

Convolutional Neural Network: back-propagation for 2D-convolution

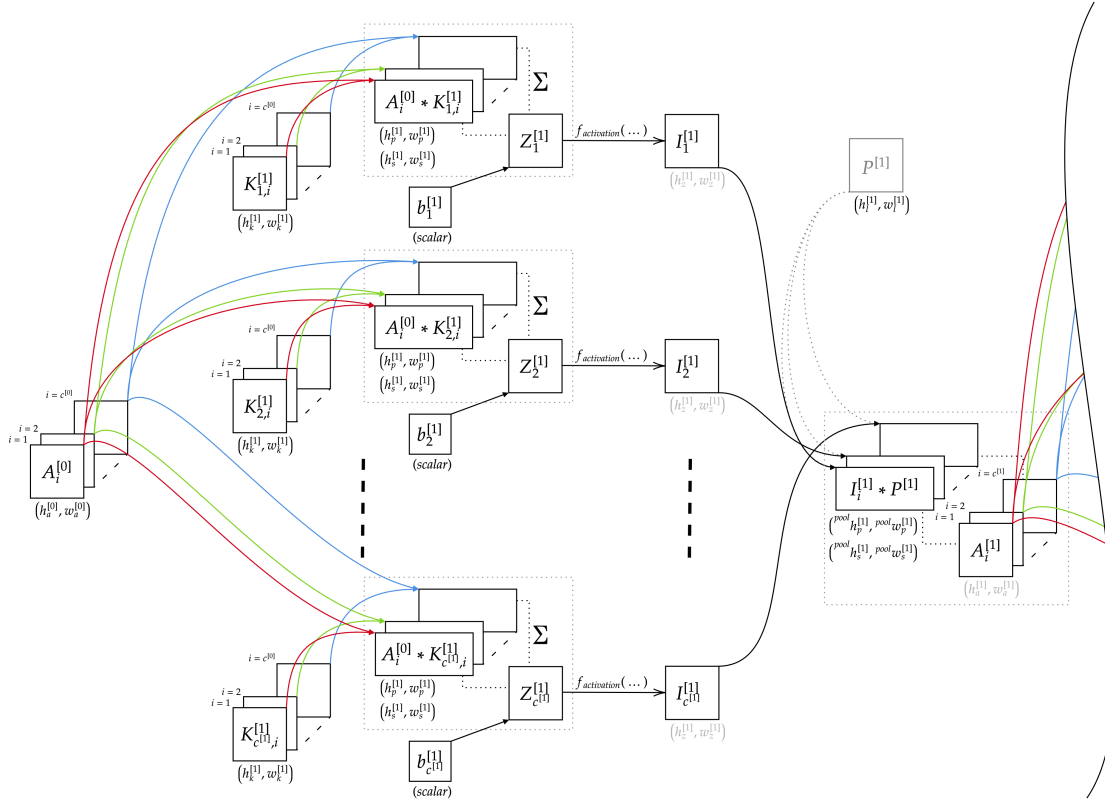
Harsha Vardhan

March 8, 2022

Abstract

This document contains derivation of the gradients for a 3-layer convolutional neural-network (with 2D-convolution), using back-propagation (or reverse-mode differentiation). For the implementation of the neural-network see the accompanying notebooks.

1 Network Architecture



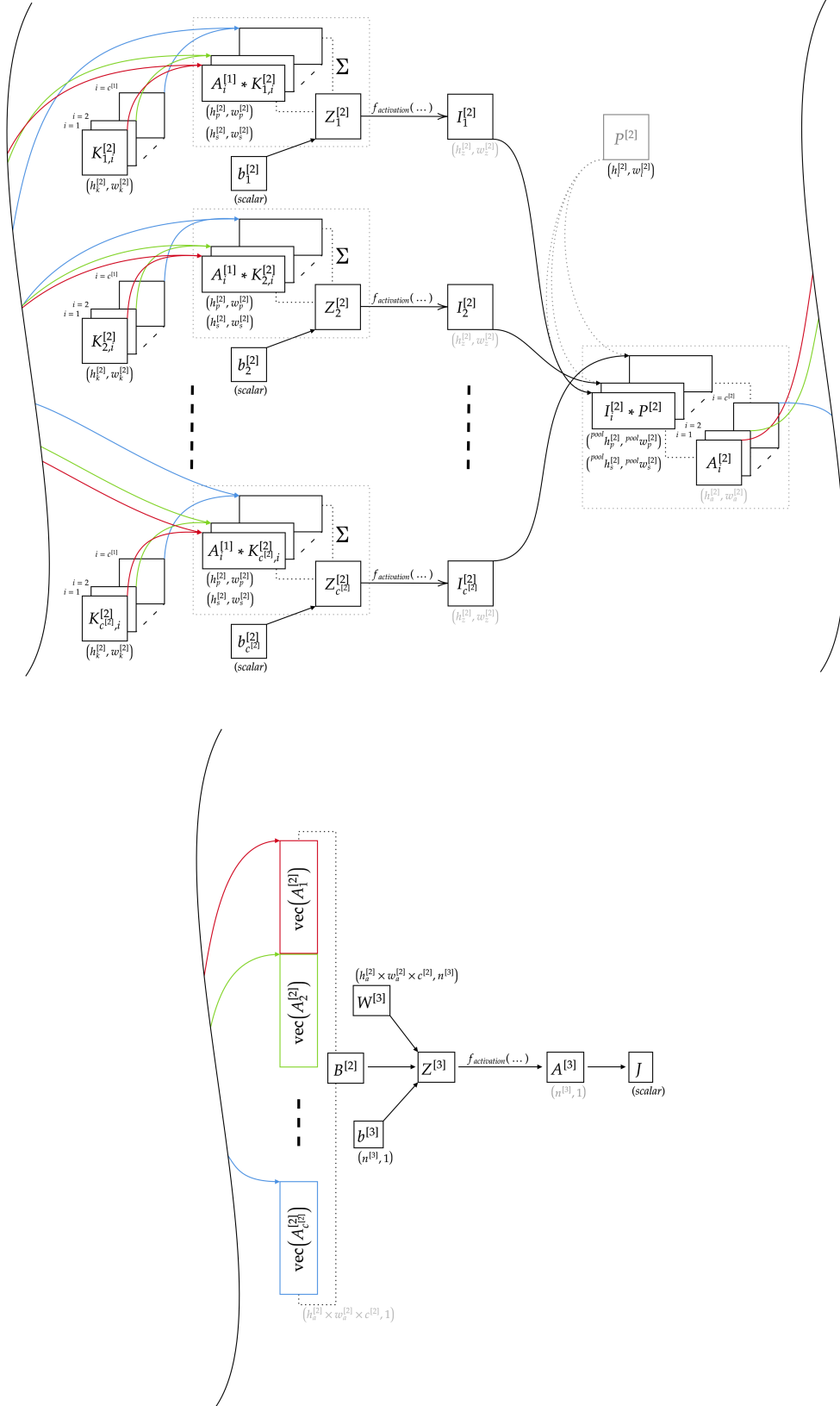


Figure 1: a three layer CNN with layers-1 & layer-2 being convolutional layers, and layer-3 a fully-connected layer

2 Forward Propagation

The equations for forward-propagation are as follows:

$$\begin{aligned}\mathbf{Z}_j^{[l]} &= \sum_{i=1}^{c^{[l-1]}} \mathbf{A}_i^{[l-1]} * \mathbf{K}_{j,i}^{[l]} + b_j^{[l]} \vec{\mathbf{1}}_{(h_z^{[l]}, w_z^{[l]})} \quad , \forall 1 \leq j \leq c^{[l]} \\ \mathbf{I}^{[l]} &= f_{activation}(\mathbf{Z}^{[l]}) \\ \mathbf{A}_j^{[l]} &= f_{pool}(\mathbf{I}_j^{[l]}) \quad , \forall 1 \leq j \leq c^{[l]}\end{aligned}$$

where,

$\mathbf{K}_{j,i}^{[l]} \in \mathbb{R}^{h_k^{[l]} \times w_k^{[l]}}$	is the i^{th} -channel of the j^{th} -kernel in l^{th} -layer, with $1 \leq i \leq c^{[l-1]}$ & $1 \leq j \leq c^{[l]}$
$b_j^{[l]} \in \mathbb{R}$	is the j^{th} -component of the bias-vector $\mathbf{b}^{[l]}$, with $1 \leq j \leq c^{[l]}$
$\mathbf{A}_i^{[l-1]} \in \mathbb{R}^{h_a^{[l-1]} \times w_a^{[l-1]}}$	is the i^{th} -channel of the output of $(l-1)^{th}$ -layer, with $1 \leq i \leq c^{[l-1]}$
$f_{activation}(\dots)$	is the activation function
$\mathbf{I}^{[l]} \in \mathbb{R}^{h_z^{[l]} \times w_z^{[l]}}$	is an intermediate result of l^{th} -layer
$f_{pool}(\dots)$	is the pooling-function. And, in the above diagram this is average/mean-pooling
$P^{[l]}$	is an abstract representation of the pooling window. And, $(^{pool}h_p^{[l]}, ^{pool}w_p^{[l]})$ & $(^{pool}h_s^{[l]}, ^{pool}w_s^{[l]})$ are the padding and stride, respectively, for the pooling operation

