Due Date : February 16th, 2019

Instructions

- *For all questions, show your work!*
- *Use a document preparation system such as LaTeX.*
- *Submit your answers electronically via Gradescope.*

Question 1 (4-4-4-2)**.** Using the following definition of the derivative and the definition of the Heaviside step function :

$$\frac{d}{dx}f(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \qquad H(x) = \begin{cases} 1 & \text{if } x > 0 \\ \frac{1}{2} & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases}$$

1. Show that the derivative of the rectified linear unit $g(x) = \max\{0, x\}$, **wherever it exists**, is equal to the Heaviside step function.
2. Give two alternative definitions of $g(x)$ using $H(x)$.
3. Show that $H(x)$ can be well approximated by the sigmoid function $\sigma(x) = \frac{1}{1+e^{-kx}}$ asymptotically (i.e for large k), where k is a parameter.

*4. Although the Heaviside step function is not differentiable, we can define its **distributional derivative**. For a function F, consider the functional $F[\phi] = \int_{\mathbb{R}} F(x)\phi(x)dx$, where ϕ is a smooth function (infinitely differentiable) with compact support ($\phi(x) = 0$ whenever $|x| \geq A$, for some $A > 0$).

Show that whenever F is differentiable, $F'[\phi] = -\int_{\mathbb{R}} F(x)\phi'(x)dx$. Using this formula as a definition in the case of non-differentiable functions, show that $H'[\phi] = \phi(0)$. ($\delta[\phi] \doteq \phi(0)$ is known as the Dirac delta function.)

Answer 1.

1. (a) For $x < 0$, when $|\epsilon| < -x$, $g(x) = g(x+\epsilon) = 0$, so

$$\lim_{\epsilon \to 0} \frac{g(x+\epsilon) - g(x)}{\epsilon} = \lim_{\epsilon \to 0} \frac{0-0}{\epsilon} = 0$$

   (b) For $x > 0$, when $|\epsilon| < x$, $g(x+\epsilon) = x + \epsilon$ and $g(x) = x$, so

$$\lim_{\epsilon \to 0} \frac{g(x+\epsilon) - g(x)}{\epsilon} = \lim_{\epsilon \to 0} \frac{x+\epsilon-x}{\epsilon} = 1$$

   (c) For $x = 0$,

$$\lim_{\epsilon \to 0^-} \frac{g(x+\epsilon) - g(x)}{\epsilon} = \lim_{\epsilon \to 0^-} \frac{0-0}{\epsilon} = 0$$

$$\lim_{\epsilon \to 0^+} \frac{g(x+\epsilon) - g(x)}{\epsilon} = \lim_{\epsilon \to 0^+} \frac{\epsilon-0}{\epsilon} = 1$$

   so g is not differentiable at 0.

2. $g(x) = xH(x) = \int_{-\infty}^{x} H(t)dt$

3. Taking $k \to \infty$, we have

$$\lim_{k\to\infty} \sigma(x) = \frac{1}{1+e^{-kx}} = \begin{cases} 1 & \text{if } x > 0 \\ \frac{1}{2} & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases}$$

4. By *integration by parts*,

$$\begin{aligned} F'[\phi] &= F(x)\phi(x)\Big|_{-\infty}^{\infty} - \int_{\mathbb{R}} F(x)\phi'(x)dx \\ &= 0 - 0 - \int_{\mathbb{R}} F(x)\phi'(x)dx \end{aligned}$$

In the case of the Heaviside step function,

$$\begin{aligned} H'[\phi] &\doteq -\int_{-\infty}^{\infty} H(x)\phi'(x) \\ &= -\int_{0}^{\infty} \phi'(x) \\ &= -\phi(x)\Big|_{0}^{\infty} = 0 + \phi(0) = \phi(0) \end{aligned}$$

