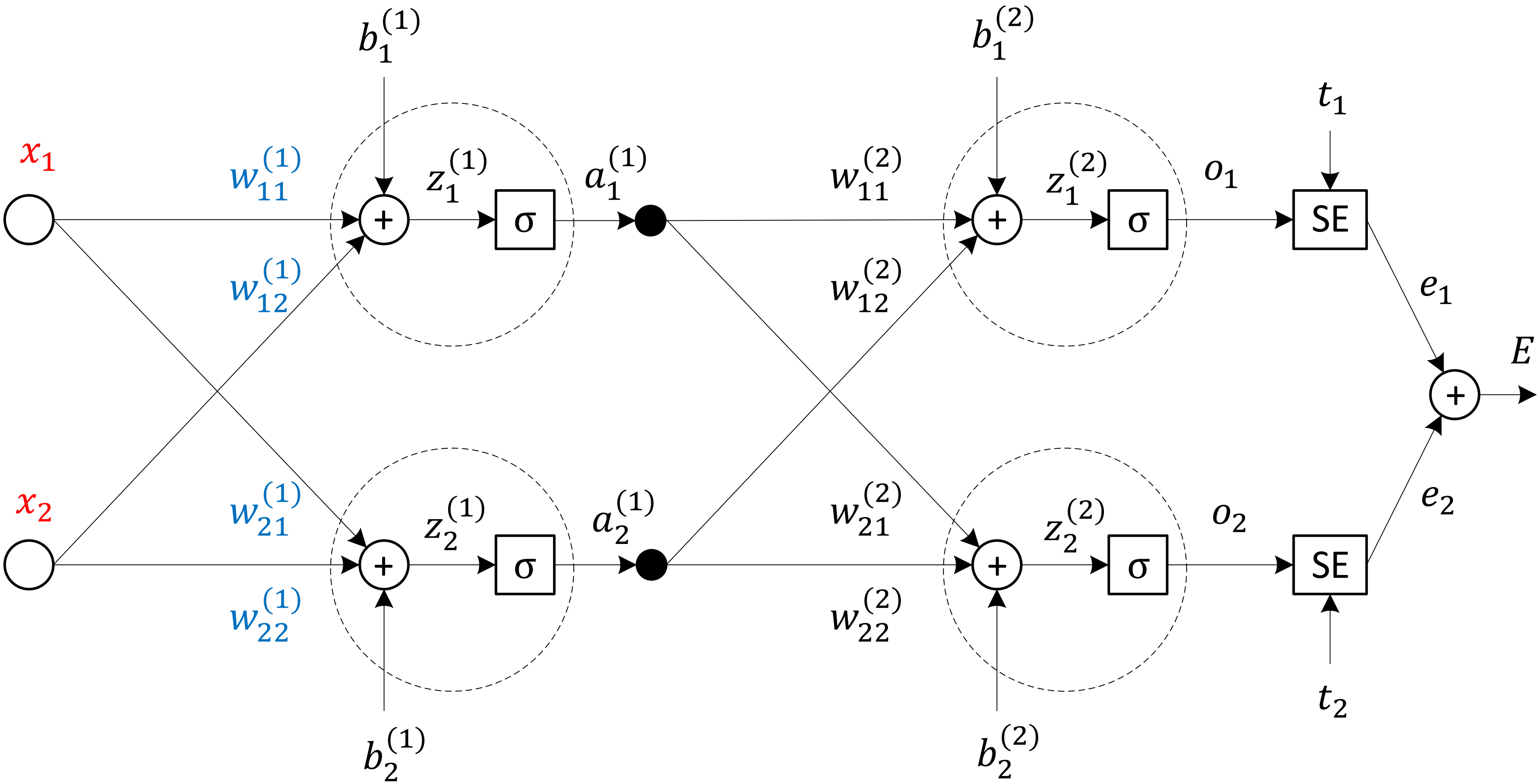
Алгоритм обратного распространения ошибки

- Алгоритм обратного распространения ошибки (**backpropagation**) – метод вычисления градиента функции потерь в нейронных сетях на основе цепного правила
  - Часто подразумевается и изменение весов, то есть градиентный спуск
- Создание:
  - Linnainmaa Seppo – The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors (Masters) (in Finnish). University of Helsinki. 1970
  - Rumelhart David, Hinton Geoffrey, Williams Ronald. Learning representations by back-propagating errors // Nature. 1986. Vol. 323

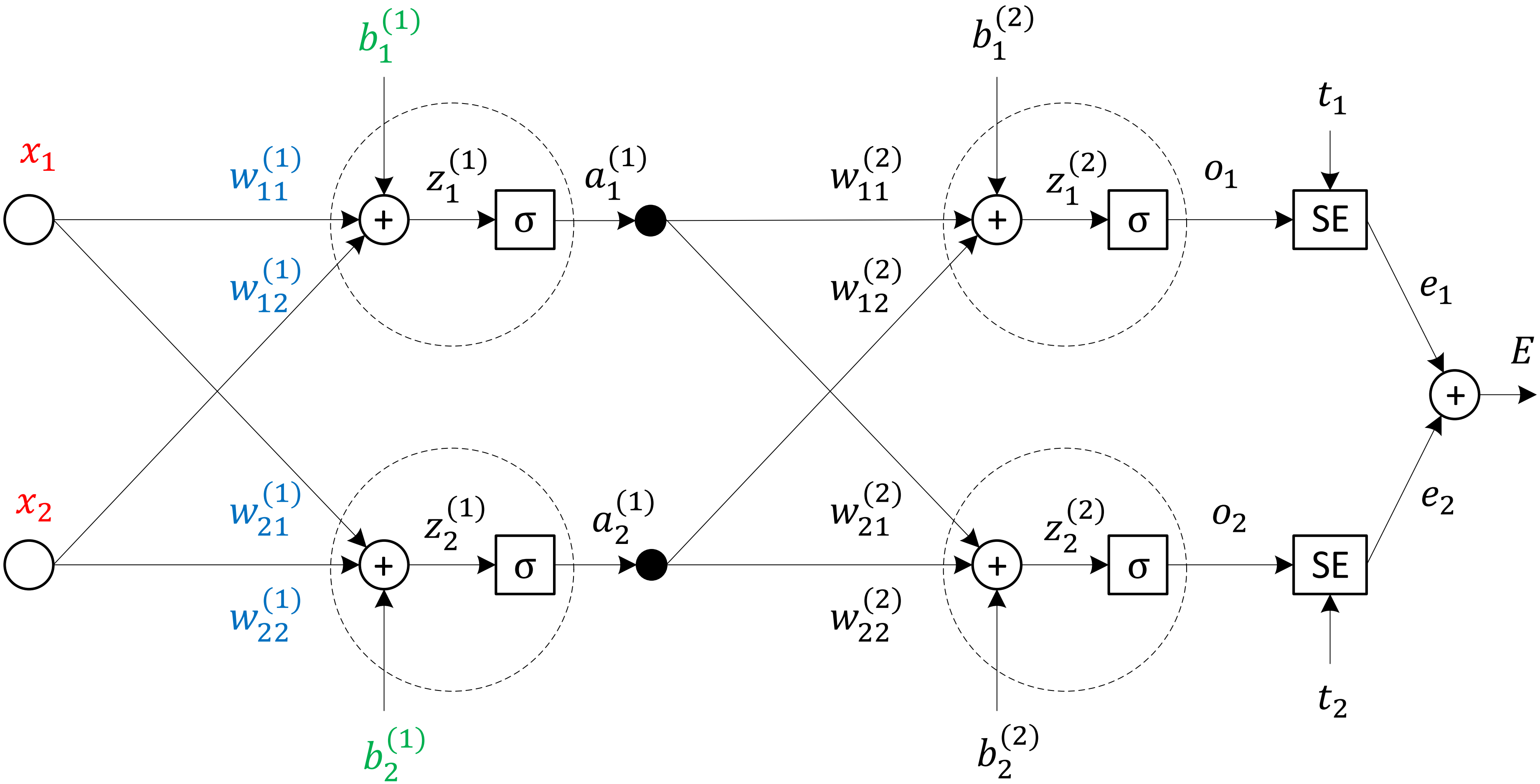

Первый
(скрытый) слой
Второй
(выходной) слой
Входы
x_1
x_2
$b_1^{(1)}$
$b_2^{(1)}$
$b_1^{(2)}$
$b_2^{(2)}$
Нейрон 1
Нейрон 2
Нейрон 1
Нейрон 2
$w_{11}^{(1)}$
$w_{12}^{(1)}$
$w_{21}^{(1)}$
$w_{22}^{(1)}$
$z_1^{(1)}$
$z_2^{(1)}$
$a_1^{(1)}$
$a_2^{(1)}$
$w_{11}^{(2)}$
$w_{12}^{(2)}$
$w_{21}^{(2)}$
$w_{22}^{(2)}$
$z_1^{(2)}$
$z_2^{(2)}$
σ
o_1
o_2
t_1
t_2
SE
e_1
e_2
E
номер слоя
$w_{12}^{(1)}$
номер нейрона
на текущем слое
номер нейрона на
предыдущем слое
SE – Squared Error



$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

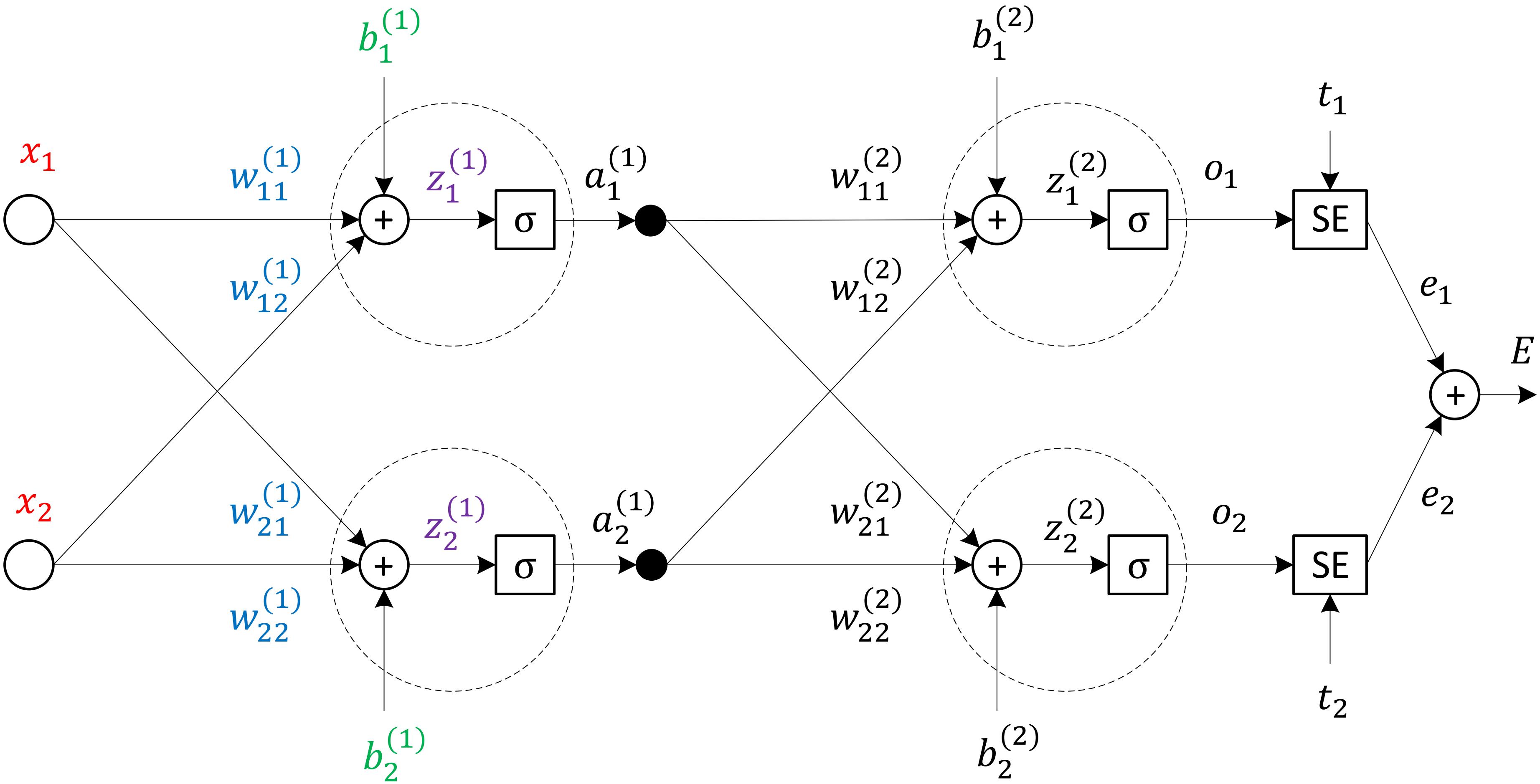


$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \boldsymbol{W}^{(1)} = \begin{array}{c l} \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{bmatrix} & \begin{array}{l} \text{– нейрон 1} \\ \text{– нейрон 2} \end{array} \\ \begin{array}{cc} \text{вход 1} & \text{вход 2} \end{array} & \end{array}$$