Note: when an input of size (h_a, w_a) is convolved with a kernel of size (h_k, w_k) , using a stride of (h_s, w_s) & padding of (h_p, w_p) , the size of the resulting output (h_z, w_z) is given by

$$h_z = \left\lfloor \frac{h_a + 2h_p - h_k}{h_s} + 1 \right\rfloor; \quad w_z = \left\lfloor \frac{w_a + 2w_p - w_k}{w_s} + 1 \right\rfloor$$

3 Optimization: gradient-descent

The optimization is performed according to the following equations:

$$\begin{aligned}\mathbf{K}_{j,i}^{[l]} &:= \mathbf{K}_{j,i}^{[l]} - \alpha \nabla_{K_{j,i}^{[l]}} J \\ \mathbf{b}^{[l]} &:= \mathbf{b}^{[l]} - \alpha \nabla_{b^{[l]}} J\end{aligned}$$

where, α is the learning-rate/step-size.

3.1 Back-propagation

The gradients in the above equations are derived using back-propagation, as follows:

- Since, J is a scalar, we can write $J = \text{tr}(J) = J^\top = \text{tr}(J^\top)$. And the derivative can be computed as follows:

$$\begin{aligned} dJ &= d(\text{tr}(J)) \\ &= \text{tr}(dJ) \end{aligned}$$

The objective while computing the above derivative, is to massage the expression to the following form:

$$dy = \text{tr}(\mathbf{A}d\mathbf{X})$$

then,

$$\frac{dy}{d\mathbf{X}} = \mathbf{A}$$

- Let \mathbf{A} and \mathbf{B} be $m \times n$ and $p \times q$ matrices. Then the essence of the derivative $\frac{d\mathbf{A}}{d\mathbf{B}}$ is to compute the derivative of every entry of \mathbf{A} w.r.t. every entry of \mathbf{B} . Therefore, we can simplify this matrix derivative by vectorizing (using the vec operator) the matrices and perform a vector-by-vector derivative.

In the following computations, by ${}_v\mathbf{A}$ we denote $\text{vec}(\mathbf{A})$; where,

- if \mathbf{A} is an $m \times n$ matrix, then ${}_v\mathbf{A}$ is the $mn \times 1$ vector obtained by stacking columns of \mathbf{A} one below the other.
- if \mathbf{A} is an $m \times n \times c$ matrix, then ${}_v\mathbf{A}$ is the $mn \times c$ dimensional matrix whose columns are the vectors ${}_v\mathbf{A}_j$, $\forall 1 \leq j \leq c$.

See [1] and [2] for more information.

3.1.1 Computing $\frac{dJ}{d\mathbf{W}^{[3]}}$, $\frac{dJ}{d\mathbf{b}^{[3]}}$, and $\frac{dJ}{d\mathbf{B}^{[2]}}$

We have,

$$\frac{dJ}{d\mathbf{Z}^{[3]}} = \frac{dJ}{d\mathbf{A}^{[3]}} \frac{d\mathbf{A}^{[3]}}{d\mathbf{Z}^{[3]}}$$

where,

$$\begin{aligned} \frac{dJ}{d\mathbf{A}^{[3]}} &= \begin{bmatrix} \frac{\partial}{\partial a_1^{[3]}} & \frac{\partial}{\partial a_2^{[3]}} & \cdots & \frac{\partial}{\partial a_{n^{[3]}}^{[3]}} \end{bmatrix} \otimes J \\ &= \begin{bmatrix} \frac{\partial J}{\partial a_1^{[3]}} & \frac{\partial J}{\partial a_2^{[3]}} & \cdots & \frac{\partial J}{\partial a_{n^{[3]}}^{[3]}} \end{bmatrix} \end{aligned}$$

and,

$$\begin{aligned} \frac{d\mathbf{A}^{[3]}}{d\mathbf{Z}^{[3]}} &= \begin{bmatrix} \frac{\partial}{\partial z_1^{[3]}} & \frac{\partial}{\partial z_2^{[3]}} & \cdots & \frac{\partial}{\partial z_{n^{[3]}}^{[3]}} \end{bmatrix} \otimes \begin{bmatrix} a_1^{[3]} \\ a_2^{[3]} \\ \vdots \\ a_{n^{[3]}}^{[3]} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial a_1^{[3]}}{\partial z_1^{[3]}} & \frac{\partial a_1^{[3]}}{\partial z_2^{[3]}} & \cdots & \frac{\partial a_1^{[3]}}{\partial z_{n^{[3]}}^{[3]}} \\ \frac{\partial a_2^{[3]}}{\partial z_1^{[3]}} & \frac{\partial a_2^{[3]}}{\partial z_2^{[3]}} & \cdots & \frac{\partial a_2^{[3]}}{\partial z_{n^{[3]}}^{[3]}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial a_{n^{[3]}}^{[3]}}{\partial z_1^{[3]}} & \frac{\partial a_{n^{[3]}}^{[3]}}{\partial z_2^{[3]}} & \cdots & \frac{\partial a_{n^{[3]}}^{[3]}}{\partial z_{n^{[3]}}^{[3]}} \end{bmatrix} \end{aligned}$$

Note: for the derivatives in this section, we have $\mathbf{A}^{[3]} = {}_v\mathbf{A}^{[3]}$, $\mathbf{Z}^{[3]} = {}_v\mathbf{Z}^{[3]}$, and $\mathbf{B}^{[2]} = {}_v\mathbf{B}^{[2]}$; since, $\mathbf{Z}^{[3]}$, $\mathbf{A}^{[3]}$, and $\mathbf{B}^{[2]}$ are already vectors.

Then,

$$dJ = \text{tr} \left(\frac{dJ}{d\mathbf{Z}^{[3]}} d\mathbf{Z}^{[3]} \right)$$

where, $d\mathbf{Z}^{[3]}$ can be expanded as,

$$\begin{aligned} d\mathbf{Z}^{[3]} &= d((\mathbf{W}^{[3]})^\top \mathbf{B}^{[2]} + \mathbf{b}^{[3]}) \\ &= d(\mathbf{W}^{[3]})^\top \mathbf{B}^{[2]} + (\mathbf{W}^{[3]})^\top d(\mathbf{B}^{[2]}) + d\mathbf{b}^{[3]} \end{aligned}$$

when differentiating w.r.t. $\mathbf{W}^{[3]}$, we have $d\mathbf{b}^{[3]} = d\mathbf{B}^{[2]} = 0$. So,

$$\begin{aligned} dJ &= \text{tr} \left(\frac{dJ}{d\mathbf{Z}^{[3]}} d(\mathbf{W}^{[3]})^\top \mathbf{B}^{[2]} \right) \\ &= \text{tr} \left((\mathbf{B}^{[2]})^\top d\mathbf{W}^{[3]} \left(\frac{dJ}{d\mathbf{Z}^{[3]}} \right)^\top \right) \\ &= \text{tr} \left(\left(\frac{dJ}{d\mathbf{Z}^{[3]}} \right)^\top (\mathbf{B}^{[2]})^\top d\mathbf{W}^{[3]} \right) \\ &\implies \frac{dJ}{d\mathbf{W}^{[3]}} = \left(\frac{dJ}{d\mathbf{Z}^{[3]}} \right)^\top (\mathbf{B}^{[2]})^\top \end{aligned}$$

when differentiating w.r.t. $\mathbf{b}^{[3]}$, we have $d(\mathbf{W}^{[3]})^\top = d\mathbf{B}^{[2]} = 0$. So,

$$\begin{aligned} dJ &= \text{tr} \left(\frac{dJ}{d\mathbf{Z}^{[3]}} d\mathbf{b}^{[3]} \right) \\ &\implies \frac{dJ}{d\mathbf{b}^{[3]}} = \frac{dJ}{d\mathbf{Z}^{[3]}} \end{aligned}$$

when differentiating w.r.t. $\mathbf{B}^{[2]}$, we have $d(\mathbf{W}^{[3]})^\top = d\mathbf{b}^{[3]} = 0$. So,

$$\begin{aligned} dJ &= \text{tr} \left(\frac{dJ}{d\mathbf{Z}^{[3]}} (\mathbf{W}^{[3]})^\top d\mathbf{B}^{[2]} \right) \\ &\implies \frac{dJ}{d\mathbf{B}^{[2]}} = \frac{dJ}{d\mathbf{Z}^{[3]}} (\mathbf{W}^{[3]})^\top \end{aligned}$$