Question 2 (5-8-5-5)**.** Let x be an n-dimensional vector. Recall the softmax function : $S : \boldsymbol{x} \in \mathbb{R}^n \mapsto S(\boldsymbol{x}) \in \mathbb{R}^n$ such that $S(\boldsymbol{x})_i = \frac{e^{\boldsymbol{x}_i}}{\sum_j e^{\boldsymbol{x}_j}}$; the diagonal function : $\text{diag}(\boldsymbol{x})_{ij} = \boldsymbol{x}_i$ if $i = j$ and $\text{diag}(\boldsymbol{x})_{ij} = 0$ if $i \neq j$; and the Kronecker delta function : $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

1. Show that the derivative of the softmax function is $\frac{dS(\boldsymbol{x})_i}{d\boldsymbol{x}_j} = S(\boldsymbol{x})_i\,(\delta_{ij} - S(\boldsymbol{x})_j)$.
2. Express the Jacobian matrix $\frac{\partial S(\boldsymbol{x})}{\partial \boldsymbol{x}}$ using matrix-vector notation. Use $\text{diag}(\cdot)$.
3. Compute the Jacobian of the sigmoid function $\sigma(\boldsymbol{x}) = 1/(1+e^{-\boldsymbol{x}})$.
4. Let $\boldsymbol{y}$ and $\boldsymbol{x}$ be $n-$dimensional vectors related by $\boldsymbol{y} = f(\boldsymbol{x})$, L be an unspecified differentiable loss function. According to the chain rule of calculus, $\nabla_{\boldsymbol{x}} L = (\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}})^\top \nabla_{\boldsymbol{y}} L$, which takes up $\mathcal{O}(n^2)$ computational time in general. Show that if $f(\boldsymbol{x}) = \sigma(\boldsymbol{x})$ or $f(\boldsymbol{x}) = S(\boldsymbol{x})$, the above matrix-vector multiplication can be simplified to a $\mathcal{O}(n)$ operation.

Answer 2.

1. Note that $\log S(\boldsymbol{x})_i = \boldsymbol{x}_i - \log \sum_{j'} e^{\boldsymbol{x}_{j'}}$. We can rewrite the gradient as

$$\begin{aligned} \frac{d \log S(\boldsymbol{x})_i}{dx_j} &= \frac{d \log S(\boldsymbol{x})_i}{dS(\boldsymbol{x})_i} \frac{dS(\boldsymbol{x})_i}{d\boldsymbol{x}_j} \\ \delta_{ij} - \frac{e^{\boldsymbol{x}_j}}{\sum_{j'} e^{\boldsymbol{x}_{j'}}} &= \frac{1}{S(\boldsymbol{x})_i} \frac{dS(\boldsymbol{x})_i}{d\boldsymbol{x}_j} \end{aligned}$$

Rearranging the terms yields $\frac{dS(\boldsymbol{x})_i}{d\boldsymbol{x}_j} = S(\boldsymbol{x})_i\,(\delta_{ij} - S(\boldsymbol{x})_j)$.

2. From the last question, we have $\frac{\partial S(\boldsymbol{x})}{\partial \boldsymbol{x}} = \text{diag}(S(\boldsymbol{x})) - S(\boldsymbol{x})S(\boldsymbol{x})^\top$.
3. For $i \neq j$, $\frac{d\sigma(\boldsymbol{x})_i}{d\boldsymbol{x}_j} = 0$. On the diagonal, we have $\frac{d\sigma(\boldsymbol{x}_i)}{d\boldsymbol{x}_i} = \sigma(\boldsymbol{x}_i)(1-\sigma(\boldsymbol{x}_i))$. Thus, $\frac{\partial \sigma(\boldsymbol{x})}{\partial \boldsymbol{x}} = \text{diag}(\sigma(\boldsymbol{x})(1-\sigma(\boldsymbol{x})))$.

