Consider a hidden layer of a neural network. The input of the layer is a 4-D vector. The layer has 16 neurons. What is the size of the weight matrix $\boldsymbol{W}$ for this layer? Assume we have $\mathbf{a} = g(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x} + \boldsymbol{b})$.

A. 4 x 4

B. 4 x 16

C. 16 x 16

D. 16 x 4

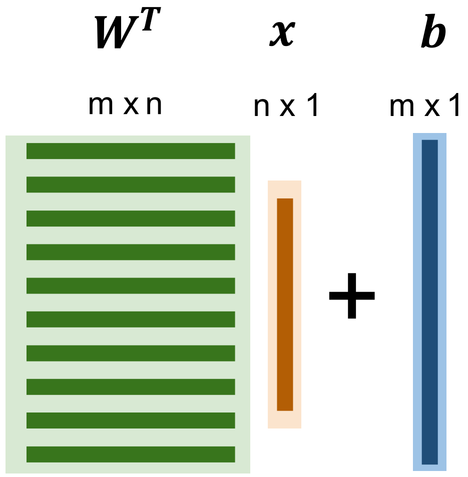
Consider a hidden layer of a neural network. The input of the layer is a 4-D vector. The layer has 16 neurons. What is the size of the weight matrix $\boldsymbol{W}$ for this layer? Assume we have $\mathbf{a} = g(\boldsymbol{W}^T\boldsymbol{x} + \boldsymbol{b})$.

A. 4 x 4

B. 4 x 16

C. 16 x 16

D. 16 x 4



Let $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Which of following functions is NOT an element-wise operation that can be used as an activation function?

A. $\mathrm{f}(\boldsymbol{x}) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

B. $\mathrm{f}(\boldsymbol{x}) = \begin{bmatrix} \max(0, x_1) \\ \max(0, x_2) \end{bmatrix}$

C. $\mathrm{f}(\boldsymbol{x}) = \begin{bmatrix} \exp(x_1) \\ \exp(x_2) \end{bmatrix}$

D. $\mathrm{f}(\boldsymbol{x}) = \begin{bmatrix} \exp(x_1 + x_2) \\ \exp(x_2) \end{bmatrix}$

Let $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Which of following functions is NOT an element-wise operation that can be used as an activation function?

A. $\mathrm{f}(\boldsymbol{x}) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

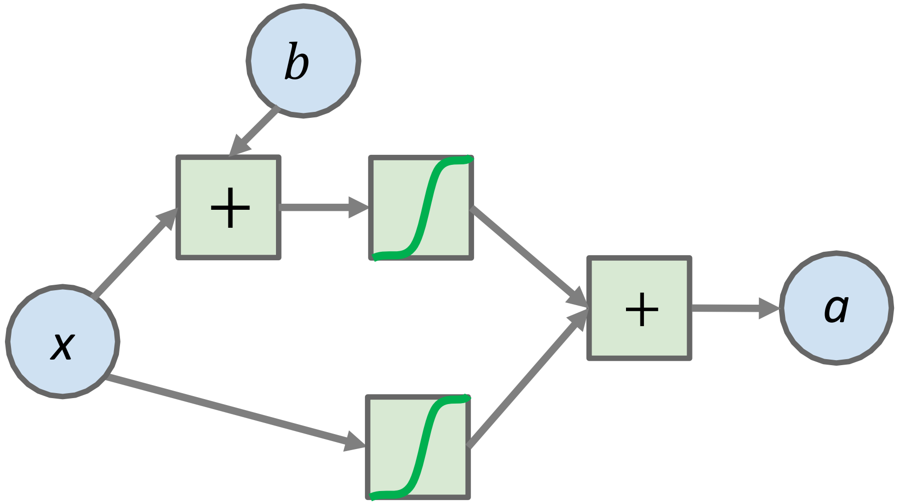
B. $\mathrm{f}(\boldsymbol{x}) = \begin{bmatrix} \max(0, x_1) \\ \max(0, x_2) \end{bmatrix}$

C. $\mathrm{f}(\boldsymbol{x}) = \begin{bmatrix} \exp(x_1) \\ \exp(x_2) \end{bmatrix}$

D. $\mathbf{f}(\boldsymbol{x}) = \begin{bmatrix} \mathbf{exp}(\boldsymbol{x}_1 + \boldsymbol{x}_2) \\ \mathbf{exp}(\boldsymbol{x}_2) \end{bmatrix}$

This is not an element-wise operation as the first output depends on both input values.

Consider the following computational graph. Which function does it represent? Assuming a sigmoid activation function.