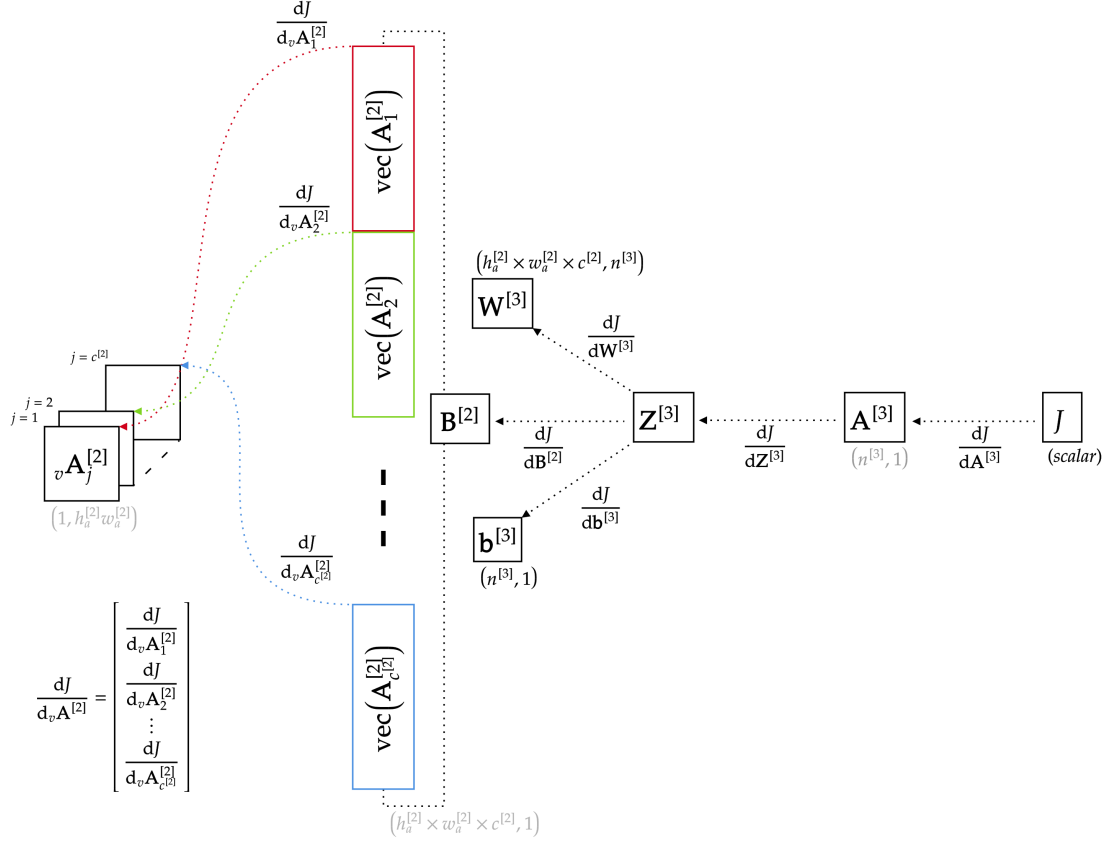


Figure 2: backward-propagation through the fully-connected layer, and the reshaping of the Jacobian $\frac{dJ}{d\mathbf{B}^{[2]}}$ into the combined (in terms of channels) Jacobian $\frac{dJ}{d_v\mathbf{A}^{[2]}}$.

3.1.2 Computing $\frac{dJ}{d_v\mathbf{A}^{[2]}}$

The vector $\mathbf{B}^{[2]}$, obtained by flattening the tensor $\mathbf{A}^{[2]}$, has the following structure:

$$B^{[2]} = \begin{bmatrix} \text{vec}(\mathbf{A}_1^{[2]}) \\ \text{vec}(\mathbf{A}_2^{[2]}) \\ \vdots \\ \text{vec}(\mathbf{A}_{c^{[2]}}^{[2]}) \end{bmatrix} = \begin{bmatrix} a_{1,11}^{[2]} \\ \vdots \\ a_{1,h_a^{[2]}w_a^{[2]}}^{[2]} \\ a_{2,11}^{[2]} \\ \vdots \\ a_{2,h_a^{[2]}w_a^{[2]}}^{[2]} \\ \vdots \\ \vdots \\ a_{c^{[2]},11}^{[2]} \\ \vdots \\ a_{c^{[2]},h_a^{[2]}w_a^{[2]}}^{[2]} \end{bmatrix}$$

where,

$$\text{vec}(\mathbf{A}_j^{[2]}) = \begin{bmatrix} a_{j,11}^{[2]} \\ a_{j,21}^{[2]} \\ \vdots \\ a_{j,h_a^{[2]}1}^{[2]} \\ a_{j,12}^{[2]} \\ a_{j,22}^{[2]} \\ \vdots \\ a_{j,h_a^{[2]}2}^{[2]} \\ \vdots \\ \vdots \\ a_{j,1w_a^{[2]}}^{[2]} \\ a_{j,2w_a^{[2]}}^{[2]} \\ \vdots \\ a_{j,h_a^{[2]}w_a^{[2]}}^{[2]} \end{bmatrix} = \begin{bmatrix} a_{j,11}^{[2]} \\ \vdots \\ a_{j,h_a^{[2]}w_a^{[2]}}^{[2]} \end{bmatrix} \quad \forall 1 \leq j \leq c^{[2]}$$

Therefore, the derivative $\frac{dJ}{d\mathbf{B}^{[2]}}$ assumes the following structure:

$$\begin{aligned}
\frac{dJ}{d\mathbf{B}^{[2]}} &= \begin{bmatrix} \frac{\partial}{\partial a_{1,11}^{[2]}} & \cdots & \frac{\partial}{\partial a_{1,h_a^{[2]}w_a^{[2]}}^{[2]}} & \frac{\partial}{\partial a_{2,11}^{[2]}} & \cdots \\ & \frac{\partial}{\partial a_{2,h_a^{[2]}w_a^{[2]}}^{[2]}} & \cdots & \cdots & \frac{\partial}{\partial a_{c^{[2]},11}^{[2]}} & \cdots & \frac{\partial}{\partial a_{c^{[2]},h_a^{[2]}w_a^{[2]}}^{[2]}} \end{bmatrix} \otimes J \\
&= \begin{bmatrix} \frac{\partial J}{\partial a_{1,11}^{[2]}} & \cdots & \frac{\partial J}{\partial a_{1,h_a^{[2]}w_a^{[2]}}^{[2]}} & \frac{\partial J}{\partial a_{2,11}^{[2]}} & \cdots \\ & \frac{\partial J}{\partial a_{2,h_a^{[2]}w_a^{[2]}}^{[2]}} & \cdots & \cdots & \frac{\partial J}{\partial a_{c^{[2]},11}^{[2]}} & \cdots & \frac{\partial J}{\partial a_{c^{[2]},h_a^{[2]}w_a^{[2]}}^{[2]}} \end{bmatrix}
\end{aligned}$$