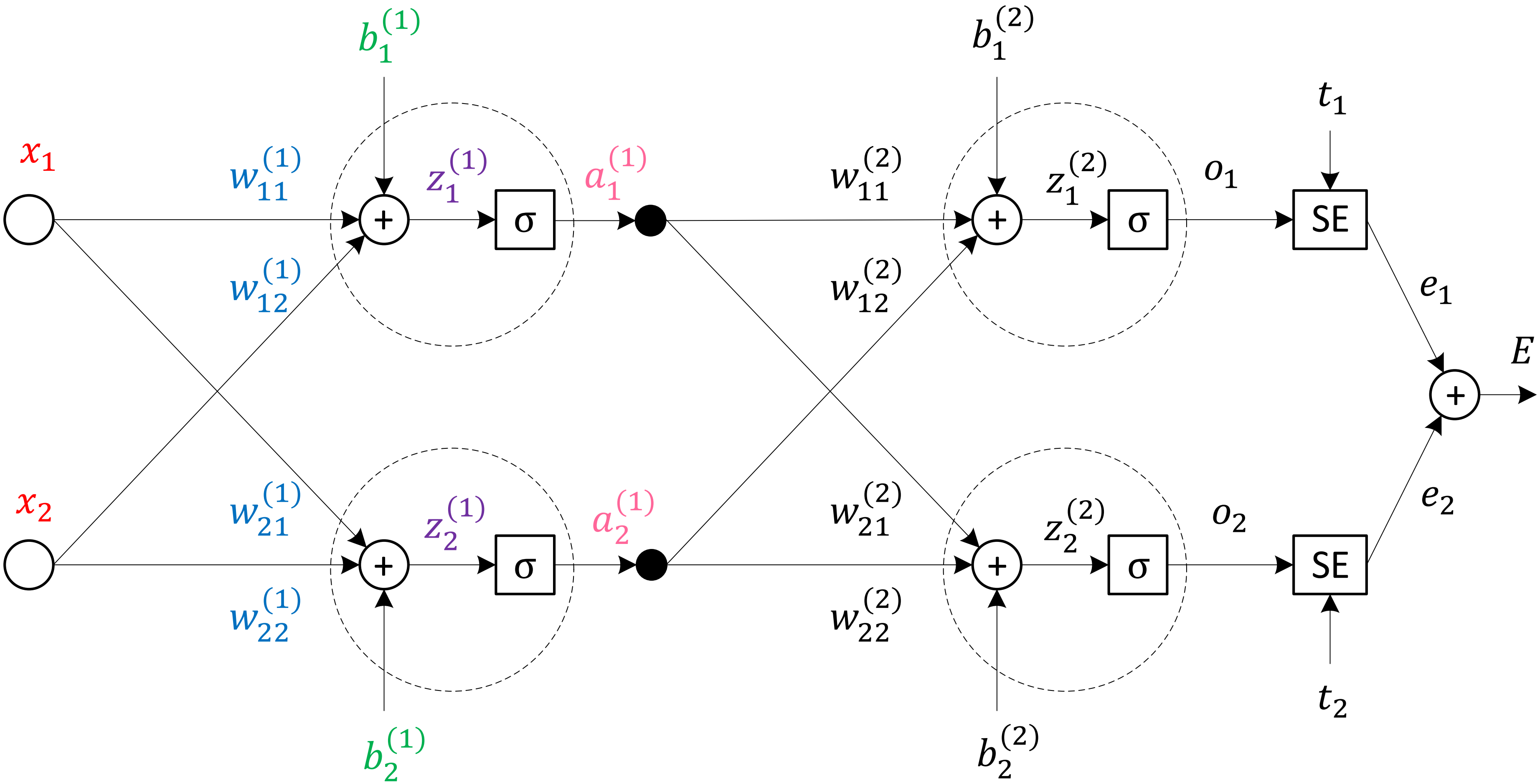
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \boldsymbol{W}^{(1)} = \begin{array}{cc} \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{bmatrix} & \begin{array}{l} \text{– нейрон 1} \\ \text{– нейрон 2} \end{array} \\ \begin{array}{cc} \text{вход 1} & \text{вход 2} \end{array} & \end{array} \qquad \boldsymbol{b}^{(1)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \end{bmatrix}$$

$$\boldsymbol{z}^{(1)} = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$$



$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \boldsymbol{W}^{(1)} = \begin{array}{cl} \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{bmatrix} & \begin{array}{l} \text{– нейрон 1} \\ \text{– нейрон 2} \end{array} \\ \begin{array}{cc} \text{вход 1} & \text{вход 2} \end{array} & \end{array} \qquad \boldsymbol{b}^{(1)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \end{bmatrix}$$

$$\boldsymbol{z}^{(1)} = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)} \qquad \boldsymbol{a}^{(1)} = \sigma\left(\boldsymbol{z}^{(1)}\right)$$



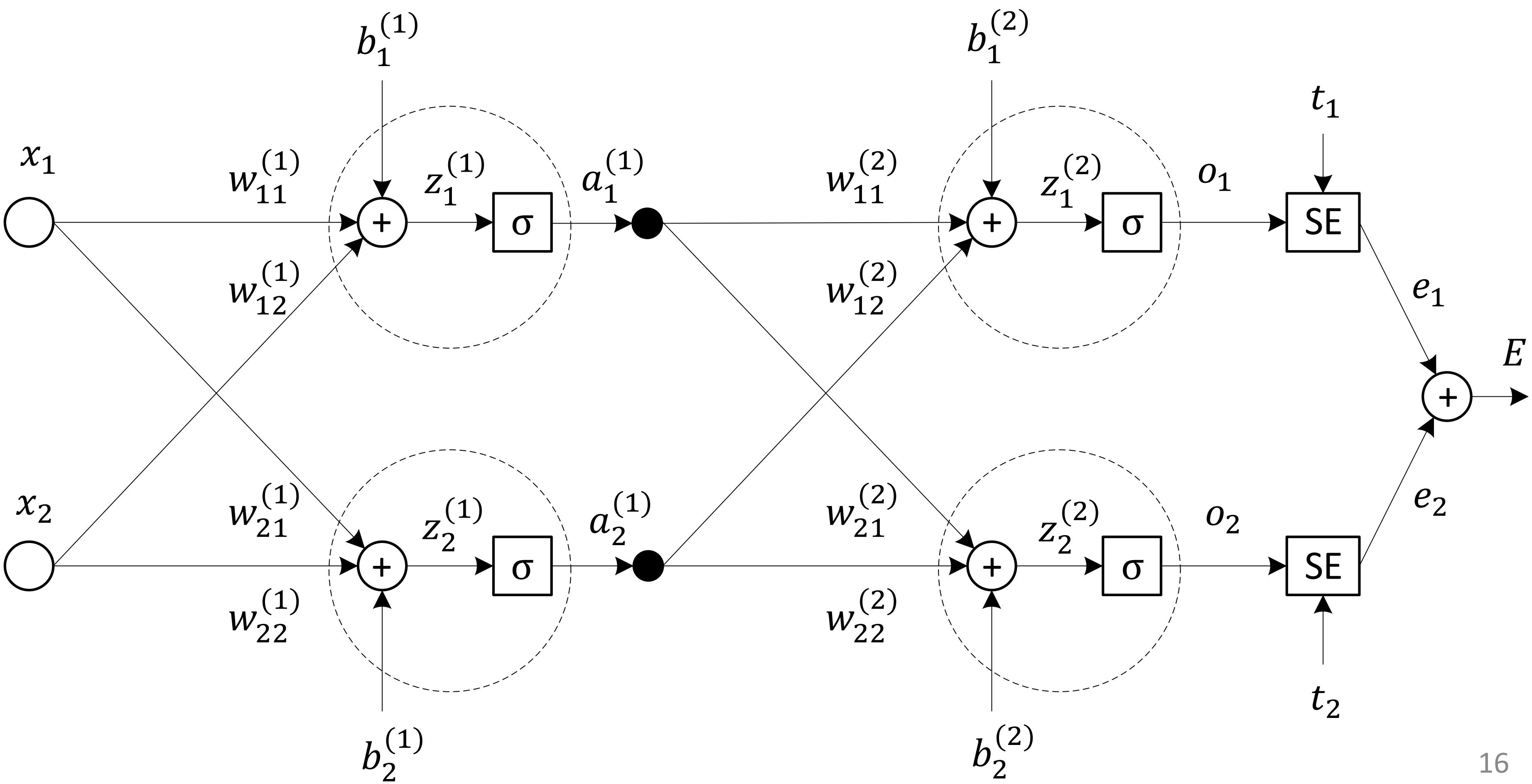
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \boldsymbol{W}^{(1)} = \begin{array}{cc} \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{bmatrix} & \begin{array}{l} \text{– нейрон 1} \\ \text{– нейрон 2} \end{array} \\ \text{вход 1} \quad \text{вход 2} & \end{array} \qquad \boldsymbol{b}^{(1)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \end{bmatrix}$$

Градиент

$$\nabla E_{\boldsymbol{w},\boldsymbol{b}} = \left[\frac{\partial E}{\partial w_{11}^{(1)}}, \frac{\partial E}{\partial w_{12}^{(1)}}, \dots, \frac{\partial E}{\partial w_{22}^{(2)}}, \frac{\partial E}{\partial b_{1}^{(1)}}, \dots, \frac{\partial E}{\partial b_{2}^{(2)}}\right]$$

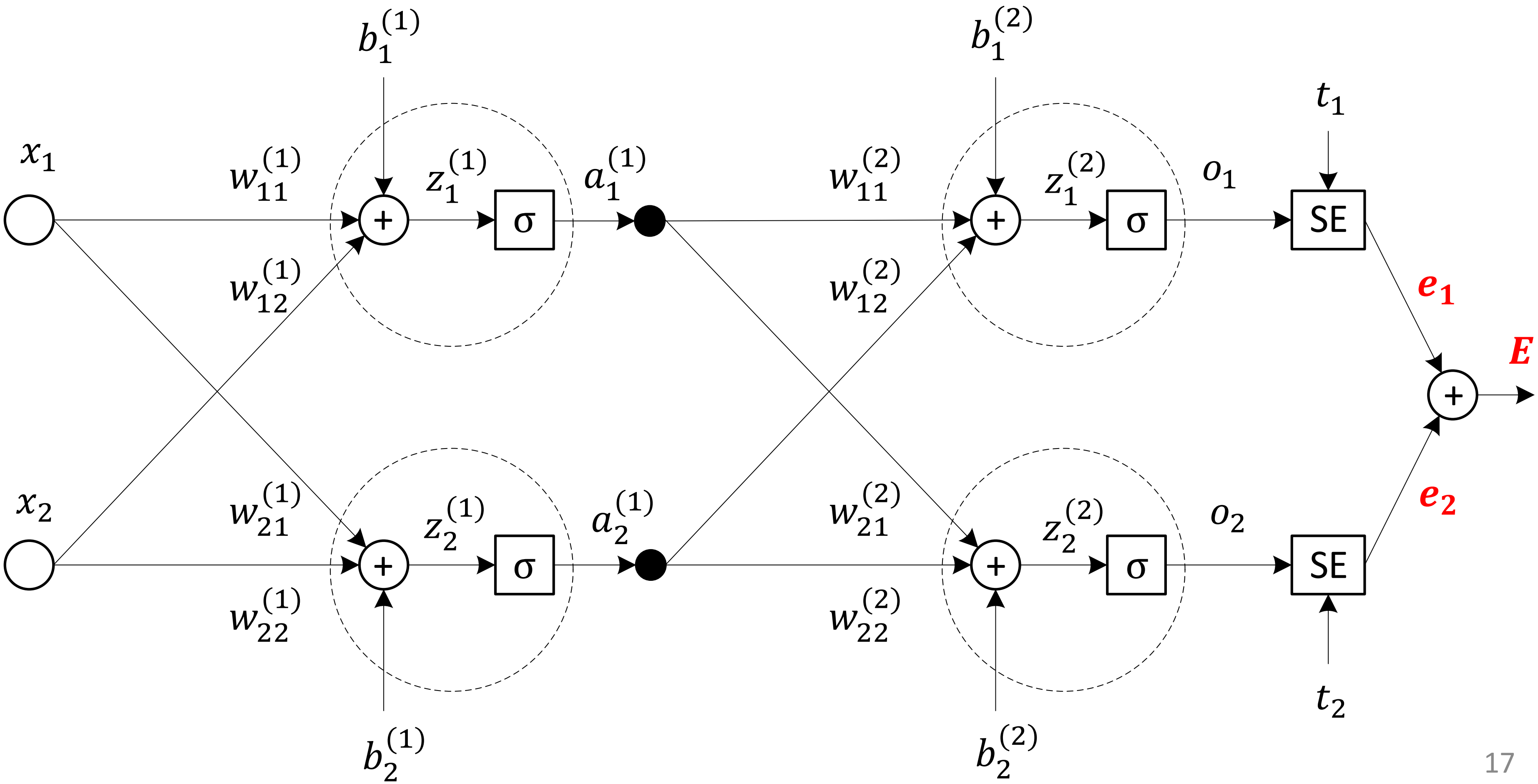
- Значения градиента (частные производные) говорят о скорости изменения (чувствительности) функции в зависимости от изменения соответствующего веса $\left(\frac{\Delta E}{\Delta w_{ij}}\right)$
- То есть градиент показывает, как наиболее эффективно изменить веса, чтобы максимально уменьшить значение функции потерь
- Backpropagation – алгоритм вычисления этого градиента

$$\nabla E_{\boldsymbol{w},\boldsymbol{b}} = \left[\frac{\partial E}{\partial w_{11}^{(1)}}, \frac{\partial E}{\partial w_{12}^{(1)}}, \dots, \frac{\partial E}{\partial w_{22}^{(2)}}, \frac{\partial E}{\partial b_{1}^{(1)}}, \dots, \frac{\partial E}{\partial b_{2}^{(2)}}\right]$$