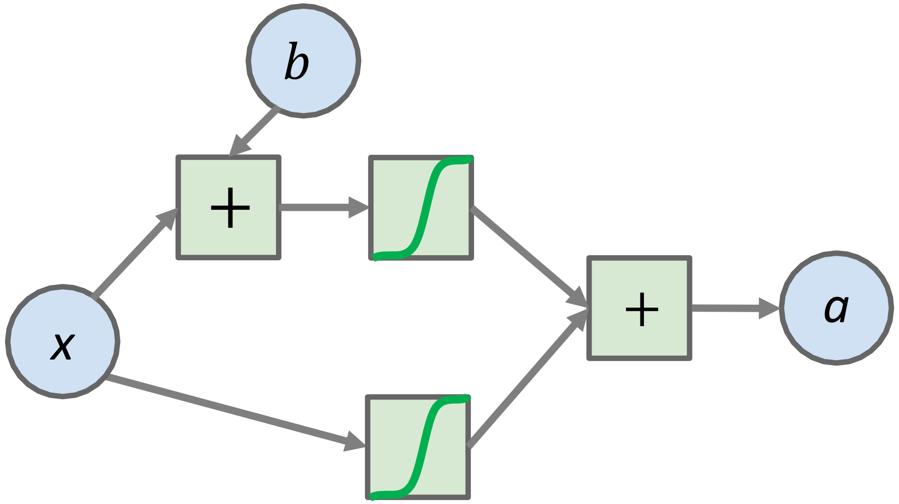
A. $\mathrm{sigmoid}(x + b)$

B. $\mathrm{sigmoid}(x)$

C. $\mathrm{sigmoid}(x + b) + \mathrm{sigmoid}(x)$

D. $x + b$

Consider the following computational graph. Which function does it represent? Assuming a sigmoid activation function.