We now reshape the vector $\frac{dJ}{d\mathbf{B}^{[2]}}$ into a $c^{[2]} \times h_a^{[2]}w_a^{[2]}$ dimensional matrix such that each row corresponds to a channel of $\mathbf{A}^{[2]}$, (figure 2) i.e. each row is a derivative of J w.r.t. $\text{vec}(\mathbf{A}_j^{[2]})$, as follows:

$$\frac{dJ}{d_v \mathbf{A}^{[2]}} = \begin{bmatrix} \frac{\partial J}{\partial a_{1,11}^{[2]}} & \frac{\partial J}{\partial a_{1,21}^{[2]}} & \cdots & \frac{\partial J}{\partial a_{1,h_a^{[2]}w_a^{[2]}}^{[2]}} \\ \frac{\partial J}{\partial a_{2,11}^{[2]}} & \frac{\partial J}{\partial a_{2,21}^{[2]}} & \cdots & \frac{\partial J}{\partial a_{2,h_a^{[2]}w_a^{[2]}}^{[2]}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial J}{\partial a_{c^{[2]},11}^{[2]}} & \frac{\partial J}{\partial a_{c^{[2]},21}^{[2]}} & \cdots & \frac{\partial J}{\partial a_{c^{[2]},h_a^{[2]}w_a^{[2]}}^{[2]}} \end{bmatrix}$$

3.1.3 Computing $\frac{dJ}{d_v \mathbf{I}^{[2]}}$

$$\frac{dJ}{d_v \mathbf{I}^{[2]}} = \frac{dJ}{d_v \mathbf{A}^{[2]}} \frac{d_v \mathbf{A}^{[2]}}{d_v \mathbf{I}^{[2]}}$$

Where, $\frac{d_v \mathbf{A}^{[2]}}{d_v \mathbf{I}^{[2]}}$ is computed based on the type of pooling, as follows:

Average Pooling: Let $P^{[2]}$ be a $h_l^{[2]} \times w_l^{[2]}$ dimensional matrix with all entries equal to $\frac{1}{h_l^{[2]}w_l^{[2]}}$. Then, the average pooled matrix $\mathbf{A}_j^{[2]}$ is obtained by convolving $P^{[2]}$ with $\mathbf{I}_j^{[2]}$, with a padding of $({}^l h_p^{[2]}, {}^l w_p^{[2]})$ and a stride of $({}^l h_s^{[2]}, {}^l w_s^{[2]})$, i.e.

$$\mathbf{A}_j^{[2]} = \mathbf{I}_j^{[2]} * P^{[2]}$$

Now, let $I_{j,pq}^{[2]}$ be the (p, q) -th entry of the matrix $\mathbf{I}_j^{[2]}$, for any $1 \leq j \leq c^{[2]}$. Then the derivative $\frac{d_v A_j^{[2]}}{dI_{j,pq}^{[2]}}$ is computed as follows:

$$\frac{d\mathbf{A}_j^{[2]}}{dI_{j,pq}^{[2]}} = \frac{d\mathbf{I}_j^{[2]}}{dI_{j,pq}^{[2]}} * P^{[2]} \implies \frac{d_v \mathbf{A}_j^{[2]}}{dI_{j,pq}^{[2]}} = \text{vec} \left(\frac{d\mathbf{I}_j^{[2]}}{dI_{j,pq}^{[2]}} * P^{[2]} \right)$$

where, the convolution is performed with the same padding and stride as that used for computing $\mathbf{A}_j^{[2]}$ from $\mathbf{I}_j^{[2]}$, i.e. a padding of $({}^l h_p^{[2]}, {}^l w_p^{[2]})$ and a stride of $({}^l h_s^{[2]}, {}^l w_s^{[2]})$.

Also, notice that the value of the derivative $\frac{d\mathbf{I}_j^{[2]}}{dI_{j,pq}^{[2]}}$ is independent of the entries in

$\mathbf{I}_j^{[2]}$, i.e. for any $1 \leq p \leq h_z^{[2]}$ & $1 \leq q \leq w_z^{[2]}$ the value of the derivative $\frac{d\mathbf{I}_j^{[2]}}{dI_{j,pq}^{[2]}}$ is the same for all $1 \leq j \leq c^{[2]}$.

Extending the above derivative w.r.t. all the entries in ${}_v\mathbf{I}_j^{[2]}$ results in a $h_a^{[2]}w_a^{[2]} \times h_z^{[2]}w_z^{[2]}$ dimensional matrix, defined as follows

$$\begin{aligned} \frac{d{}_v\mathbf{A}_j^{[2]}}{d{}_v\mathbf{I}_j^{[2]}} &= \begin{bmatrix} \frac{\partial}{\partial I_{j,11}^{[2]}} & \frac{\partial}{\partial I_{j,21}^{[2]}} & \cdots & \frac{\partial}{\partial I_{j,h_z^{[2]}w_z^{[2]}}^{[2]}} \end{bmatrix} \otimes {}_v\mathbf{A}_j^{[2]} \\ &= \begin{bmatrix} \frac{\partial {}_v\mathbf{A}_j^{[2]}}{\partial I_{j,11}^{[2]}} & \frac{\partial {}_v\mathbf{A}_j^{[2]}}{\partial I_{j,21}^{[2]}} & \cdots & \frac{\partial {}_v\mathbf{A}_j^{[2]}}{\partial I_{j,h_z^{[2]}w_z^{[2]}}^{[2]}} \end{bmatrix} \\ &= \begin{bmatrix} \text{vec}\left(\frac{d\mathbf{I}_j^{[2]}}{dI_{j,11}^{[2]}} * P^{[2]}\right) & \text{vec}\left(\frac{d\mathbf{I}_j^{[2]}}{dI_{j,21}^{[2]}} * P^{[2]}\right) & \cdots & \text{vec}\left(\frac{d\mathbf{I}_j^{[2]}}{dI_{j,h_z^{[2]}w_z^{[2]}}^{[2]}} * P^{[2]}\right) \end{bmatrix} \\ &\quad \text{and, is equal for all channels } 1 \leq j \leq c^{[2]}. \text{ Hence, we have,} \\ &= \frac{d{}_v\mathbf{A}^{[2]}}{d{}_v\mathbf{I}^{[2]}} \end{aligned}$$

Max-pooling: the backward pass for a max-pooling operation is performed by routing the gradient to the input that had the highest value in the forward pass. Hence, during the forward pass of a pooling layer it is common to keep track of the index of the max activation (sometimes also called the switches) so that gradient routing is efficient during back-propagation. [3]

L^2 -norm pooling: TODO

3.1.4 Computing $\frac{dJ}{d{}_v\mathbf{Z}^{[2]}}$

$$\frac{dJ}{d{}_v\mathbf{Z}^{[2]}} = \frac{dJ}{d{}_v\mathbf{I}^{[2]}} \frac{d{}_v\mathbf{I}^{[2]}}{d{}_v\mathbf{Z}^{[2]}}$$

Here, the derivative $\frac{d{}_v\mathbf{I}^{[2]}}{d{}_v\mathbf{Z}^{[2]}}$ has to be computed channel-wise because the value of the derivative is dependent on the values of the entries of $\mathbf{I}_j^{[2]}$, i.e. for some channel $1 \leq j \leq c^{[2]}$, we have

$$\frac{dJ}{d{}_v\mathbf{Z}_j^{[2]}} = \frac{dJ}{d{}_v\mathbf{I}_j^{[2]}} \frac{d{}_v\mathbf{I}_j^{[2]}}{d{}_v\mathbf{Z}_j^{[2]}}$$

and then stacked as follows

$$\frac{dJ}{d{}_v\mathbf{Z}^{[2]}} = \begin{bmatrix} \frac{dJ}{d{}_v\mathbf{Z}_1^{[2]}} \\ \frac{dJ}{d{}_v\mathbf{Z}_2^{[2]}} \\ \vdots \\ \frac{dJ}{d{}_v\mathbf{Z}_{c^{[2]}}^{[2]}} \end{bmatrix} = \begin{bmatrix} \frac{dJ}{d{}_v\mathbf{I}_1^{[2]}} \frac{d{}_v\mathbf{I}_1^{[2]}}{d{}_v\mathbf{Z}_1^{[2]}} \\ \frac{dJ}{d{}_v\mathbf{I}_2^{[2]}} \frac{d{}_v\mathbf{I}_2^{[2]}}{d{}_v\mathbf{Z}_2^{[2]}} \\ \vdots \\ \frac{dJ}{d{}_v\mathbf{I}_{c^{[2]}}^{[2]}} \frac{d{}_v\mathbf{I}_{c^{[2]}}^{[2]}}{d{}_v\mathbf{Z}_{c^{[2]}}^{[2]}} \end{bmatrix}$$