4. For $f = \sigma$, since the Jacobian is a diagonal matrix $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \text{diag}(\sigma(\boldsymbol{x})(1 - \sigma(\boldsymbol{x})))$,

$$(\nabla_{\boldsymbol{x}} L)_i = \left(\left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \right)^\top \nabla_{\boldsymbol{y}} L \right)_i = \sum_j \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \right)_{ji} (\nabla_{\boldsymbol{y}} L)_j = \sigma(\boldsymbol{x}_i)(1 - \sigma(\boldsymbol{x}_i)) \, (\nabla_{\boldsymbol{y}} L)_i$$

So the dot product can be replaced with an elementwise product. If $f = S$, the gradient can be written as

$$\nabla_x L = (\text{diag}(S(\boldsymbol{x})) - S(\boldsymbol{x}) S(\boldsymbol{x})^\top)^\top \nabla_{\boldsymbol{y}} L = S(\boldsymbol{x}) \odot \nabla_{\boldsymbol{y}} L - (S(\boldsymbol{x})^\top \nabla_{\boldsymbol{y}} L) S(\boldsymbol{x})$$

where $\odot$ is Hadamard product. Both terms can be computed in $\mathcal{O}(n)$ time.

Question 3 (3-3-3-3)**.** Recall the definition of the softmax function : $S(\boldsymbol{x})_i = e^{\boldsymbol{x}_i} / \sum_j e^{\boldsymbol{x}_j}$.

1. Show that softmax is translation-invariant, that is : $S(\boldsymbol{x} + c) = S(\boldsymbol{x})$, where c is a scalar constant.
2. Show that softmax is not invariant under scalar multiplication. Let $S_c(\boldsymbol{x}) = S(c\boldsymbol{x})$ where $c \geq 0$. What are the effects of taking c to be 0 and arbitrarily large ?
3. Let $\boldsymbol{x}$ be a 2-dimensional vector. One can represent a 2-class categorical probability using softmax $S(\boldsymbol{x})$. Show that $S(\boldsymbol{x})$ can be reparameterized using sigmoid function, i.e. $S(\boldsymbol{x}) = [\sigma(z), 1 - \sigma(z)]^\top$ where z is a scalar function of $\boldsymbol{x}$.
4. Let $\boldsymbol{x}$ be a K-dimensional vector ($K \geq 2$). Show that $S(\boldsymbol{x})$ can be represented using $K - 1$ parameters, i.e. $S(\boldsymbol{x}) = S([0, y_1, y_2, ..., y_{K-1}]^\top)$ where y_i is a scalar function of $\boldsymbol{x}$ for $i \in \{1, ..., K - 1\}$.

Answer 3.

1. Let K be the dimensionality of $\boldsymbol{x}$. For $i \in \{1, ..., K\}$,

$$\begin{aligned} S(\boldsymbol{x} + c)_i &= \frac{\exp(x_i + c)}{\sum_{j=1}^K \exp(x_k + c)} \\ &= \frac{\exp(c) \exp(x_i)}{\sum_{j=1}^K \exp(c) \exp(x_k)} \\ &= \frac{\exp(c) \exp(x_i)}{\exp(c) \sum_{j=1}^K \exp(x_k)} \\ &= S(\boldsymbol{x})_i \end{aligned}$$

2. First, $\boldsymbol{x} = [0, \log 2]^\top$ and $c = 2$ is a counterexample, since $S(\boldsymbol{x})_1 = \frac{1}{1+2}$ whereas $S(c\boldsymbol{x})_1 = \frac{1}{1+4}$. Second, setting c to be 0 yields a uniform output, as $S(0\boldsymbol{x})_i = \frac{1}{K}$ for all $i \in \{0, ..., K\}$. Lastly, assume all elements of $\boldsymbol{x}$ are distinct. Since softmax is translation-invariant, one can subtract cx^* from the exponent, where $x^* = ||\boldsymbol{x}||_\infty$ is the maximum norm :

$$S(c\boldsymbol{x} - cx^*)_i = \frac{\exp(cx_i - cx^*)}{\sum_{j=1}^K \exp(cx_j - cx^*)} \xrightarrow{c \to \infty} \begin{cases} 1 & \text{if } x_i = x^* \\ 0 & \text{if } x_i < x^* \end{cases}$$

When the norm of the input of softmax is finite, it behaves like a soft version of the arg max operator.