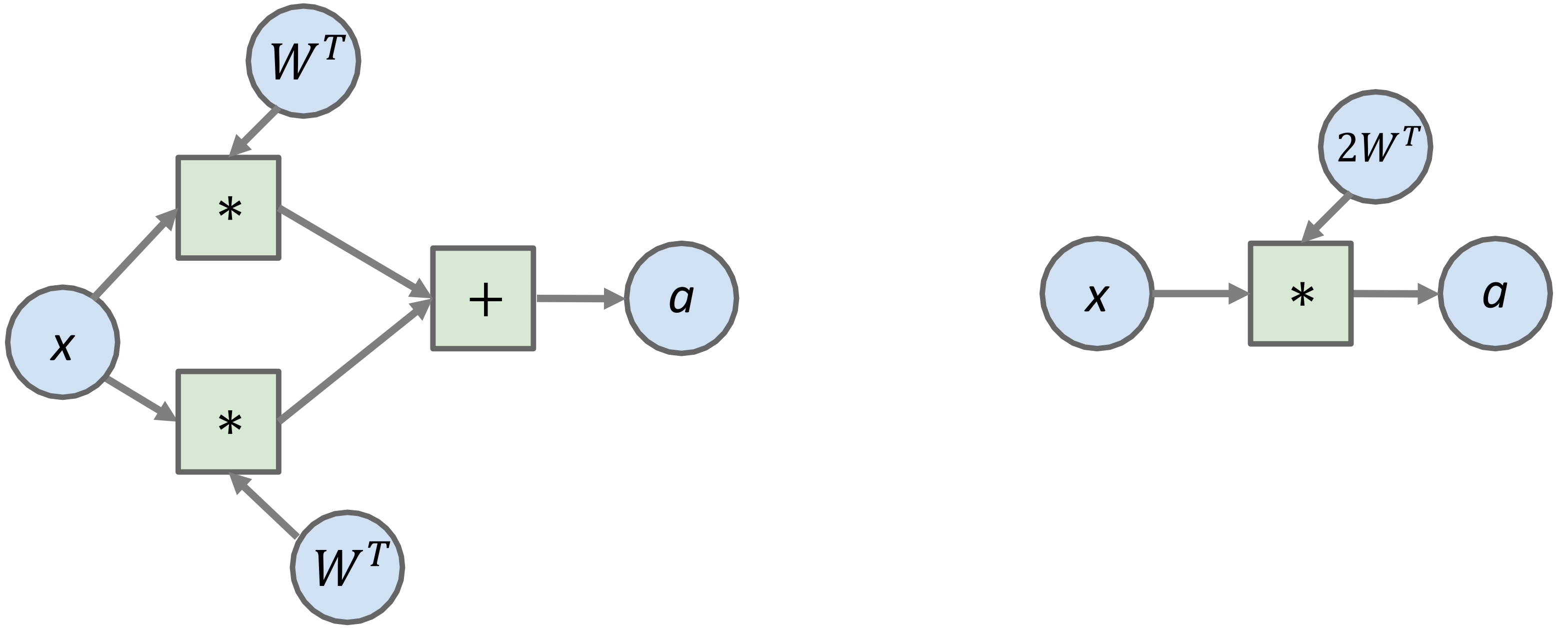
A. $\mathrm{sigmoid}(x + b)$

B. $\mathrm{sigmoid}(x)$

C. $\mathrm{sigmoid}(x + b) + \mathrm{sigmoid}(x)$

D. $x + b$

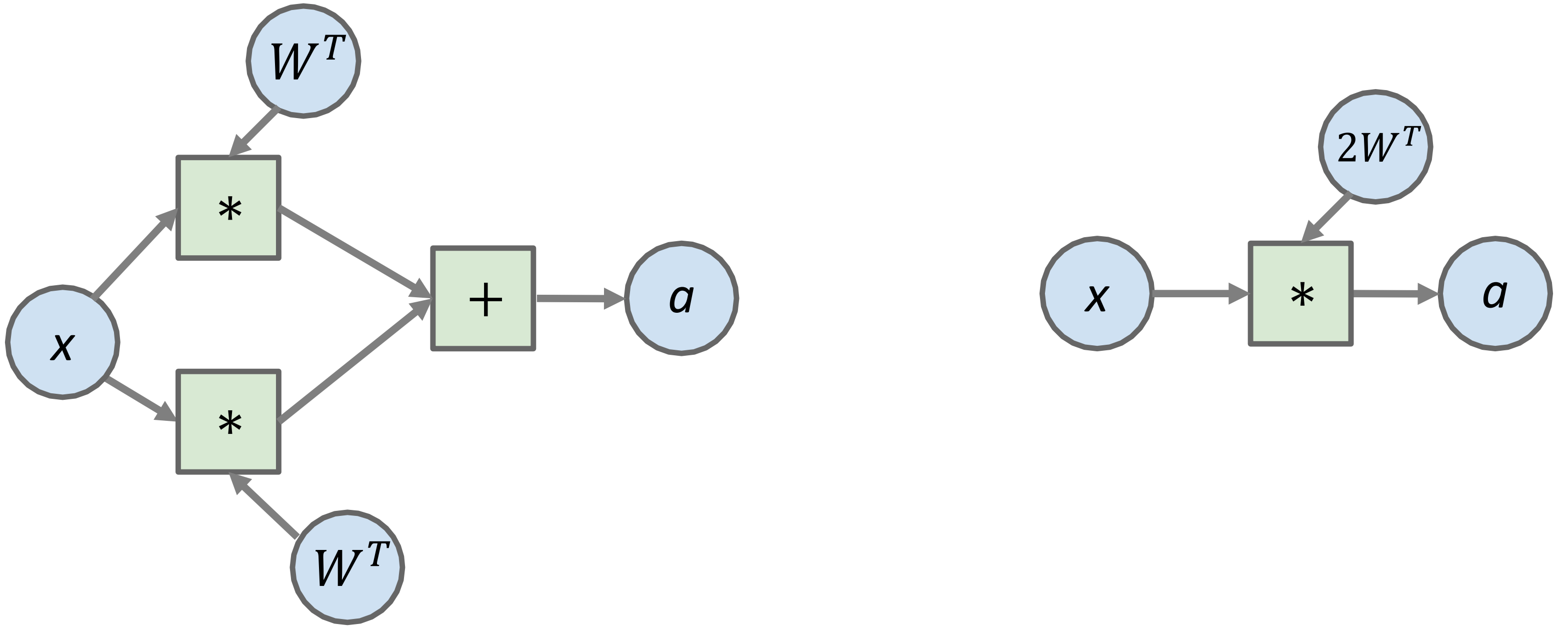
Consider the following two computational graphs. Do they represent the same function?



Yes

No

Consider the following two computational graphs. Do they represent the same function?



Yes. The first graph is $W^T x + W^T x$ and the second graph is $2W^T x$.

Let $f(x) = \begin{cases} -1 & x < 0.5 \\ 1 & x \geq 0.5 \end{cases}$. Can we use this function as an operation on a computational graph that supports backward propagation?

Yes

No

Let $f(x) = \begin{cases} -1 & x < 0.5 \\ 1 & x \geq 0.5 \end{cases}$. Can we use this function as an operation on a computational graph that supports backward propagation?

No. The function is not continuous and not differentiable when x=0.5.

Let $f(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}$. Can we use this function as an operation on a computational graph that supports backward propagation? Assume that we define the "gradient" $f'(0) = 0$.

Yes

No

Let $f(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}$. Can we use this function as an operation on a computational graph that supports backward propagation? Assume that we define the “gradient” $f'(0) = 0$.

Yes. The function is continuous but not differentiable at 0. With the patch, we can compute a “gradient” (known as sub-gradient) for this function and thus use this function as an operation on the graph.