Квадратичная функция ошибки для одного примера:

$$E = \frac{1}{2}\left\|\vec{t} - \vec{o}\right\|^2 = \frac{1}{2}\sum_j \left(t_j - o_j\right)^2 = \frac{1}{2}(t_1 - o_1)^2 + \frac{1}{2}(t_2 - o_2)^2 = e_1 + e_2$$

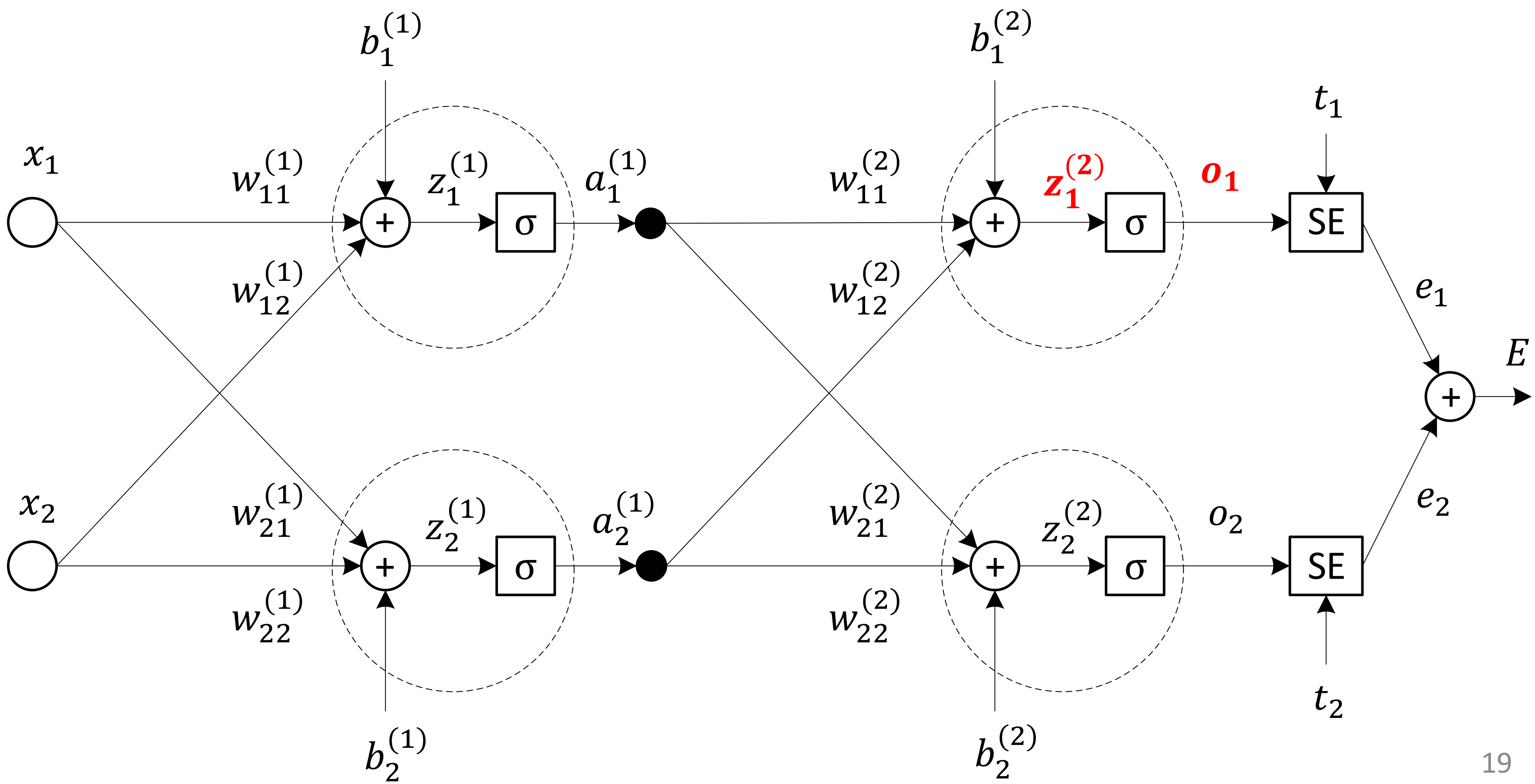


Градиент

$$\nabla E_{\boldsymbol{w},\boldsymbol{b}} = \left[\frac{\partial E}{\partial w_{11}^{(1)}}, \frac{\partial E}{\partial w_{12}^{(1)}}, \dots, \frac{\partial E}{\partial w_{22}^{(2)}}, \frac{\partial E}{\partial b_{1}^{(1)}}, \dots, \frac{\partial E}{\partial b_{2}^{(2)}}\right]$$

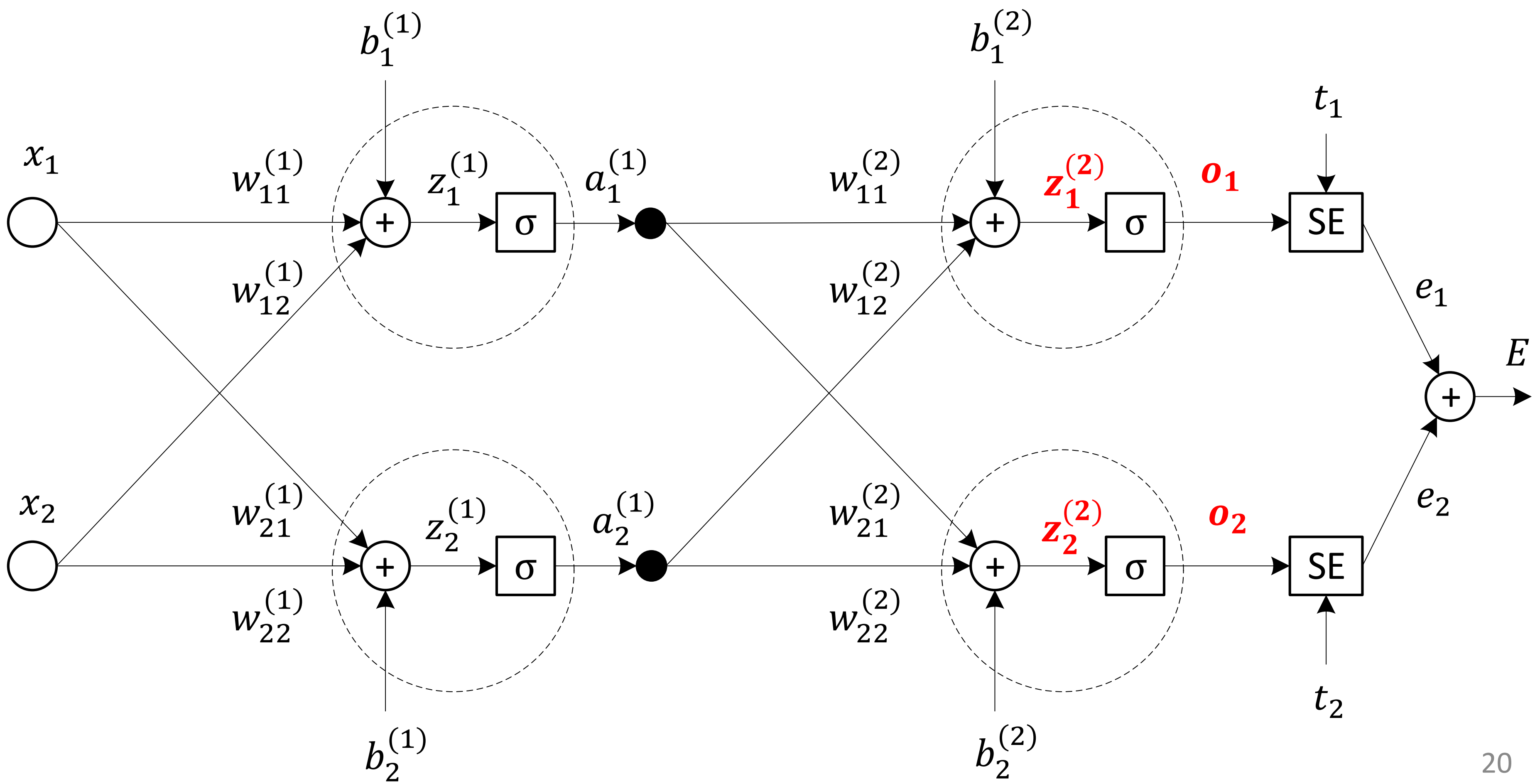
$$E = e_1 + e_2 = \frac{1}{2}(t_1 - o_1)^2 + \frac{1}{2}(t_2 - o_2)^2$$

$$o_1 = \sigma\left(z_1^{(2)}\right) = \sigma\left(w_{11}^{(2)} a_1^{(1)} + w_{12}^{(2)} a_2^{(1)} + b_1^{(2)}\right)$$



$$o_1 = \sigma\left(z_1^{(2)}\right) = \sigma\left(w_{11}^{(2)} a_1^{(1)} + w_{12}^{(2)} a_2^{(1)} + b_1^{(2)}\right)$$

$$o_2 = \sigma\left(z_2^{(2)}\right) = \sigma\left(w_{21}^{(2)} a_1^{(1)} + w_{22}^{(2)} a_2^{(1)} + b_2^{(2)}\right)$$



Градиент

$$\nabla E_{\boldsymbol{w},\boldsymbol{b}} = \left[\frac{\partial E}{\partial w_{11}^{(1)}}, \frac{\partial E}{\partial w_{12}^{(1)}}, \dots, \frac{\partial E}{\partial w_{22}^{(2)}}, \frac{\partial E}{\partial b_{1}^{(1)}}, \dots, \frac{\partial E}{\partial b_{2}^{(2)}}\right]$$

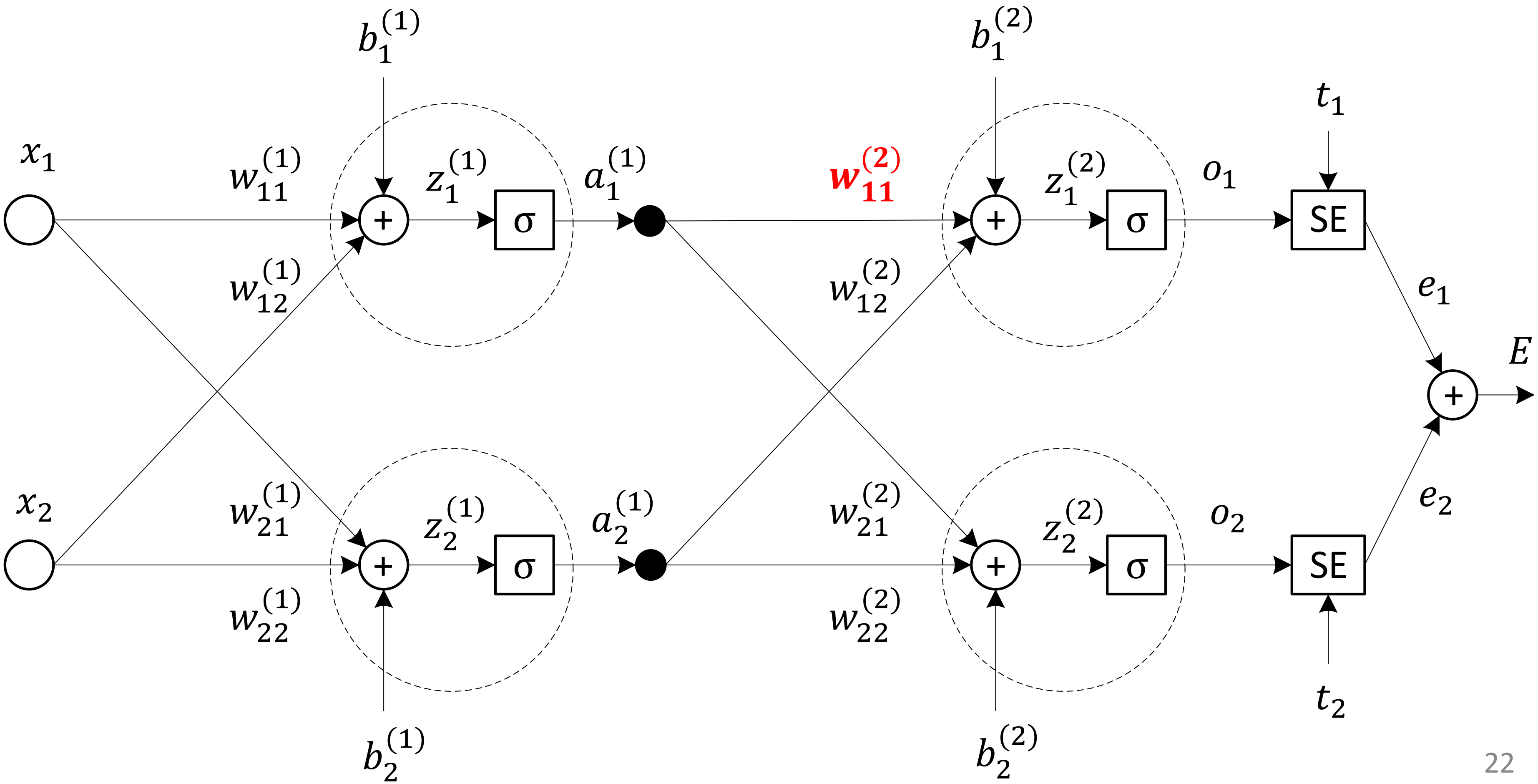
$$E = e_1 + e_2 = \frac{1}{2}(t_1 - o_1)^2 + \frac{1}{2}(t_2 - o_2)^2$$

$$o_1 = \sigma\left(z_1^{(2)}\right) = \sigma\left(w_{11}^{(2)} a_1^{(1)} + w_{12}^{(2)} a_2^{(1)} + b_1^{(2)}\right)$$