3. Let $z = x_1 - x_2$.

$$\sigma(z) = \frac{1}{1+\exp(-x_1+x_2)} = \frac{\exp(x_1)}{\exp(x_1)+\exp(x_2)} = S(\boldsymbol{x})_1$$

   Similarly, $1-\sigma(z) = S(\boldsymbol{x})_2$.

4. For $i \in \{1, ..., K-1\}$, let $y_i = x_{i+1} - x_1$. By the translation-invariance,

$$S(\boldsymbol{x}) = S(\boldsymbol{x} - x_1) = S([0, x_2 - x_1, x_3 - x_1 ..., x_K - x_1]^\top) = S([0, y_1, ..., y_{K-1}]^\top)$$

Question 4 (15)**.** Consider a 2-layer neural network $y : \mathbb{R}^D \to \mathbb{R}^K$ of the form :

$$y(x, \Theta, \sigma)_k = \sum_{j=1}^{M} \omega_{kj}^{(2)} \sigma\left(\sum_{i=1}^{D} \omega_{ji}^{(1)} x_i + \omega_{j0}^{(1)}\right) + \omega_{k0}^{(2)}$$

for $1 \leq k \leq K$, with parameters $\Theta = (\omega^{(1)}, \omega^{(2)})$ and logistic sigmoid activation function σ. Show that there exists an equivalent network of the same form, with parameters $\Theta' = (\tilde{\omega}^{(1)}, \tilde{\omega}^{(2)})$ and tanh activation function, such that $y(x, \Theta', \tanh) = y(x, \Theta, \sigma)$ for all $x \in \mathbb{R}^D$, and express Θ' as a function of Θ.

Answer 4. First since $\tanh(x) = 2\sigma(2x) - 1$, we have $\sigma(x) = \frac{1}{2}\left(\tanh(\frac{x}{2}) + 1\right)$. Thus,

$$\begin{aligned}
y(x, \Theta, \sigma)_k &= \sum_{j=1}^{M} \omega_{kj}^{(2)} \sigma\left(\sum_{i=1}^{D} \omega_{ji}^{(1)} x_i + \omega_{j0}^{(1)}\right) + \omega_{k0}^{(2)} \\
&= \sum_{j=1}^{M} \omega_{kj}^{(2)} \cdot \frac{1}{2}\left(1 + \tanh\left(\sum_{i=1}^{D} \frac{\omega_{ji}^{(1)}}{2} x_i + \frac{\omega_{j0}^{(1)}}{2}\right)\right) + \omega_{k0}^{(2)} \\
&= \sum_{j=1}^{M} \frac{\omega_{kj}^{(2)}}{2} \tanh\left(\sum_{i=1}^{D} \frac{\omega_{ji}^{(1)}}{2} x_i + \frac{\omega_{j0}^{(1)}}{2}\right) + \left(\omega_{k0}^{(2)} + \sum_{j=1}^{M} \frac{\omega_{kj}^{(2)}}{2}\right) \\
&= \sum_{j=1}^{M} \tilde{\omega_{kj}}^{(2)} \sigma\left(\sum_{i=1}^{D} \tilde{\omega_{ji}}^{(1)} x_i + \tilde{\omega_{j0}}^{(1)}\right) + \tilde{\omega_{k0}}^{(2)} = y(x, \Theta', \tanh)_k
\end{aligned}$$

where $\tilde{\omega_{ji}}^{(1)} = \frac{\omega_{ji}^{(1)}}{2}$, $\tilde{\omega_{kj}}^{(2)} = \frac{\omega_{kj}^{(2)}}{2}$ for $j \geq 1$, and $\tilde{\omega_{k0}}^{(2)} = \left(\omega_{k0}^{(2)} + \sum_{j=1}^{M} \frac{\omega_{kj}^{(2)}}{2}\right)$.