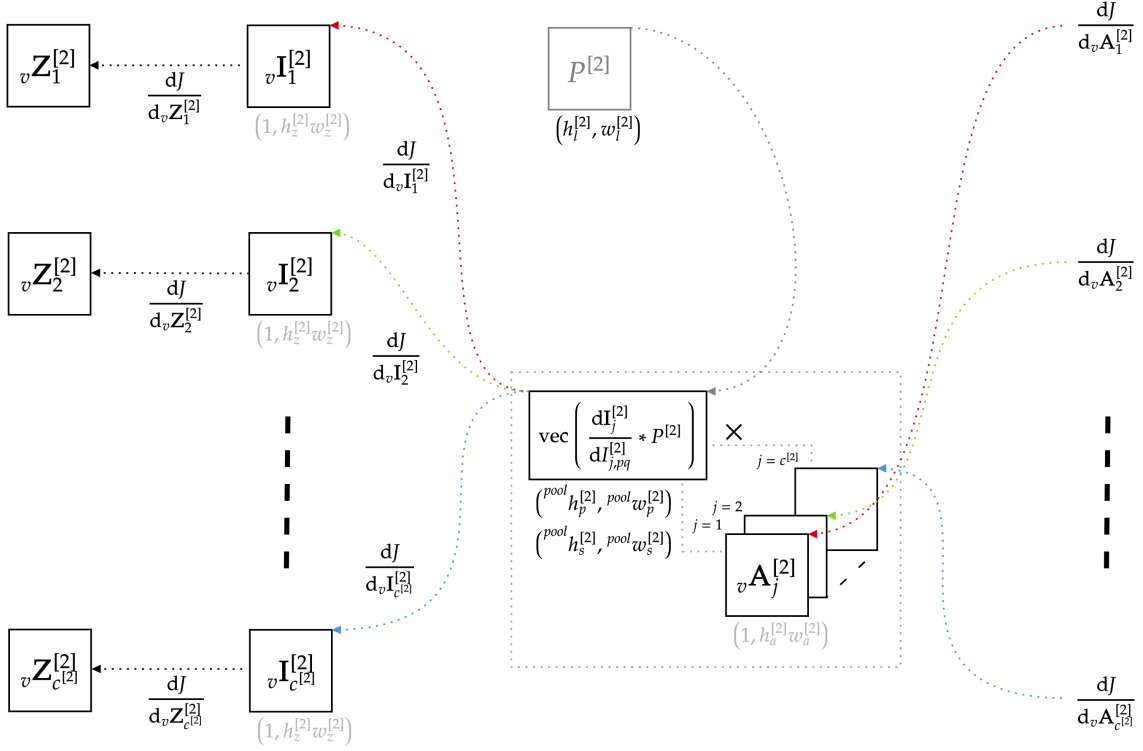


Figure 3: backward-propagation through an average-pooling layer. The derivative $\frac{d\mathbf{I}_j^{[2]}}{dI_{j,pq}^{[2]}}$ is computed $\forall 1 \leq p \leq h_z^{[2]} \ \& \ \forall 1 \leq q \leq w_z^{[2]}$. It is independent of the choice of the channel $1 \leq j \leq c^{[2]}$ for $\mathbf{I}_j^{[2]}$, and hence is equal for all the channels.

3.1.5 Computing $\frac{dJ}{d_v \mathbf{K}^{[2]}}$, $\frac{dJ}{d\mathbf{b}^{[2]}}$, and $\frac{dJ}{d_v \mathbf{A}^{[1]}}$

The matrix $\mathbf{Z}_j^{[2]}$ is computed as per the following equation

$$\mathbf{Z}_j^{[2]} = \sum_{i=1}^{c^{[1]}} \mathbf{A}_i^{[1]} * \mathbf{K}_{j,i}^{[2]} + b_j^{[2]} \vec{\mathbf{1}}_{(h_z^{[2]}, w_z^{[2]})}, \forall 1 \leq j \leq c^{[2]}$$

where, the convolution is performed with a padding of $(h_p^{[2]}, w_p^{[2]})$ and a stride of $(h_s^{[2]}, w_s^{[2]})$.

Derivative w.r.t. $_v \mathbf{K}^{[2]}$: For computing the derivative of J w.r.t. $_v \mathbf{K}^{[2]}$, let's start by considering each channel of each kernel in layer-2 separately, as follows

$$\frac{dJ}{d_v \mathbf{K}_{j,i}^{[2]}} = \frac{dJ}{d_v \mathbf{Z}_j^{[2]}} \frac{d_v \mathbf{Z}_j^{[2]}}{d_v \mathbf{K}_{j,i}^{[2]}}$$

where,

$_v \mathbf{K}_{j,i}^{[2]}$ is the i^{th} -channel of the j^{th} -kernel in layer-2, and $1 \leq j \leq c^{[2]} \ \& \ 1 \leq i \leq c^{[1]}$

$\frac{dJ}{d_v \mathbf{Z}_j^{[2]}}$ is the derivative of J w.r.t. the j^{th} -channel of $\mathbf{Z}^{[2]}$, i.e. the j^{th} -row of the matrix $\frac{dJ}{d_v \mathbf{Z}^{[2]}}$