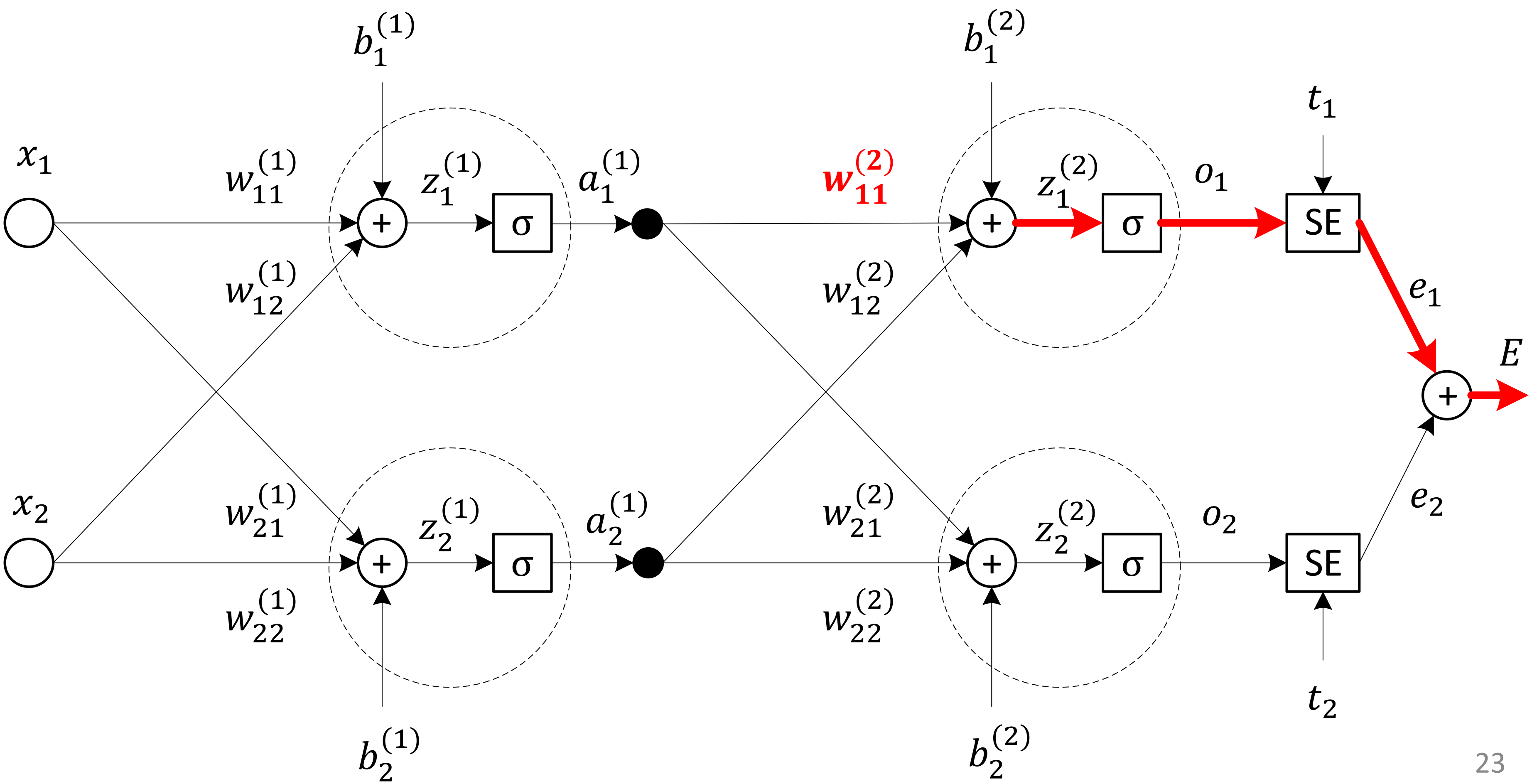
$$o_2 = \sigma\left(z_2^{(2)}\right) = \sigma\left(w_{21}^{(2)} a_1^{(1)} + w_{22}^{(2)} a_2^{(1)} + b_2^{(2)}\right)$$

Градиент выходного слоя

$$\frac{\partial E}{\partial w_{11}^{(2)}} = ?$$



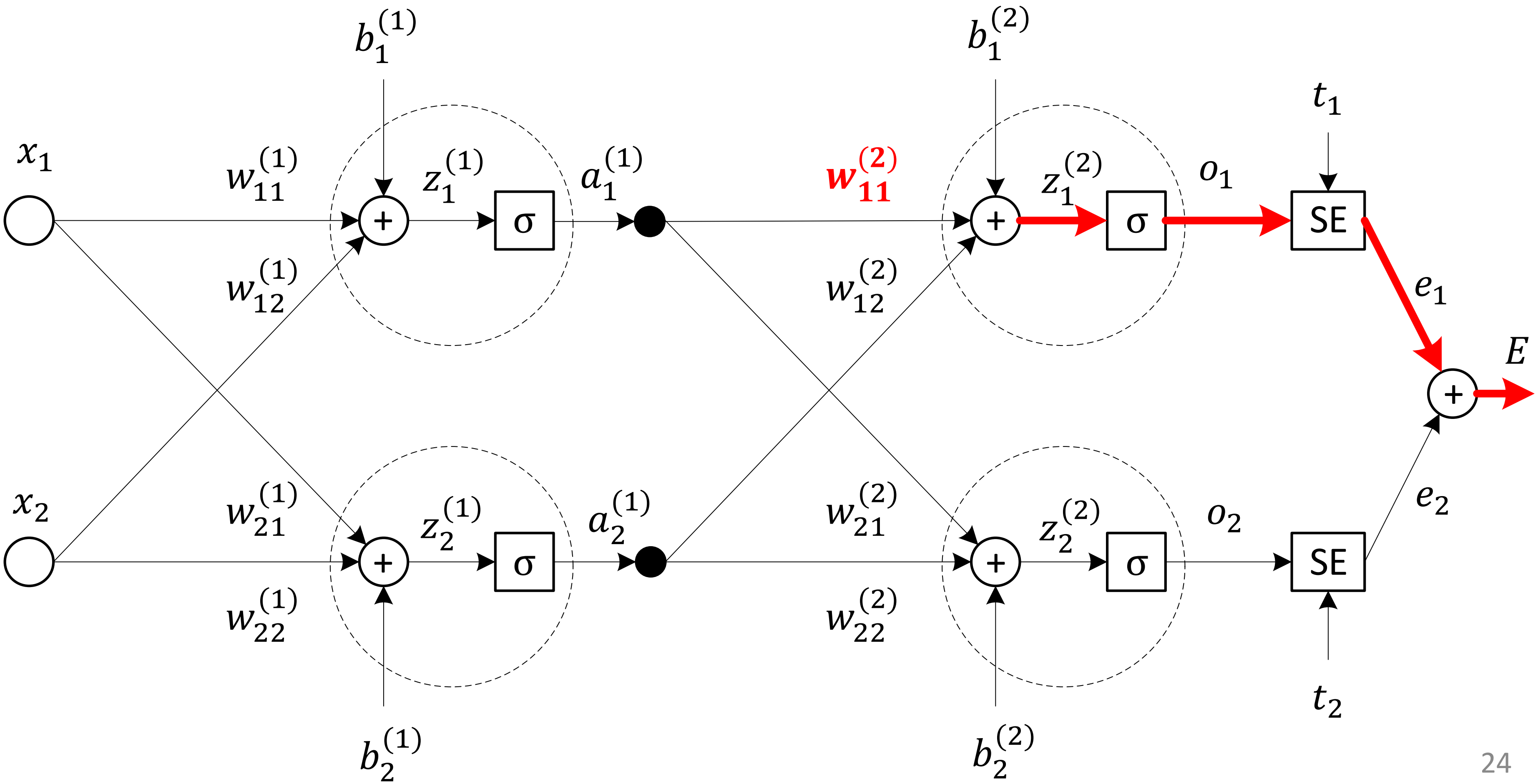
$$\frac{\partial E}{\partial w_{11}^{(2)}} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial z_1^{(2)}}\frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}}$$ – цепное правило, (chain rule)



$$\frac{\partial E}{\partial w_{11}^{(2)}} = \frac{\partial E}{\partial o_1} \frac{\partial o_1}{\partial z_1^{(2)}} \frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}}$$

$$E = e_1 + e_2 = \frac{1}{2}(t_1 - o_1)^2 + \frac{1}{2}(t_2 - o_2)^2$$

$$\frac{\partial E}{\partial o_1} = -(t_1 - o_1)$$

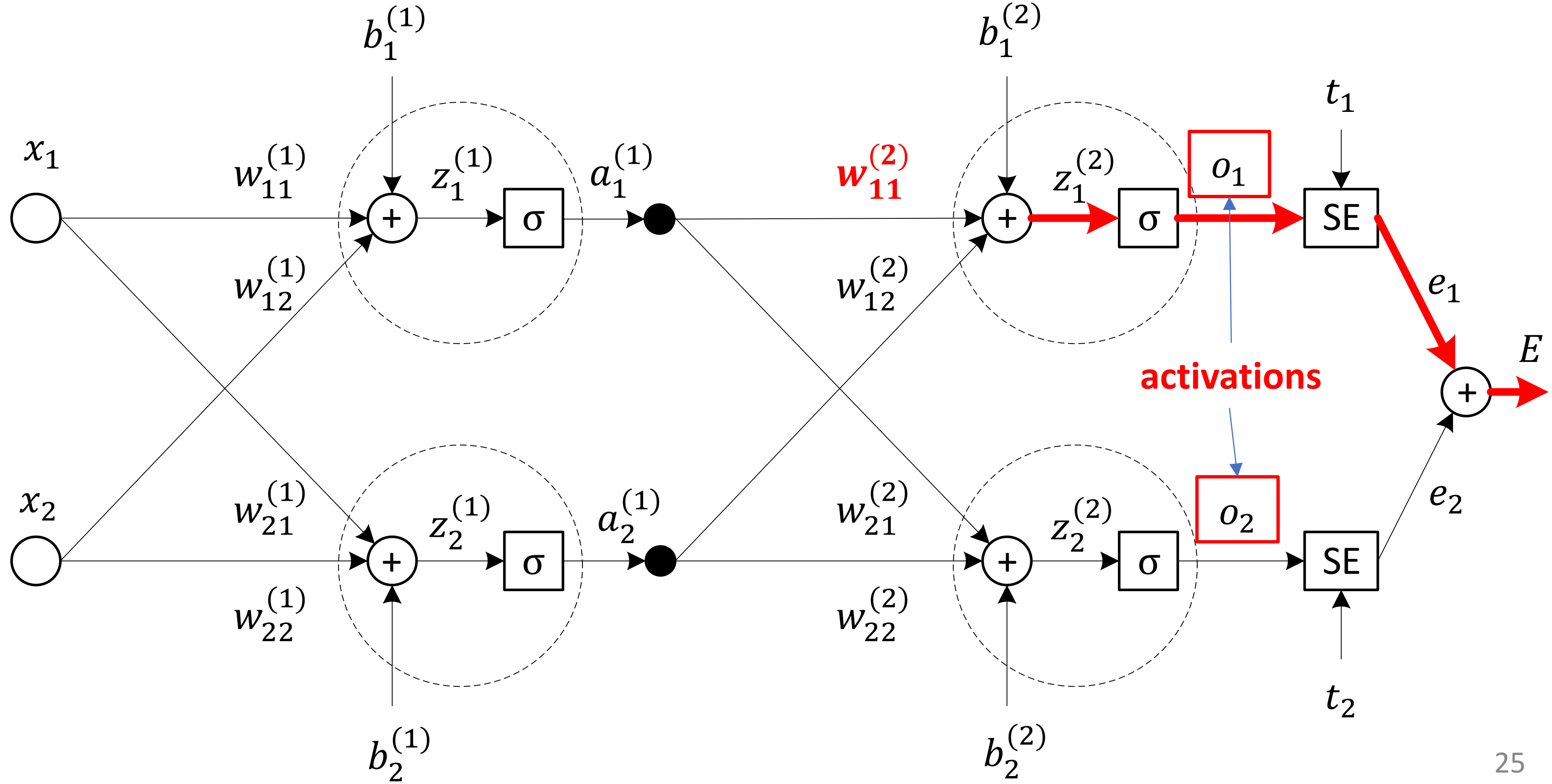


$$\frac{\partial E}{\partial w_{11}^{(2)}} = \frac{\partial E}{\partial o_1} \frac{\boldsymbol{\partial o_1}}{\boldsymbol{\partial z_1^{(2)}}} \frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}}$$

$$o_1 = \sigma\left(z_1^{(2)}\right)$$

Храним активации (а $z_j^{(i)}$ – не храним)