Question 5 (2-2-2-2)**.** Given $N \in \mathbb{Z}^+$, we want to show that for any $f : \mathbb{R}^n \to \mathbb{R}^m$ and any sample set $\mathcal{S} \subset \mathbb{R}^n$ of size N, there is a set of parameters for a two-layer network such that the output $y(\boldsymbol{x})$ matches $f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{S}$. That is, we want to interpolate f with y on any finite set of samples $\mathcal{S}$.

1. Write the generic form of the function $y : \mathbb{R}^n \to \mathbb{R}^m$ defined by a 2-layer network with $N-1$ hidden units, with linear output and activation function ϕ, in terms of its weights and biases $(\boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})$ and $(\boldsymbol{W}^{(2)}, \boldsymbol{b}^{(2)})$.
2. In what follows, we will restrict $\boldsymbol{W}^{(1)}$ to be $\boldsymbol{W}^{(1)} = [\boldsymbol{w}, \cdots, \boldsymbol{w}]^T$ for some $\boldsymbol{w} \in \mathbb{R}^n$ (so the rows of $\boldsymbol{W}^{(1)}$ are all the same). Show that the interpolation problem on the sample set $\mathcal{S} = \{\boldsymbol{x}^{(1)}, \cdots \boldsymbol{x}^{(N)}\} \subset \mathbb{R}^n$ can be reduced to solving a matrix equation : $\boldsymbol{M}\tilde{\boldsymbol{W}}^{(2)} = \boldsymbol{F}$, where $\tilde{\boldsymbol{W}}^{(2)}$ and $\boldsymbol{F}$ are both $N \times m$, given by

$$\tilde{\boldsymbol{W}}^{(2)} = [\boldsymbol{W}^{(2)}, \boldsymbol{b}^{(2)}]^\top \qquad\qquad \boldsymbol{F} = [f(\boldsymbol{x}^{(1)}), \cdots, f(\boldsymbol{x}^{(N)})]^\top$$

   Express the $N \times N$ matrix $\boldsymbol{M}$ in terms of $\boldsymbol{w}$, $\boldsymbol{b}^{(1)}$, ϕ and $\boldsymbol{x}^{(i)}$.

*3. **Proof with Relu activation.** Assume $\boldsymbol{x}^{(i)}$ are all distinct. Choose $\boldsymbol{w}$ such that $\boldsymbol{w}^\top \boldsymbol{x}^{(i)}$ are also all distinct (Try to prove the existence of such a $\boldsymbol{w}$, although this is not required for the assignment - See Assignment 0). Set $\boldsymbol{b}_j^{(1)} = -\boldsymbol{w}^\top \boldsymbol{x}^{(j)} + \epsilon$, where $\epsilon > 0$. Find a value of ϵ such that $\boldsymbol{M}$ is triangular with non-zero diagonal elements. Conclude. (Hint : assume an ordering of $\boldsymbol{w}^\top \boldsymbol{x}^{(i)}$.)

*4. **Proof with sigmoid-like activations**. Assume ϕ is continuous, bounded, $\phi(-\infty) = 0$ and $\phi(0) > 0$. Decompose $\boldsymbol{w}$ as $\boldsymbol{w} = \lambda \boldsymbol{u}$. Set $\boldsymbol{b}_j^{(1)} = -\lambda \boldsymbol{u}^\top \boldsymbol{x}^{(j)}$. Fixing $\boldsymbol{u}$, show that $\lim_{\lambda \to +\infty} \boldsymbol{M}$ is triangular with non-zero diagonal elements. Conclude. (Note that doing so preserves the distinctness of $\boldsymbol{w}^\top \boldsymbol{x}^{(i)}$.)