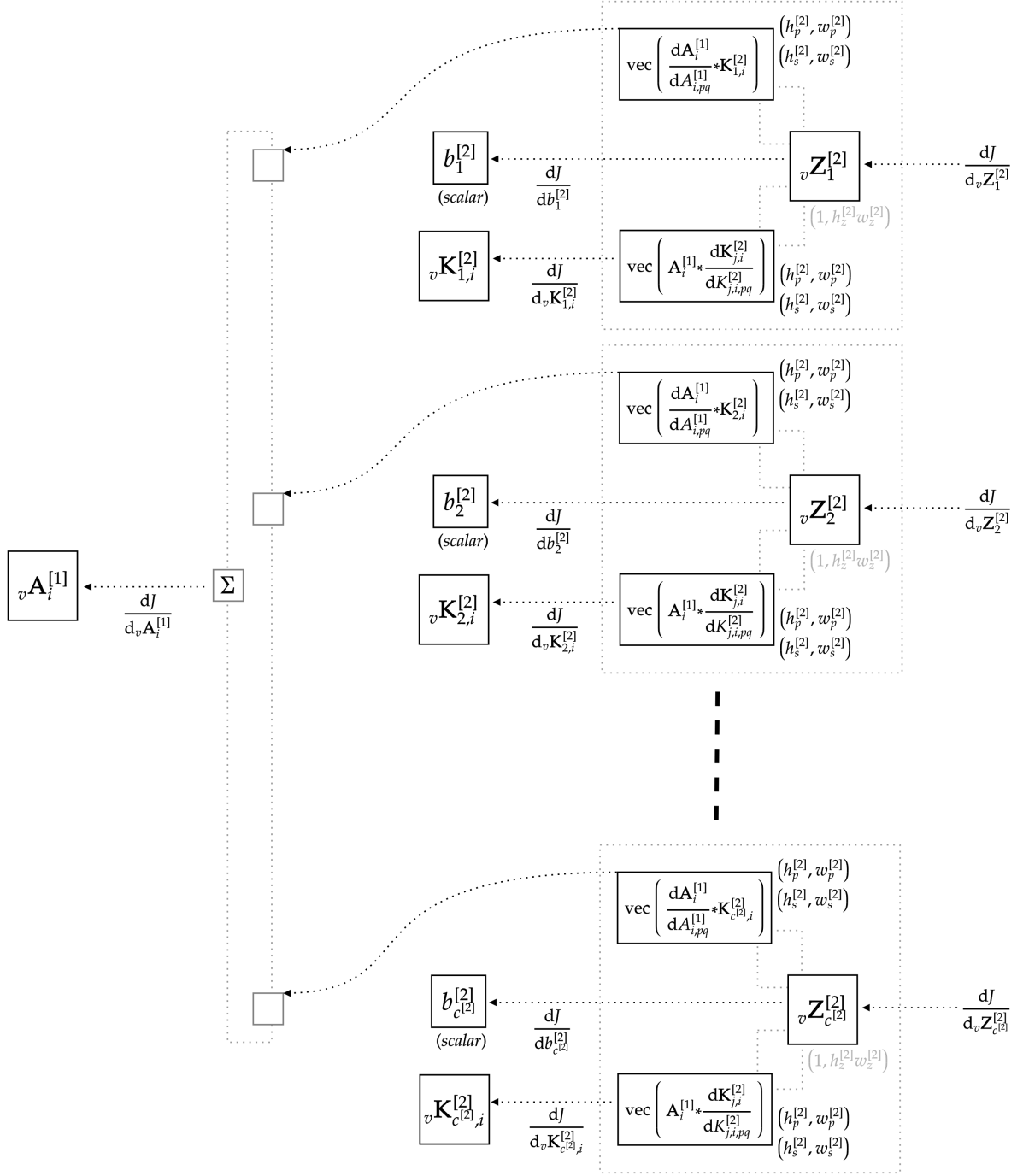


Figure 4: back-propagation through a convolutional for some channel- i , and must be performed for each $1 \leq i \leq c^{[1]}$. The derivative $\frac{d\mathbf{K}_{j,i}^{[2]}}{dK_{j,i,pq}^{[2]}}$ must be computed $\forall 1 \leq p \leq h_k^{[2]}$ & $\forall 1 \leq q \leq w_k^{[2]}$. Also, the value of this derivative is independent of the choice of the kernel, and hence is equal for all $1 \leq j \leq c^{[2]}$.

Then the derivative of $\mathbf{Z}_j^{[2]}$ w.r.t. $K_{j,i,pq}^{[2]}$, the $(p, q)^{th}$ -entry of $\mathbf{K}_{j,i}^{[2]}$, is computed as follows

$$\frac{d\mathbf{Z}_j^{[2]}}{dK_{j,i,pq}^{[2]}} = \mathbf{A}_i^{[1]} * \frac{d\mathbf{K}_{j,i}^{[2]}}{dK_{j,i,pq}^{[2]}} \implies \frac{d_v \mathbf{Z}_j^{[2]}}{dK_{j,i,pq}^{[2]}} = \text{vec} \left(\mathbf{A}_i^{[1]} * \frac{d\mathbf{K}_{j,i}^{[2]}}{dK_{j,i,pq}^{[2]}} \right) \quad \text{See section-4}$$

Extending the above derivative w.r.t. all the entries of ${}_v \mathbf{K}_{j,i}^{[2]}$ results in a $h_z^{[2]} w_z^{[2]} \times h_k^{[2]} w_k^{[2]}$ dimensional matrix, defined as follows

$$\begin{aligned} \frac{d_v \mathbf{Z}_j^{[2]}}{d {}_v \mathbf{K}_{j,i}^{[2]}} &= \begin{bmatrix} \frac{\partial}{\partial K_{j,i,11}^{[2]}} & \frac{\partial}{\partial K_{j,i,21}^{[2]}} & \cdots & \frac{\partial}{\partial K_{j,i,h_k^{[2]} w_k^{[2]}}^{[2]}} \end{bmatrix} \otimes {}_v \mathbf{Z}_j^{[2]} \\ &= \begin{bmatrix} \frac{\partial {}_v \mathbf{Z}_j^{[2]}}{\partial K_{j,i,11}^{[2]}} & \frac{\partial {}_v \mathbf{Z}_j^{[2]}}{\partial K_{j,i,21}^{[2]}} & \cdots & \frac{\partial {}_v \mathbf{Z}_j^{[2]}}{\partial K_{j,i,h_k^{[2]} w_k^{[2]}}^{[2]}} \end{bmatrix} \\ &= \begin{bmatrix} \text{vec} \left(\mathbf{A}_i^{[1]} * \frac{d\mathbf{K}_{j,i}^{[2]}}{dK_{j,i,11}^{[2]}} \right) & \text{vec} \left(\mathbf{A}_i^{[1]} * \frac{d\mathbf{K}_{j,i}^{[2]}}{dK_{j,i,21}^{[2]}} \right) & \cdots & \text{vec} \left(\mathbf{A}_i^{[1]} * \frac{d\mathbf{K}_{j,i}^{[2]}}{dK_{j,i,h_k^{[2]} w_k^{[2]}}^{[2]}} \right) \end{bmatrix} \end{aligned}$$

The value of the above derivative, given a channel- i , is equal for all kernels $1 \leq j \leq c^{[2]}$, i.e. the derivative is independent of the values of the entries of $\mathbf{K}_{j,i}^{[2]}$ and depends only on the entries of i^{th} -channel of $\mathbf{A}^{[2]}$. Therefore,

$$\frac{dJ}{d {}_v \mathbf{K}_{:,i}^{[2]}} = \frac{dJ}{d {}_v \mathbf{Z}^{[2]}} \frac{d_v \mathbf{Z}^{[2]}}{d {}_v \mathbf{K}_{j,i}^{[2]}} \quad \text{for any } 1 \leq j \leq c^{[2]}$$

gives the derivative of J w.r.t. the i^{th} -channel of all the kernels in layer-2. And, the above computation must be performed for each $1 \leq i \leq c^{[1]}$ to get the gradient of J w.r.t. all the layer-2 kernels' entries.

Derivative w.r.t. $\mathbf{b}^{[2]}$: Vectorizing the above equation and taking the derivative of ${}_v \mathbf{Z}_j^{[2]}$ w.r.t. $b_j^{[2]}$, we get

$$\begin{aligned} {}_v \mathbf{Z}_j^{[2]} &= \text{vec} \left(\sum_{i=1}^{c^{[1]}} \mathbf{A}_i^{[1]} * \mathbf{K}_{j,i}^{[2]} \right) + b_j^{[2]} \text{vec} \left(\mathbf{1}_{(h_z^{[2]}, w_z^{[2]})} \right) \\ &= \text{vec} \left(\sum_{i=1}^{c^{[1]}} \mathbf{A}_i^{[1]} * \mathbf{K}_{j,i}^{[2]} \right) + b_j^{[2]} \mathbf{1}_{(h_z^{[2]} w_z^{[2]}, 1)} \\ \implies \frac{d_v \mathbf{Z}_j^{[2]}}{db_j^{[2]}} &= \mathbf{1}_{(h_z^{[2]} w_z^{[2]}, 1)} \quad \forall 1 \leq j \leq c^{[2]} \end{aligned}$$

This derivative is independent of the value of $b_j^{[2]}$, and hence is equal for all the kernels in layer-2. Therefore, we have

$$\begin{aligned}
\frac{dJ}{db_j^{[2]}} &= \frac{dJ}{d_v \mathbf{Z}_j^{[2]}} \frac{d_v \mathbf{Z}_j^{[2]}}{db_j^{[2]}} \\
&= \frac{dJ}{d_v \mathbf{Z}_j^{[2]}} \mathbf{1}_{(h_z^{[2]} w_z^{[2]}, 1)}
\end{aligned}$$

Now, expanding the above derivative w.r.t. all components of $\mathbf{b}^{[2]}$, we get

$$\begin{aligned}
\frac{dJ}{d\mathbf{b}^{[2]}} &= \begin{bmatrix} \frac{dJ}{d_v \mathbf{Z}_1^{[2]}} \mathbf{1}_{(h_z^{[2]} w_z^{[2]}, 1)} & \frac{dJ}{d_v \mathbf{Z}_2^{[2]}} \mathbf{1}_{(h_z^{[2]} w_z^{[2]}, 1)} & \cdots & \frac{dJ}{d_v \mathbf{Z}_{c^{[1]}}^{[2]}} \mathbf{1}_{(h_z^{[2]} w_z^{[2]}, 1)} \end{bmatrix} \\
&= \left(\frac{dJ}{d_v \mathbf{Z}^{[2]}} \mathbf{1}_{(h_z^{[2]} w_z^{[2]}, 1)} \right)^\top
\end{aligned}$$

Derivative w.r.t. $\mathbf{A}^{[1]}$: Each channel of $\mathbf{A}^{[1]}$ is convolved by the corresponding channels of all the kernels in layer-2. Therefore, the gradient w.r.t. $\mathbf{A}_i^{[2]}$, for any $1 \leq i \leq c^{[1]}$, is the sum of gradients flowing in from all the kernels, i.e.