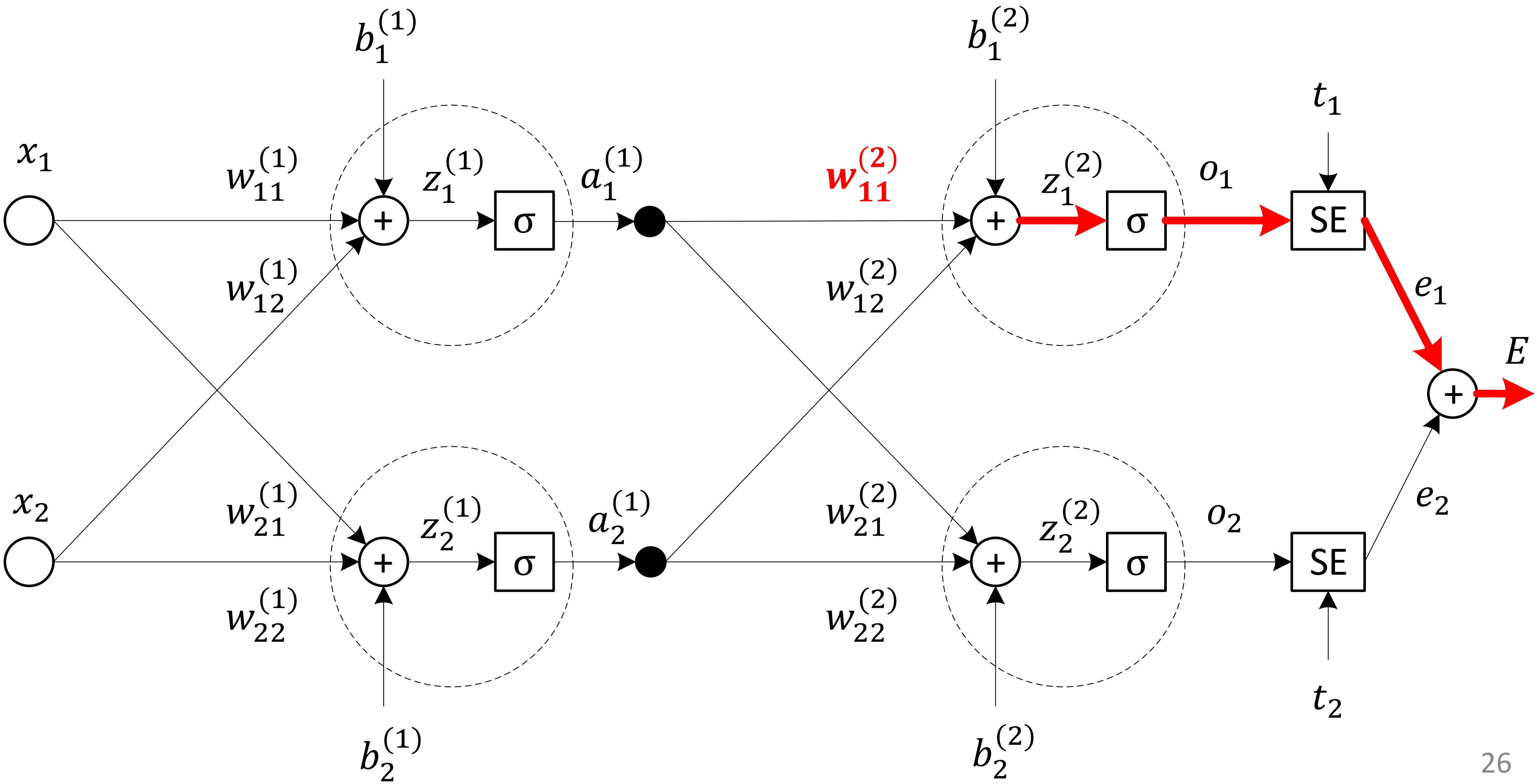
$$\frac{\partial o_1}{\partial z_1^{(2)}} = o_1(1 - o_1)$$



$$\frac{\partial E}{\partial w_{11}^{(2)}} = \frac{\partial E}{\partial o_1} \frac{\partial o_1}{\partial z_1^{(2)}} \frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}}$$

$$z_1^{(2)} = w_{11}^{(2)} a_1^{(1)} + w_{12}^{(2)} a_2^{(1)} + b_1^{(2)}$$

$$\frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}} = a_1^{(1)}$$



Градиент выходного слоя

$$\frac{\partial E}{\partial w_{11}^{(2)}} = \frac{\partial E}{\partial o_1} \frac{\partial o_1}{\partial z_1^{(2)}} \frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}} = -(t_1 - o_1)o_1(1 - o_1)a_1^{(1)}$$

Градиент выходного слоя

$$\frac{\partial E}{\partial w_{11}^{(2)}} = \frac{\partial E}{\partial o_1} \frac{\partial o_1}{\partial z_1^{(2)}} \frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}} = -(t_1 - o_1)o_1(1 - o_1)a_1^{(1)}$$

$$\frac{\partial E}{\partial w_{12}^{(2)}} = \frac{\partial E}{\partial o_1} \frac{\partial o_1}{\partial z_1^{(2)}} \frac{\partial z_1^{(2)}}{\partial w_{12}^{(2)}} = -(t_1 - o_1)o_1(1 - o_1)a_2^{(1)}$$

$$\frac{\partial E}{\partial b_1^{(2)}} = \frac{\partial E}{\partial o_1} \frac{\partial o_1}{\partial z_1^{(2)}} \frac{\partial z_1^{(2)}}{\partial b_1^{(2)}} = -(t_1 - o_1)o_1(1 - o_1)$$

Градиент выходного слоя

$$\frac{\partial E}{\partial w_{11}^{(2)}} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial z_1^{(2)}}\frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}} = -(t_1 - o_1)o_1(1 - o_1)\, a_1^{(1)} = \delta_1^{(2)} a_1^{(1)}$$

$$\frac{\partial E}{\partial w_{12}^{(2)}} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial z_1^{(2)}}\frac{\partial z_1^{(2)}}{\partial w_{12}^{(2)}} = -(t_1 - o_1)o_1(1 - o_1)\, a_2^{(1)} = \delta_1^{(2)} a_2^{(1)}$$

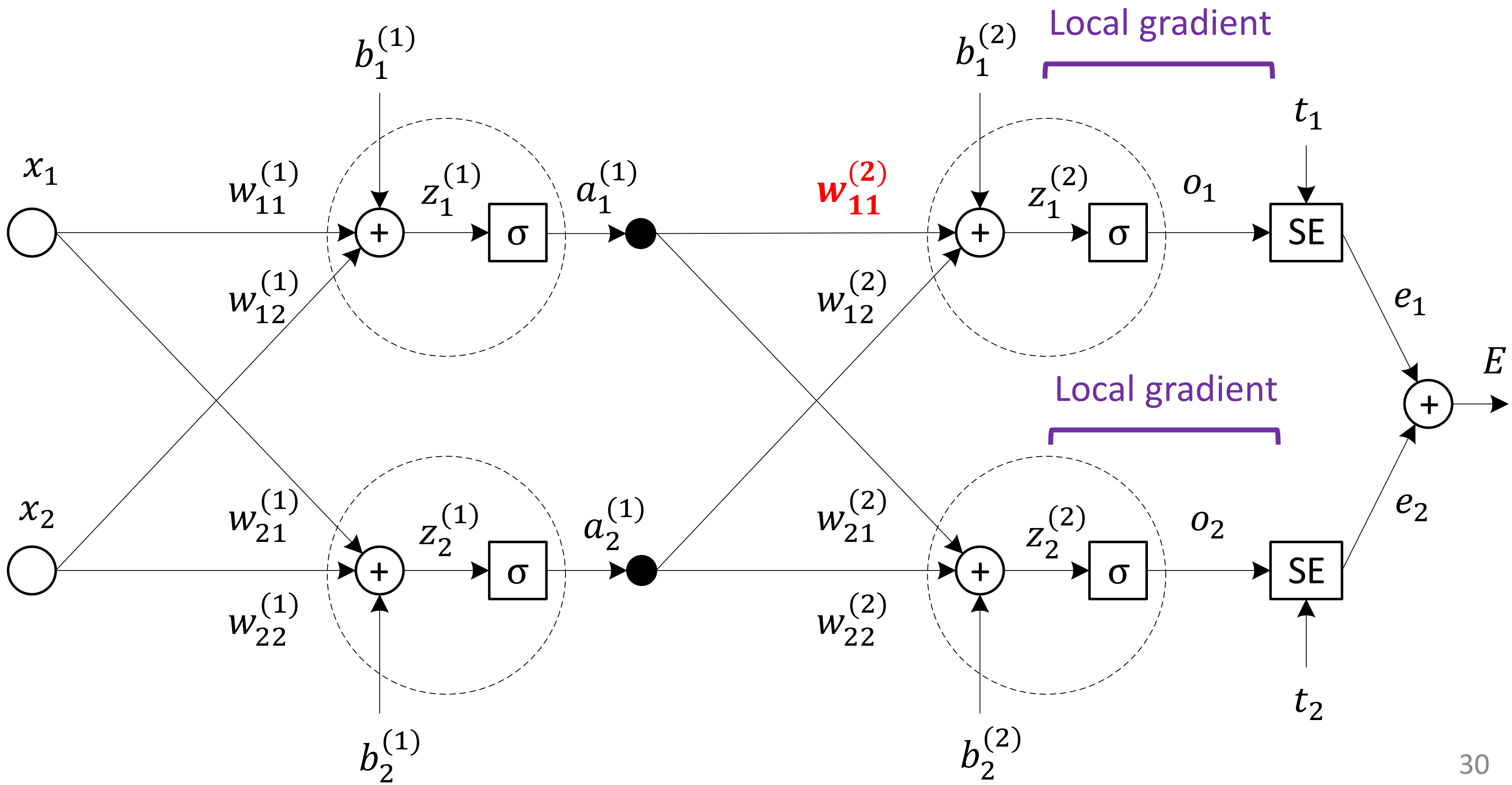
$$\frac{\partial E}{\partial b_1^{(2)}} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial z_1^{(2)}}\frac{\partial z_1^{(2)}}{\partial b_1^{(2)}} = -(t_1 - o_1)o_1(1 - o_1) = \delta_1^{(2)}$$

$$\delta_1^{(2)} = \frac{\partial E}{\partial z_1^{(2)}}$$

Local gradient (или *ошибка нейрона*)

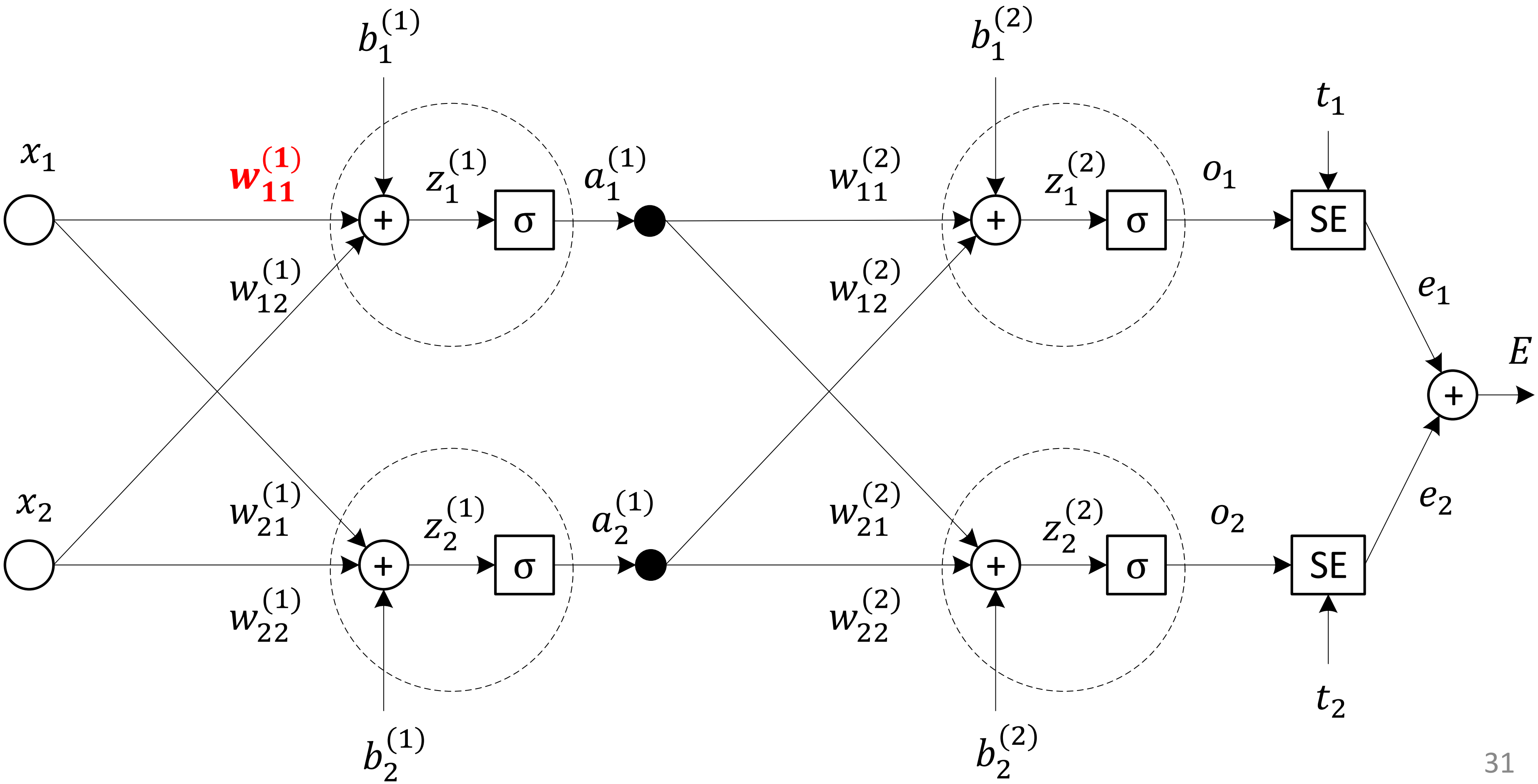
(вычисляем для каждого нейрона однократно и сохраняем)

$$\frac{\partial E}{\partial w_{11}^{(2)}} = \frac{\partial E}{\partial o_1}\frac{\partial o_1}{\partial z_1^{(2)}}\frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}} = \frac{\partial E}{\partial z_1^{(2)}}\frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}}$$