Answer 5.

1. $\boldsymbol{W}^{(1)}$ is $N-1 \times n$, $\boldsymbol{W}^{(2)}$ is $m \times N-1$, $\boldsymbol{b}^{(1)} \in \mathbb{R}^{N-1}$ and $\boldsymbol{b}^{(2)} \in \mathbb{R}^m$. We have

$$y(\boldsymbol{x}) = \boldsymbol{W}^{(2)} \phi(\boldsymbol{W}^{(1)} \boldsymbol{x} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)}$$

2. Let the last column of $\boldsymbol{M}$ be a vector of ones (dummy variable). For $j < N$, $\boldsymbol{M}_{ij} = \phi(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + \boldsymbol{b}_j)$. Or in matrix form ($\mathbf{1}_n$ represents a vector of n ones),

$$\begin{aligned}
\boldsymbol{M} &= \begin{bmatrix} \phi(\boldsymbol{X}\boldsymbol{W}^{(1)\top} + \mathbf{1}_N \boldsymbol{b}^{(1)\top}) & \mathbf{1}_N \end{bmatrix} \\
&= \begin{bmatrix} \phi(\boldsymbol{w}^\top \boldsymbol{x}^{(1)} \mathbf{1}_{N-1} + \boldsymbol{b}^{(1)})^\top & 1 \\ \phi(\boldsymbol{w}^\top \boldsymbol{x}^{(2)} \mathbf{1}_{N-1} + \boldsymbol{b}^{(1)})^\top & 1 \\ \vdots & \vdots \\ \phi(\boldsymbol{w}^\top \boldsymbol{x}^{(3)} \mathbf{1}_{N-1} + \boldsymbol{b}^{(1)})^\top & 1 \end{bmatrix}
\end{aligned}$$

3. With the proposed form for $\boldsymbol{b}_j^{(1)}$, we have for $j < N$,

$$\boldsymbol{M}_{ij} = \max(\boldsymbol{w}^\top (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}) + \epsilon, 0)$$

According to Exercise **??**, $\boldsymbol{w}$ can be chosen such that all $\boldsymbol{w}^\top \boldsymbol{x}^{(i)}$ are distinct. Since the last column of $\boldsymbol{M}$ is a vector of ones, we assume $\boldsymbol{w}^\top \boldsymbol{x}^{(i)}$ has a decreasing order, such that if we choose ϵ such that $0 < \epsilon < \inf_{i \neq j} |\boldsymbol{w}^\top (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})|$, $\boldsymbol{M}$ is an upper triangular matrix with non-zero diagonal elements. It is thus invertible, we can solve the linear system by inverting $\boldsymbol{M}$.

4. With the proposed form for $\boldsymbol{b}_j^{(1)}$, we have for $j < N$,

$$\boldsymbol{M}_{ij} = \phi(\lambda \boldsymbol{u}^\top (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}))$$

With the same ordering assumption, $\boldsymbol{u}^\top (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}) < 0$ for $i > j$, implying that $\lim_{\lambda \to +\infty} \boldsymbol{M}$ is upper triangular, with non-zero diagonal elements (since $\phi(0) > 0$). Also, the upper right part of the limiting matrix is bounded, since ϕ is assumed to be bounded, implying invertibility. By continuity of the determinant and the mapping to adjugate, $\boldsymbol{M}$ is invertible by choosing sufficiently large λ, and thus the linear equation is solvable.

Question 6 (6)**.** Compute the *full*, *valid*, and *same* convolution (with kernel flipping) for the following 1D matrices : $\begin{bmatrix}1, 2, 3, 4\end{bmatrix} * \begin{bmatrix}1, 0, 2\end{bmatrix}$

Answer 6. Full : $\begin{bmatrix}1, 2, 5, 8, 6, 8\end{bmatrix}$; Valid : $\begin{bmatrix}5, 8\end{bmatrix}$; Same : $\begin{bmatrix}2, 5, 8, 6\end{bmatrix}$.