$$\frac{dJ}{d_v \mathbf{A}_i^{[1]}} = \sum_{j=1}^{c^{[2]}} \frac{dJ}{d_v \mathbf{Z}_j^{[2]}} \frac{d_v \mathbf{Z}_j^{[2]}}{d_v \mathbf{A}_i^{[1]}}$$

Let, $A_{i,pq}^{[1]}$ be the $(p, q)^{th}$ -entry of the i^{th} -channel of $\mathbf{A}^{[1]}$. Then, the derivative of $\mathbf{Z}_j^{[2]}$ w.r.t. $A_{i,pq}^{[1]}$ is computed as follows

$$\frac{d\mathbf{Z}_j^{[2]}}{dA_{i,pq}^{[1]}} = \frac{d\mathbf{A}_i^{[1]}}{dA_{i,pq}^{[1]}} * \mathbf{K}_{j,i}^{[2]} \implies \frac{d_v \mathbf{Z}_j^{[2]}}{dA_{i,pq}^{[1]}} = \text{vec} \left(\frac{d\mathbf{A}_i^{[1]}}{dA_{i,pq}^{[1]}} * \mathbf{K}_{j,i}^{[2]} \right) \quad \text{See section-4}$$

where, the above convolution is performed with the same padding and stride as that used to compute $\mathbf{Z}_j^{[2]}$, i.e. with a padding of $(h_p^{[2]}, w_p^{[2]})$ and a stride of $(h_s^{[2]}, w_s^{[2]})$.

Extending the above derivative w.r.t. all the entries of $\mathbf{A}_i^{[1]}$ results in a $h_z^{[2]} w_z^{[2]} \times h_a^{[2]} w_a^{[2]}$ dimensional matrix, defined as follows

$$\begin{aligned}
\frac{d_v \mathbf{Z}_j^{[2]}}{d_v \mathbf{A}_i^{[2]}} &= \begin{bmatrix} \frac{\partial}{\partial A_{i,11}^{[1]}} & \frac{\partial}{\partial A_{i,21}^{[1]}} & \cdots & \frac{\partial}{\partial A_{i,h_a^{[1]} w_a^{[1]}}^{[1]}} \end{bmatrix} \otimes {}_v \mathbf{Z}_j^{[2]} \\
&= \begin{bmatrix} \frac{\partial_v \mathbf{Z}_j^{[2]}}{\partial A_{i,11}^{[1]}} & \frac{\partial_v \mathbf{Z}_j^{[2]}}{\partial A_{i,21}^{[1]}} & \cdots & \frac{\partial_v \mathbf{Z}_j^{[2]}}{\partial A_{i,h_a^{[1]} w_a^{[1]}}^{[1]}} \end{bmatrix} \\
&= \begin{bmatrix} \text{vec} \left(\frac{d\mathbf{A}_i^{[1]}}{dA_{i,11}^{[1]}} * \mathbf{K}_{j,i}^{[2]} \right) & \text{vec} \left(\frac{d\mathbf{A}_i^{[1]}}{dA_{i,21}^{[1]}} * \mathbf{K}_{j,i}^{[2]} \right) & \cdots & \text{vec} \left(\frac{d\mathbf{A}_i^{[1]}}{dA_{i,h_a^{[1]} w_a^{[1]}}^{[1]}} * \mathbf{K}_{j,i}^{[2]} \right) \end{bmatrix}
\end{aligned}$$

After computing $\frac{dJ}{d_v \mathbf{A}_i^{[1]}}$ for all $1 \leq i \leq c^{[1]}$, the derivative $\frac{dJ}{d_v \mathbf{A}^{[1]}}$, which is a $c^{[1]} \times h_a^{[1]} w_a^{[1]}$ dimensional matrix, is computed as follows

$$\frac{dJ}{d_v \mathbf{A}^{[1]}} = \begin{bmatrix} \frac{dJ}{d_v \mathbf{A}_1^{[1]}} \\ \frac{dJ}{d_v \mathbf{A}_2^{[1]}} \\ \vdots \\ \frac{dJ}{d_v \mathbf{A}_{c^{[1]}}^{[1]}} \end{bmatrix}$$

3.1.6 Computing $\frac{dJ}{d_v \mathbf{Z}^{[1]}}$, $\frac{dJ}{d_v \mathbf{K}^{[1]}}$, and $\frac{dJ}{d_v \mathbf{b}^{[1]}}$

Given the derivative $\frac{dJ}{d_v \mathbf{A}^{[1]}}$, we can compute the subsequent derivatives as follows

$$\frac{dJ}{d_v \mathbf{I}^{[1]}} = \frac{dJ}{d_v \mathbf{A}^{[1]}} \frac{d_v \mathbf{A}^{[1]}}{d_v \mathbf{I}^{[1]}}$$

the derivative $\frac{d_v \mathbf{A}^{[1]}}{d_v \mathbf{I}^{[1]}}$ is computed based on the pooling strategy, as described in the above *section-3.1.3*

$$\frac{dJ}{d_v \mathbf{Z}^{[1]}} = \frac{dJ}{d_v \mathbf{I}^{[1]}} \frac{d_v \mathbf{I}^{[1]}}{d_v \mathbf{Z}^{[1]}}$$

the derivative $\frac{d_v \mathbf{I}^{[1]}}{d_v \mathbf{Z}^{[1]}}$ is computed based on the activation function, as shown in *section-3.1.4*

$$\frac{dJ}{d_v \mathbf{K}_{:,i}^{[2]}} = \frac{dJ}{d_v \mathbf{Z}^{[2]}} \frac{d_v \mathbf{Z}^{[2]}}{d_v \mathbf{K}_{j,i}^{[2]}}$$

the derivative $\frac{d_v \mathbf{Z}^{[2]}}{d_v \mathbf{K}_{j,i}^{[2]}}$ is computed as shown in *section-3.1.5*. And, the derivative $\frac{dJ}{d_v \mathbf{K}_{:,i}^{[2]}}$ must be computed for each $1 \leq i \leq c^{[1]}$

⟨straightforward-to-compute⟩

$$\frac{dJ}{d_v \mathbf{b}^{[1]}} = \left(\frac{dJ}{d_v \mathbf{Z}^{[1]}} \mathbf{1}_{(h_z^{[1]} w_z^{[1]}, 1)} \right)^\top$$

3.2 Jacobian or Gradient?

In the above derivations, we have used the numerator layout while performing matrix-derivatives. One of the consequences of this decision is that the derivatives that we have computed are in-fact jacobians and not gradients. Fortunately, gradients are just transpose of jacobians. So, based on our derivations the gradients would be the following:

$$\begin{aligned} \nabla_{K_{j,i}^{[l]}} J &= \left(\frac{dJ}{d\mathbf{K}_{j,i}^{[l]}} \right)^\top \\ \nabla_{b^{[l]}} J &= \left(\frac{d_v \mathbf{Z}^{[2]}}{d_v \mathbf{b}^{[l]}} \right)^\top \end{aligned}$$

4 Appendix: derivative of a convolution

Here, we compute the derivative of a convolution w.r.t. its constituents. Let \mathbf{A} be a $(3, 4)$ dimensional matrix, β be a $(2, 2)$ dimensional matrix, and \mathbf{Z} be a $(4, 3)$ dimensional matrix obtained by convolving \mathbf{A} by β with a padding of $(1, 1)$ and a stride of $(1, 2)$, i.e.