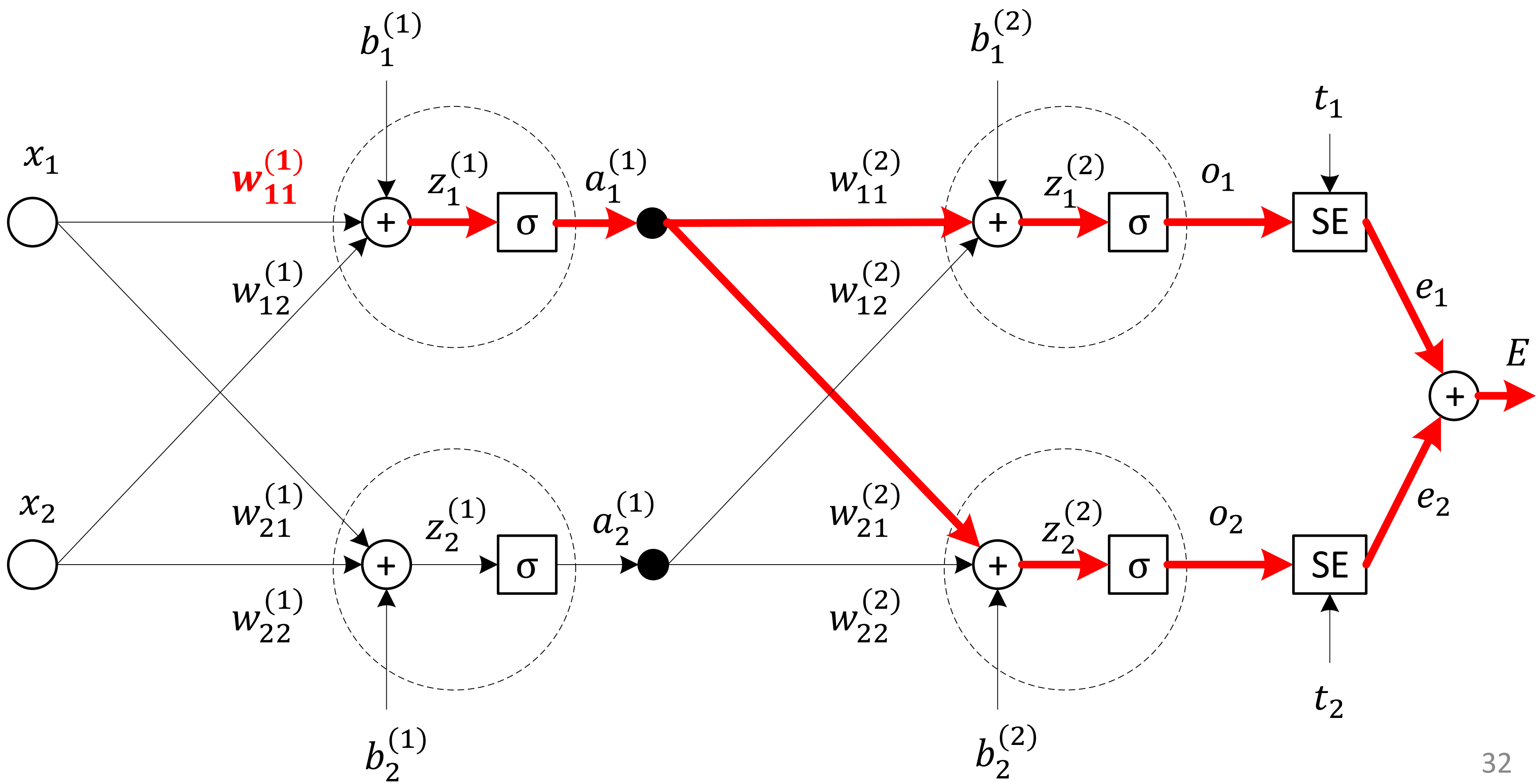
Градиент скрытого слоя

$$\frac{\partial E}{\partial w_{11}^{(1)}} =$$



$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}}$$

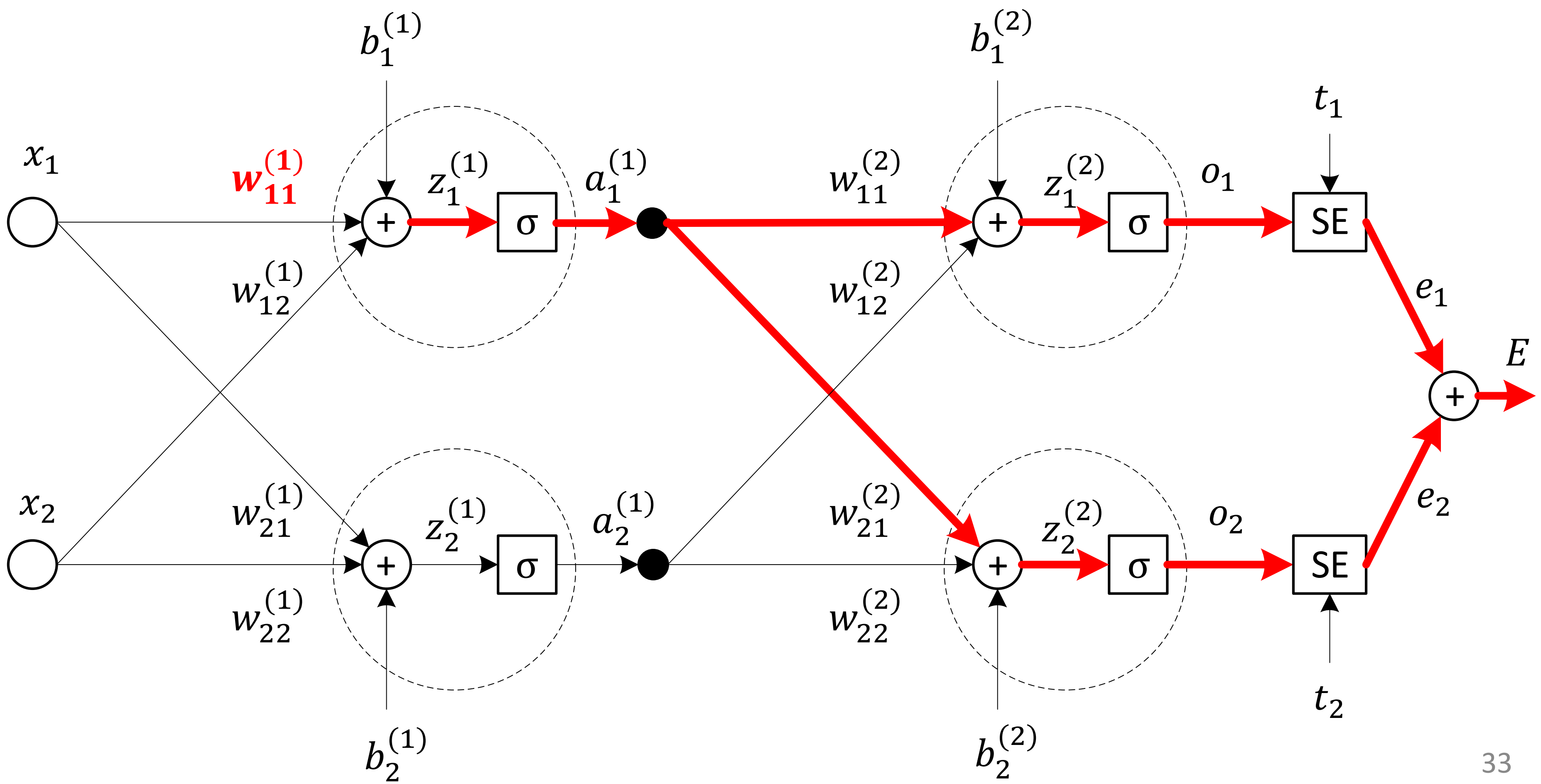
– цепное правило, (chain rule)



$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}}$$

$$E = e_1 + e_2 = \frac{1}{2}(t_1 - o_1)^2 + \frac{1}{2}(t_2 - o_2)^2$$

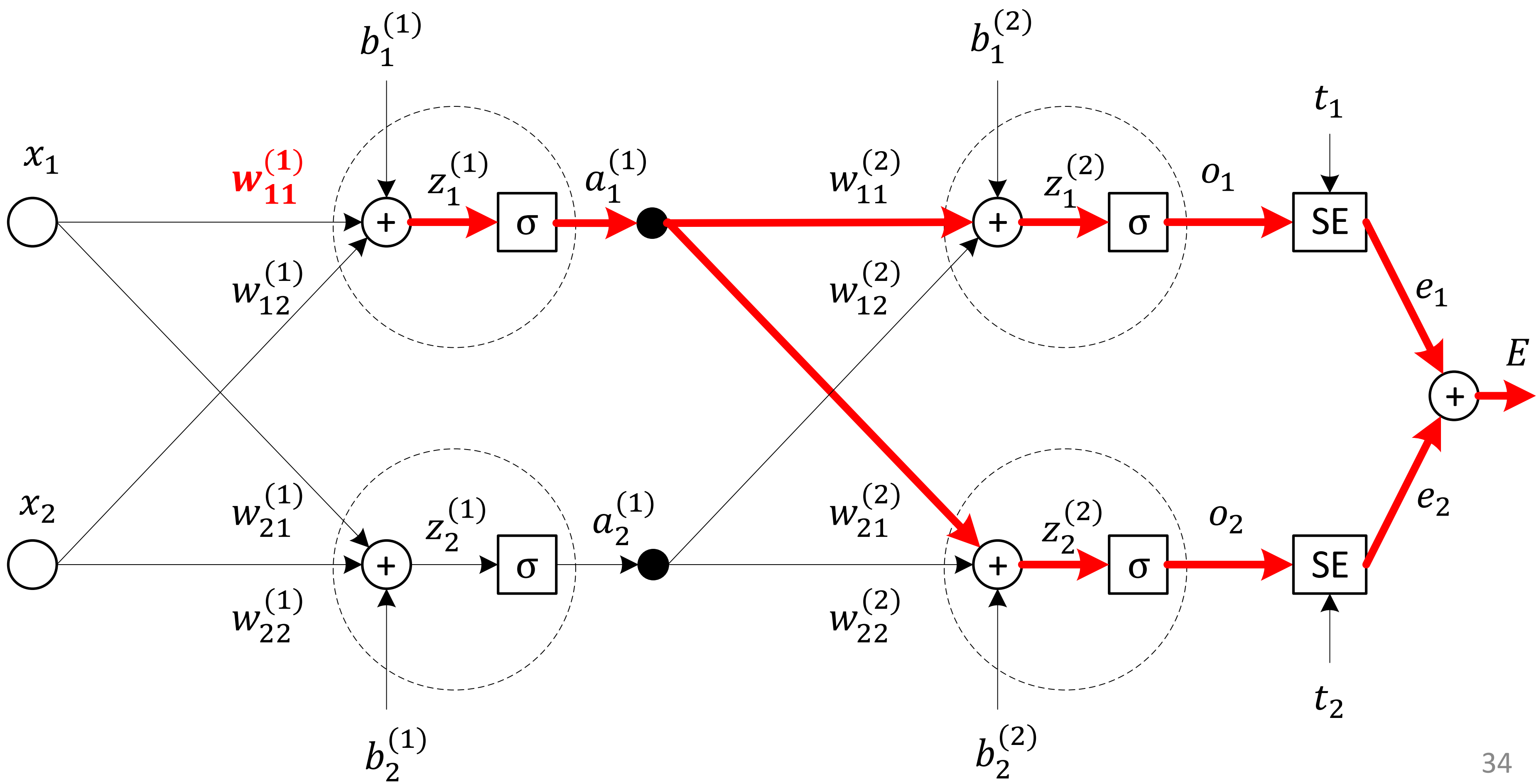
$$\frac{\partial E}{\partial a_1^{(1)}} = \frac{\partial e_1}{\partial a_1^{(1)}} + \frac{\partial e_2}{\partial a_1^{(1)}}$$



$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}}$$

$$\frac{\partial E}{\partial a_1^{(1)}} = \frac{\partial e_1}{\partial a_1^{(1)}} + \frac{\partial e_2}{\partial a_1^{(1)}}$$