Question 7 (5-5)**.** Consider a convolutional neural network. Assume the input is a colorful image of size 256×256 in the RGB representation. The first layer convolves 64 8×8 kernels with the input, using a stride of 2 and no padding. The second layer downsamples the output of the first layer with a 5×5 non-overlapping max pooling. The third layer convolves 128 4×4 kernels with a stride of 1 and a zero-padding of size 1 on each border.

1. What is the dimensionality (scalar) of the output of the last layer ?
2. Not including the biases, how many parameters are needed for the last layer ?

Answer 7.

1. The output shape of a convolutional layer is
$$o = \lfloor \frac{i + 2p - k}{s} \rfloor + 1$$
where i, p, k, s are the input size, padding size, kernel size, and stride size, respectively. Initially, the input is of shape $(3, 64, 64)$. After the first layer, the representation is of shape $(64, 125, 125)$. After the second layer, the representation is of shape $(64, 25, 25)$. After the last layer, the output is of shape $(128, 24, 24)$. Thus the output has $128 \times 24 \times 24 = 73728$ dimensions.
2. $128 \times 64 \times 4 \times 4 = 131072$.

Question 8 (4-4-4)**.** Assume we are given data of size $3 \times 64 \times 64$. In what follows, provide a correct configuration of a convolutional neural network layer that satisfies the specified assumption. Answer with the window size of kernel (k), stride (s), padding (p), and dilation (d, with convention $d = 1$ for no dilation). Use square windows only (e.g. same k for both width and height).

1. The output shape of the first layer is $(64, 32, 32)$.
   (a) Assume $k = 8$ without dilation.
   (b) Assume $d = 7$, and $s = 2$.
2. The output shape of the second layer is $(64, 8, 8)$. Assume $p = 0$ and $d = 1$.
   (a) Specify k and s for pooling with non-overlapping window.
   (b) What is output shape if $k = 8$ and $s = 4$ instead ?
3. The output shape of the last layer is $(128, 4, 4)$.
   (a) Assume we are not using padding or dilation.
   (b) Assume $d = 2$, $p = 2$.
   (c) Assume $p = 1$, $d = 1$.

Answer 8. Let i and o be the size of the input and output. The general formula for output size is :
$$o = \lfloor \frac{i + 2p - d(k-1) - 1}{s} \rfloor + 1$$

1. (a) Given $k = 8$ and $d = 1$, one solution to $32 = \lfloor \frac{64+2p-8}{s} \rfloor + 1$ is $p = 3$ and $s = 2$.
   (b) Given $d = 7$ and $s = 2$, one solution to $32 = \lfloor \frac{64+2p-7(k-1)-1}{2} \rfloor + 1$ is $p = 3$ and $k = 2$.
2. (a) Given $p = 0$, $d = 1$, and $k = s$ (to have non-overlapping window), $8 = \lfloor \frac{32-(k-1)-1}{k} \rfloor + 1$ yields $k = s = 4$.
   (b) $o = \lfloor \frac{32-(8-1)-1}{4} \rfloor + 1 = 7$.
3. (a) Given $p = 0$ and $d = 1$, one solution to $4 = \lfloor \frac{8-(k-1)-1}{s} \rfloor + 1$ is $k = 2$ and $s = 2$.
   (b) Given $d = 2$ and $p = 2$, one solution to $4 = \lfloor \frac{8+2\times 2-2(k-1)-1}{s} \rfloor + 1$ is $k = 3$ and $s = 2$.
   (c) Given $p = 1$ and $d = 1$, one solution to $4 = \lfloor \frac{8+2\times 1-(k-1)-1}{s} \rfloor + 1$ is $k = 7$ and $s = 1$ (or $k = 4$ and $s = 2$).