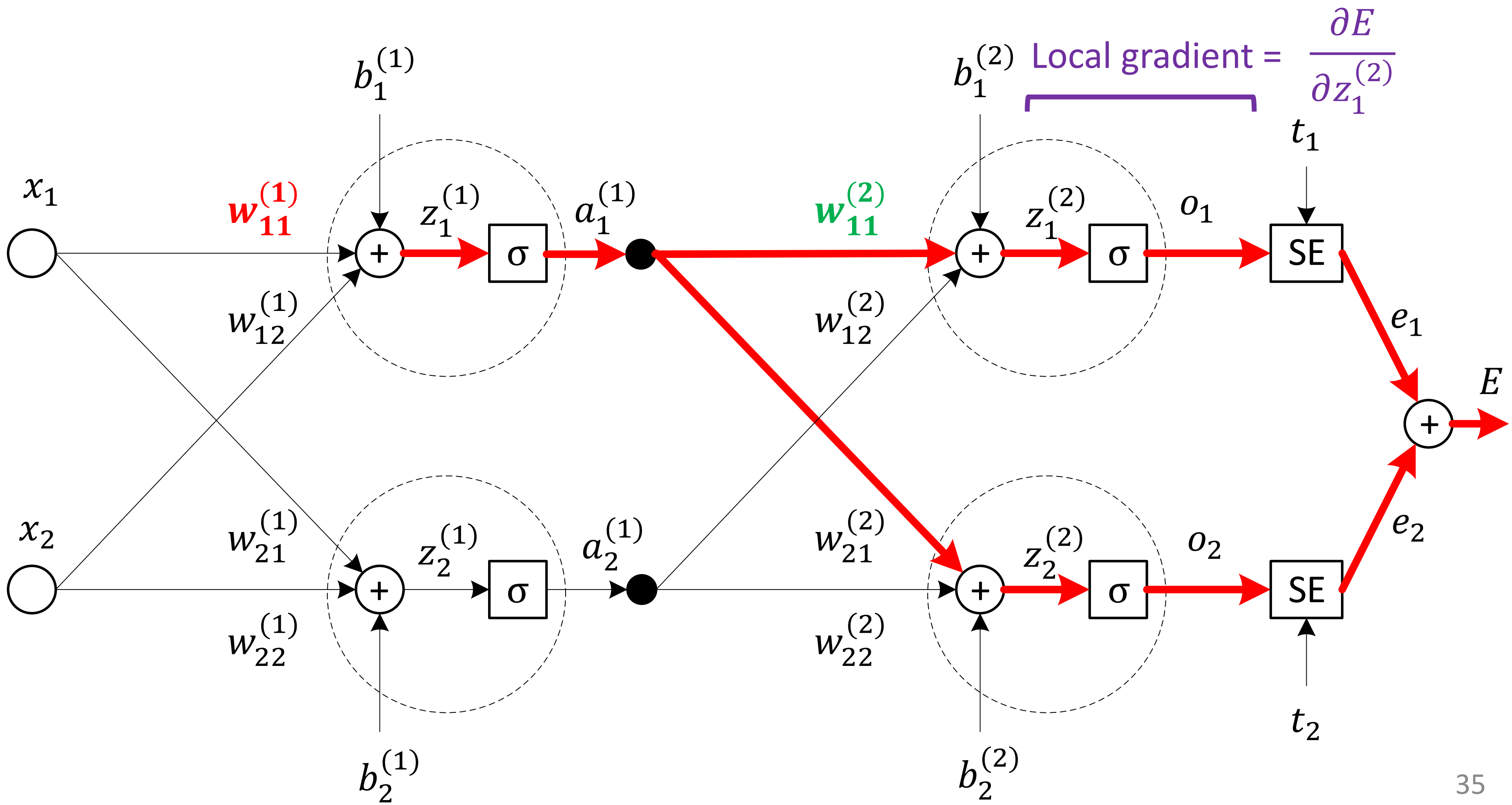
$$\frac{\partial e_1}{\partial a_1^{(1)}} = \frac{\partial e_1}{\partial z_1^{(2)}} \frac{\partial z_1^{(2)}}{\partial a_1^{(1)}}$$



$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}}$$

$$\frac{\partial E}{\partial a_1^{(1)}} = \frac{\partial e_1}{\partial a_1^{(1)}} + \frac{\partial e_2}{\partial a_1^{(1)}}$$

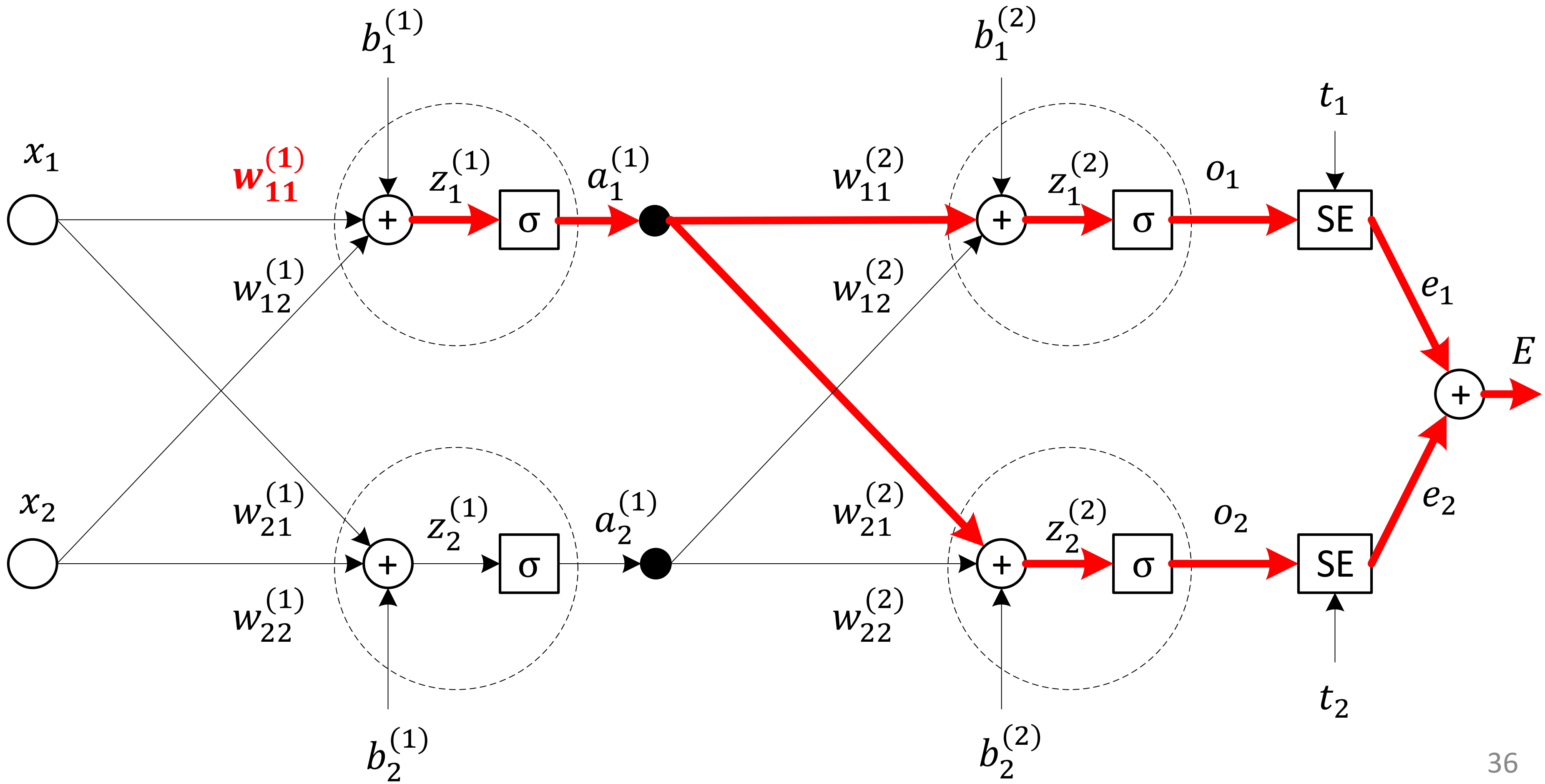
$$\frac{\partial e_1}{\partial a_1^{(1)}} = \frac{\partial e_1}{\partial z_1^{(2)}} \frac{\partial z_1^{(2)}}{\partial a_1^{(1)}} = \frac{\partial E}{\partial z_1^{(2)}} w_{11}^{(2)}$$



$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}}$$

$$\frac{\partial E}{\partial a_1^{(1)}} = \frac{\partial e_1}{\partial a_1^{(1)}} + \frac{\partial e_2}{\partial a_1^{(1)}}$$

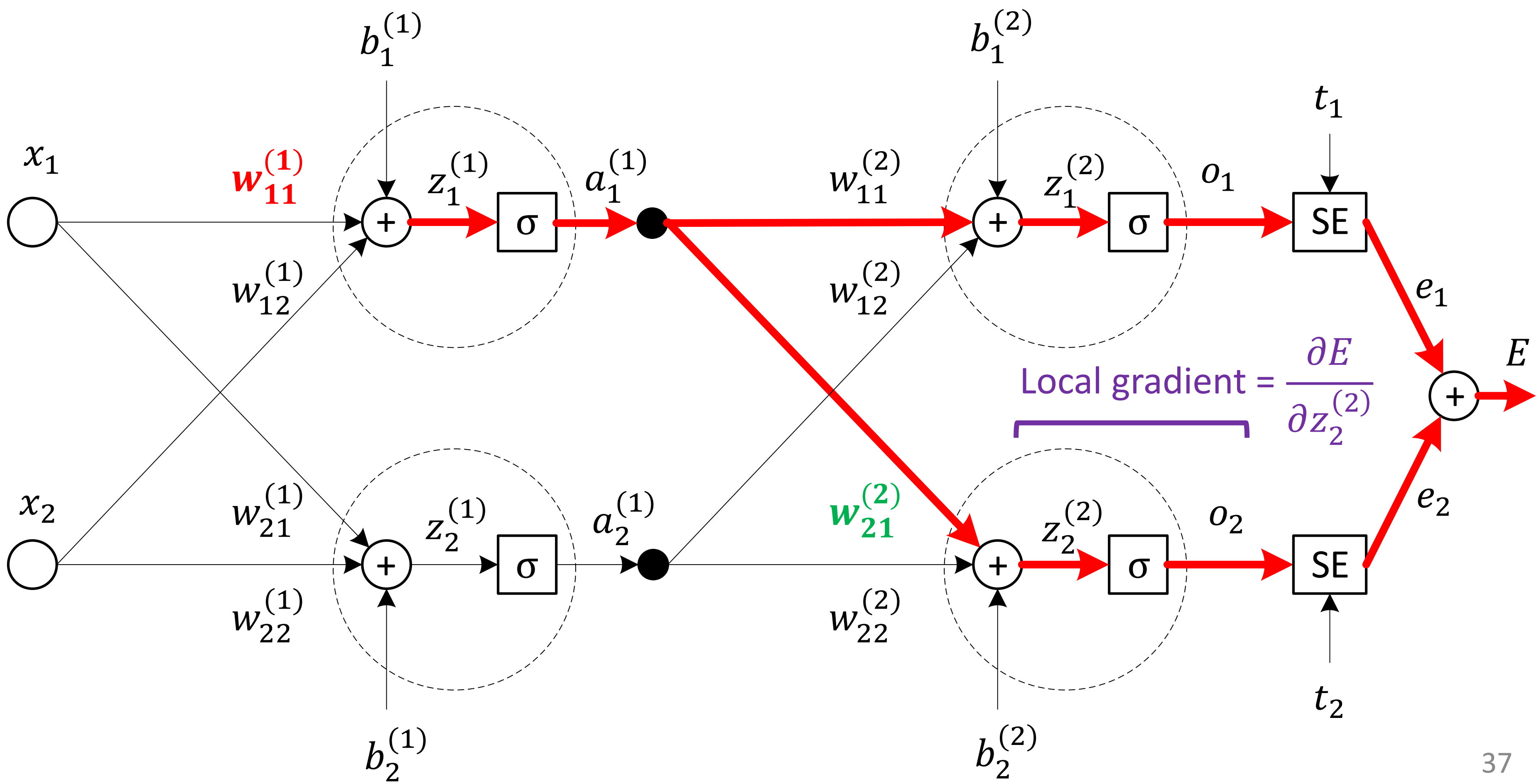
$$\frac{\partial e_2}{\partial a_1^{(1)}} = \frac{\partial e_2}{\partial z_2^{(2)}} \frac{\partial z_2^{(2)}}{\partial a_1^{(1)}}$$



$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}}$$

$$\frac{\partial E}{\partial a_1^{(1)}} = \frac{\partial e_1}{\partial a_1^{(1)}} + \frac{\partial e_2}{\partial a_1^{(1)}}$$

$$\frac{\partial e_2}{\partial a_1^{(1)}} = \frac{\partial e_2}{\partial z_2^{(2)}} \frac{\partial z_2^{(2)}}{\partial a_1^{(1)}} = \frac{\partial E}{\partial z_2^{(2)}} w_{21}^{(2)}$$

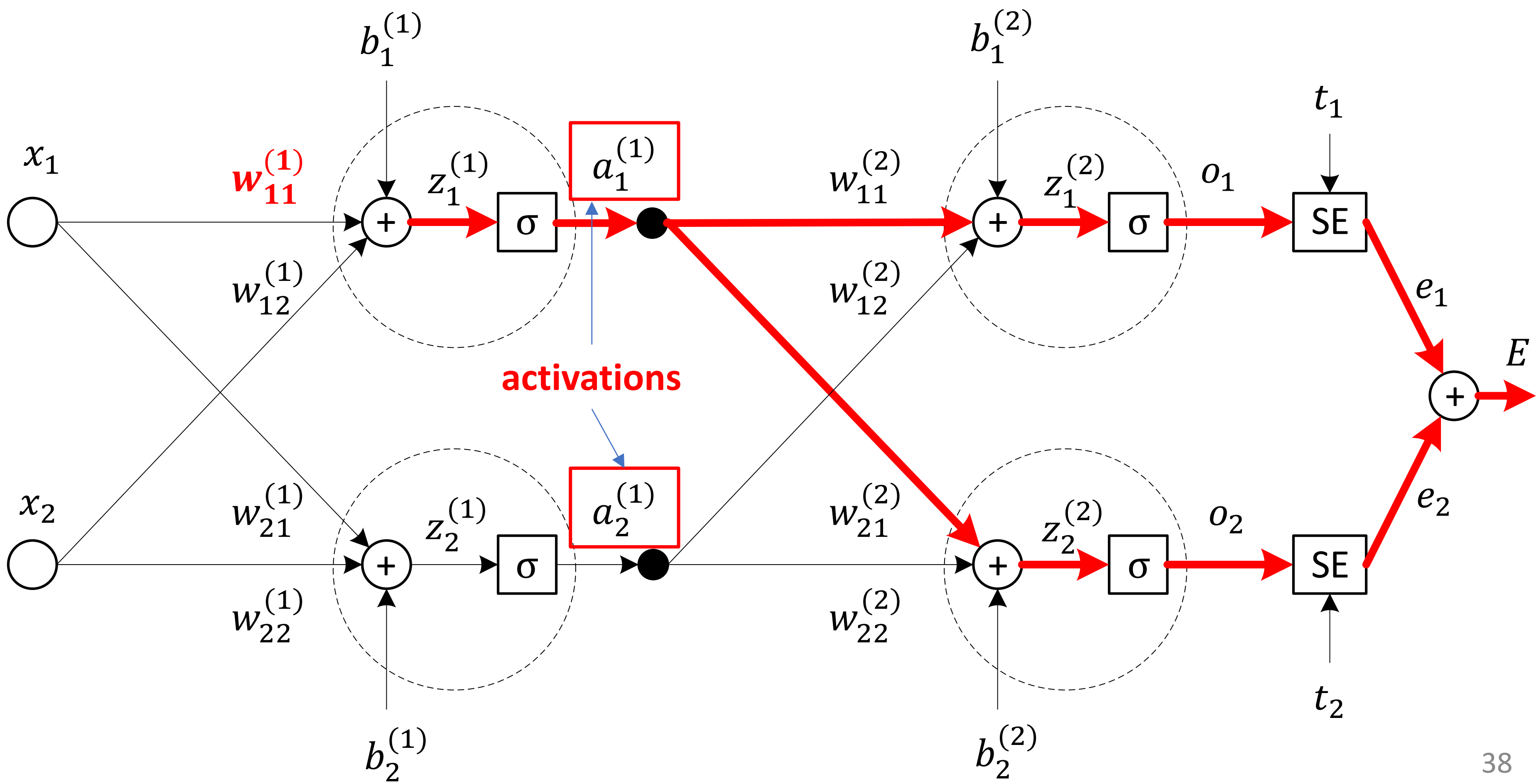


$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}}$$

$$a_1^{(1)} = \sigma\left(z_1^{(1)}\right)$$

Храним активации (а $z_j^{(i)}$ – не храним)

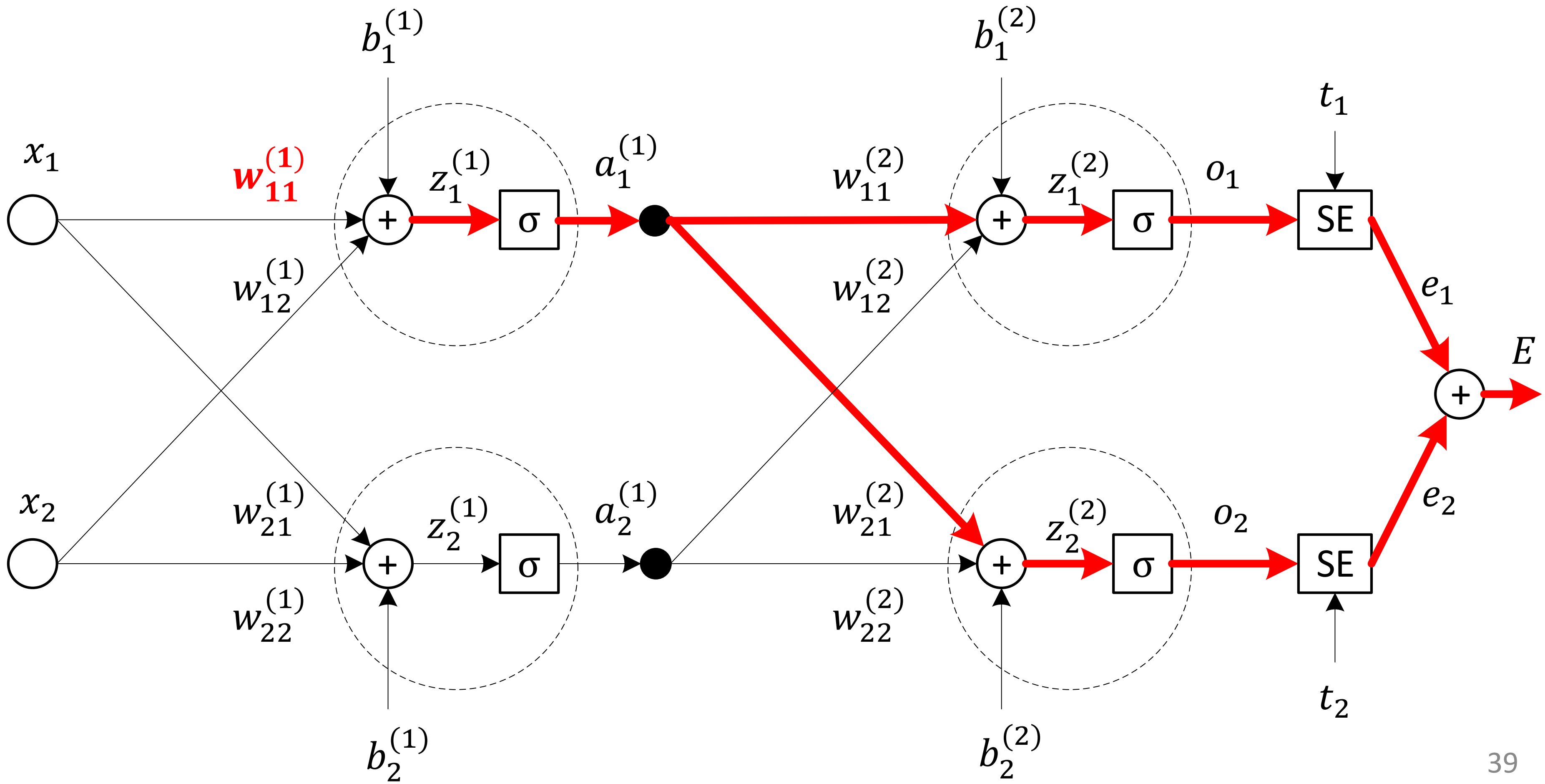
$$\frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} = a_1^{(1)}(1 - a_1^{(1)})$$



$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}}$$

$$z_1^{(1)} = w_{11}^{(2)} x_1 + w_{12}^{(2)} x_2 + b_1^{(1)}$$

$$\frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}} = x_1$$



Градиент скрытого слоя

$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}} = \left(\frac{\partial e_1}{\partial a_1^{(1)}} + \frac{\partial e_2}{\partial a_1^{(1)}}\right) \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}} =$$

Градиент скрытого слоя

$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}} = \left(\frac{\partial e_1}{\partial a_1^{(1)}} + \frac{\partial e_2}{\partial a_1^{(1)}} \right) \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}} =$$

$$= \left(\frac{\partial E}{\partial z_1^{(2)}} w_{11}^{(2)} + \frac{\partial E}{\partial z_2^{(2)}} w_{21}^{(2)} \right) \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}} =$$

Градиент скрытого слоя

$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}} = \left(\frac{\partial e_1}{\partial a_1^{(1)}} + \frac{\partial e_2}{\partial a_1^{(1)}} \right) \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}} =$$

$$= \left(\frac{\partial E}{\partial z_1^{(2)}} w_{11}^{(2)} + \frac{\partial E}{\partial z_2^{(2)}} w_{21}^{(2)} \right) \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}} =$$

$$= \left(\frac{\partial E}{\partial z_1^{(2)}} w_{11}^{(2)} + \frac{\partial E}{\partial z_2^{(2)}} w_{21}^{(2)} \right) a_1^{(1)} \left(1 - a_1^{(1)} \right) x_1$$

Градиент скрытого слоя

$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}} = \left(\frac{\partial E}{\partial z_1^{(2)}} w_{11}^{(2)} + \frac{\partial E}{\partial z_2^{(2)}} w_{21}^{(2)}\right) a_1^{(1)} \left(1 - a_1^{(1)}\right) x_1$$

$$\frac{\partial E}{\partial w_{12}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{12}^{(1)}} = \left(\frac{\partial E}{\partial z_1^{(2)}} w_{11}^{(2)} + \frac{\partial E}{\partial z_2^{(2)}} w_{21}^{(2)}\right) a_1^{(1)} \left(1 - a_1^{(1)}\right) x_2$$

$$\frac{\partial E}{\partial b_1^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial b_1^{(1)}} = \left(\frac{\partial E}{\partial z_1^{(2)}} w_{11}^{(2)} + \frac{\partial E}{\partial z_2^{(2)}} w_{21}^{(2)}\right) a_1^{(1)} \left(1 - a_1^{(1)}\right)$$

Градиент скрытого слоя

$$\frac{\partial E}{\partial w_{11}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}} = \left(\frac{\partial E}{\partial z_1^{(2)}} w_{11}^{(2)} + \frac{\partial E}{\partial z_2^{(2)}} w_{21}^{(2)} \right) a_1^{(1)} \left(1 - a_1^{(1)}\right) x_1$$

$$\frac{\partial E}{\partial w_{12}^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial w_{12}^{(1)}} = \left(\frac{\partial E}{\partial z_1^{(2)}} w_{11}^{(2)} + \frac{\partial E}{\partial z_2^{(2)}} w_{21}^{(2)} \right) a_1^{(1)} \left(1 - a_1^{(1)}\right) x_2$$

$$\frac{\partial E}{\partial b_1^{(1)}} = \frac{\partial E}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial b_1^{(1)}} = \left(\frac{\partial E}{\partial z_1^{(2)}} w_{11}^{(2)} + \frac{\partial E}{\partial z_2^{(2)}} w_{21}^{(2)} \right) a_1^{(1)} \left(1 - a_1^{(1)}\right)$$

$$\delta_1^{(1)} = \frac{\partial E}{\partial z_1^{(1)}}$$

Local gradient (или *ошибка нейрона*)

(вычисляем для каждого нейрона однократно и сохраняем)

Алгоритм обратного распространения ошибки

1. Инициализация весов
2. Цикл по эпохам:
   - Цикл по батчам:
     - Forward propagation для каждого примера из батча
     - Вычисление ошибки по батчу
     - Back propagation:

$$\delta_i^{(l)} = \begin{cases} -e_i^{(L)} \sigma_i' \left(z_i^{(L)} \right) = -(t_i - o_i) o_i (1 - o_i) & \text{– для выходного слоя } L \\ \sigma_i' \left(z_i^{(L)} \right) \sum\limits_{k \in Children(i)} \delta_k^{(l+1)} w_{ki}^{(l+1)} & \text{– для других слоев} \end{cases}$$

Алгоритм обратного распространения ошибки

- Обновление весов:

$$w_{ij}^{(l)}(n+1) = w_{ij}^{(l)}(n) - \eta\delta_i^{(l)}a_j^{(l-1)},$$

$$b_i^{(l)}(n+1) = b_i^{(l)}(n) - \eta\delta_i^{(l)},$$

где n – номер итерации,

$a_j^{(l-1)}$ – активация на предыдущем слое; если $l = 1$, то $a_j^0 = x_j$

- Критерии останова:
  - $\|\nabla E\| < \varepsilon$ – в точке минимума градиент близок к нулю
  - $|E(n+1) - E(n)| < \varepsilon$ – ошибка перестает изменяться

Алгоритм обратного распространения ошибки

1. Инициализация весов
2. Цикл по эпохам: **for** epoch **in** range(num_epochs):
   - Цикл по батчам: **for** i, (images, labels) in **enumerate**(train_loader):
     - Forward propagation для каждого примера из батча

       outputs = model(images)
     - Вычисление ошибки по батчу loss = loss_fn(outputs, labels)
     - Back propagation: optimizer.zero_grad()

       loss.backward()

$$\delta_i^{(l)} = \begin{cases} -e_i^{(L)} \sigma_i' \left(z_i^{(L)} \right) = -(t_i - o_i) o_i (1 - o_i) \\ \sigma_i' \left(z_i^{(L)} \right) \sum\limits_{k \in Children(i)} \delta_k^{(l+1)} w_{ki}^{(l+1)} \end{cases}$$

     - Обновление весов: optimizer.step()

$$w_{ij}^{(l)}(n+1) = w_{ij}^{(l)}(n) + \eta \delta_i^{(l)} a_j^{(l-1)},$$

$$b_i^{(l)}(n+1) = b_i^{(l)}(n) + \eta \delta_i^{(l)}$$

Ссылки

- 3Blue1Brown – Что на самом деле делает обратное распространение ошибки?
  - https://www.youtube.com/watch?v=Ilg3gGewQ5U
  - https://www.youtube.com/watch?v=tIeHLnjs5U8
- Matt Mazur – A Step by Step Backpropagation Example
  - https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example
- Jason Brownlee – How to Code a Neural Network with Backpropagation In Python (from scratch)
  - https://machinelearningmastery.com/implement-backpropagation-algorithm-scratch-python/
- Neural Networks and Deep Learning – How the backpropagation algorithm works
  - http://neuralnetworksanddeeplearning.com/chap2.html