$$\mathbf{Z} = \mathbf{A} * \beta$$

$$\begin{aligned}
 &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & a_{11} & a_{12} & a_{13} & a_{14} & 0 \\ 0 & a_{21} & a_{22} & a_{23} & a_{24} & 0 \\ 0 & a_{31} & a_{32} & a_{33} & a_{34} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \\
 &= \left[\begin{array}{ccc} \begin{bmatrix} 0 & 0 \\ 0 & a_{11} \end{bmatrix} * \beta & \begin{bmatrix} 0 & 0 \\ a_{12} & a_{13} \end{bmatrix} * \beta & \begin{bmatrix} 0 & 0 \\ a_{14} & 0 \end{bmatrix} * \beta \\ \begin{bmatrix} 0 & a_{11} \\ 0 & a_{21} \end{bmatrix} * \beta & \begin{bmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{bmatrix} * \beta & \begin{bmatrix} a_{14} & 0 \\ a_{24} & 0 \end{bmatrix} * \beta \\ \begin{bmatrix} 0 & a_{21} \\ 0 & a_{31} \end{bmatrix} * \beta & \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} * \beta & \begin{bmatrix} a_{24} & 0 \\ a_{34} & 0 \end{bmatrix} * \beta \\ \begin{bmatrix} 0 & 0 \\ 0 & a_{31} \end{bmatrix} * \beta & \begin{bmatrix} a_{32} & a_{33} \\ 0 & 0 \end{bmatrix} * \beta & \begin{bmatrix} a_{34} & 0 \\ 0 & 0 \end{bmatrix} * \beta \end{array} \right] \quad (\text{rep.1}) \\
 &= \begin{bmatrix} a_{11}\beta_{22} & a_{12}\beta_{21} + a_{13}\beta_{22} & a_{14}\beta_{21} \\ a_{11}\beta_{12} + a_{21}\beta_{22} & a_{12}\beta_{11} + a_{13}\beta_{12} + a_{22}\beta_{21} + a_{23}\beta_{22} & a_{14}\beta_{11} + a_{24}\beta_{21} \\ a_{21}\beta_{12} + a_{31}\beta_{22} & a_{22}\beta_{11} + a_{23}\beta_{12} + a_{32}\beta_{21} + a_{33}\beta_{22} & a_{24}\beta_{11} + a_{34}\beta_{21} \\ a_{31}\beta_{12} & a_{32}\beta_{11} + a_{33}\beta_{12} & a_{34}\beta_{11} \end{bmatrix} \quad (\text{rep.2}) \\
 &= \begin{bmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \\ z_{41} & z_{42} & z_{43} \end{bmatrix} \quad (\text{rep.3})
 \end{aligned}$$

Also, the vectorized forms of these matrices are as follows

$${}_v\mathbf{A} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{12} \\ a_{22} \\ a_{32} \\ a_{13} \\ a_{23} \\ a_{33} \\ a_{14} \\ a_{24} \\ a_{34} \end{bmatrix}; \quad {}_v\beta = \begin{bmatrix} \beta_{11} \\ \beta_{21} \\ \beta_{12} \\ \beta_{22} \end{bmatrix}; \quad {}_v\mathbf{Z} = \begin{bmatrix} z_{11} \\ z_{21} \\ z_{31} \\ z_{41} \\ z_{12} \\ z_{22} \\ z_{32} \\ z_{42} \\ z_{13} \\ z_{23} \\ z_{33} \\ z_{43} \end{bmatrix}$$

Essence of matrix/vector derivative: Let \mathbf{M} and \mathbf{N} be any two vectors/matrices. Then, the objective of the derivative $\frac{d\mathbf{M}}{d\mathbf{N}}$ is to compute the derivative of each entry of \mathbf{M} w.r.t. each entry of \mathbf{N} . And, the representation of a group of entries as matrices, vectors, or tensors is merely a matter of notation.

Now, consider the following derivatives:

$$\frac{d\mathbf{Z}}{d\beta} = \text{reshape}\left(\frac{d{}_v\mathbf{Z}}{d{}_v\beta}\right):$$

$$\begin{aligned} \frac{d{}_v\mathbf{Z}}{d{}_v\beta} &= \begin{bmatrix} \frac{\partial}{\partial\beta_{11}} & \frac{\partial}{\partial\beta_{21}} & \frac{\partial}{\partial\beta_{12}} & \frac{\partial}{\partial\beta_{22}} \end{bmatrix} \otimes {}_v\mathbf{Z} \\ &= \begin{bmatrix} \frac{\partial{}_v\mathbf{Z}}{\partial\beta_{11}} & \frac{\partial{}_v\mathbf{Z}}{\partial\beta_{21}} & \frac{\partial{}_v\mathbf{Z}}{\partial\beta_{12}} & \frac{\partial{}_v\mathbf{Z}}{\partial\beta_{22}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial z_{11}}{\partial\beta_{11}} & \frac{\partial z_{11}}{\partial\beta_{21}} & \frac{\partial z_{11}}{\partial\beta_{12}} & \frac{\partial z_{11}}{\partial\beta_{22}} \\ \frac{\partial z_{21}}{\partial\beta_{11}} & \frac{\partial z_{21}}{\partial\beta_{21}} & \frac{\partial z_{21}}{\partial\beta_{12}} & \frac{\partial z_{21}}{\partial\beta_{22}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial z_{43}}{\partial\beta_{11}} & \frac{\partial z_{43}}{\partial\beta_{21}} & \frac{\partial z_{43}}{\partial\beta_{12}} & \frac{\partial z_{43}}{\partial\beta_{22}} \end{bmatrix} \end{aligned} \quad \text{eq.a1-1}$$

Let's compute the value of an arbitrary element of this matrix, say derivative of z_{32} w.r.t. β_{21} (you should try out with some other combination of entries)

$$\begin{aligned}
\frac{\partial z_{32}}{\partial \beta_{21}} &= \frac{\partial(a_{22}\beta_{11} + a_{23}\beta_{12} + a_{32}\beta_{21} + a_{33}\beta_{22})}{\partial \beta_{21}} && \text{using rep.2} \\
&= a_{32} \\
&= \frac{\partial \left(\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} * \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \right)}{\partial \beta_{21}} && \text{using rep.1} \\
&= \frac{\partial \left(\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} \right)}{\partial \beta_{21}} * \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} + \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} * \frac{\partial \left(\begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \right)}{\partial \beta_{21}} \\
&= \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} * \frac{\partial \left(\begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \right)}{\partial \beta_{21}} = \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} * \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} && \text{eq.a1-2}
\end{aligned}$$

In the matrix in eq.a1-1 above, each column involves the derivative of ${}_v\mathbf{Z}$ w.r.t. a single entry of β , hence, the value of the derivative $\frac{\partial \beta}{\partial \beta_{21}}$ is equal for all entries in the column corresponding to β_{21} . Extending the derivative in eq.a1-2 to all the entries z_{pq} , $\forall 1 \leq p \leq 4$ & $\forall 1 \leq q \leq 3$ and using rep.2, we get the following

$$\frac{d\mathbf{Z}}{d\beta_{21}} = \mathbf{A} * \frac{d\beta}{d\beta_{21}} = \mathbf{A} * \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

where, the convolution is performed with the same padding and stride as that used for computing \mathbf{Z} . And, vectorizing the above derivative we get

$$\frac{d_v \mathbf{Z}}{d\beta_{21}} = \text{vec} \left(\mathbf{A} * \frac{d\beta}{d\beta_{21}} \right) \quad \text{eq.a1-3}$$

Extending the derivative in eq.a1-3 to all the entries β_{rs} , $\forall 1 \leq r \leq 2$ & $\forall 1 \leq s \leq 2$, we get the following

$$\frac{d_v \mathbf{Z}}{d_v \beta} = \begin{bmatrix} \text{vec} \left(\mathbf{A} * \frac{d\beta}{d\beta_{11}} \right) & \text{vec} \left(\mathbf{A} * \frac{d\beta}{d\beta_{12}} \right) & \text{vec} \left(\mathbf{A} * \frac{d\beta}{d\beta_{21}} \right) & \text{vec} \left(\mathbf{A} * \frac{d\beta}{d\beta_{22}} \right) \end{bmatrix}$$

$$\frac{d\mathbf{Z}}{d\mathbf{A}} = \text{reshape}\left(\frac{d_v\mathbf{Z}}{d_v\mathbf{A}}\right):$$

$$\begin{aligned} \frac{d_v\mathbf{Z}}{d_v\mathbf{A}} &= \begin{bmatrix} \frac{\partial}{\partial A_{11}} & \frac{\partial}{\partial A_{21}} & \cdots & \frac{\partial}{\partial \beta_{34}} \end{bmatrix} \otimes {}_v\mathbf{Z} \\ &= \begin{bmatrix} \frac{\partial_v\mathbf{Z}}{\partial A_{11}} & \frac{\partial_v\mathbf{Z}}{\partial A_{21}} & \cdots & \frac{\partial_v\mathbf{Z}}{\partial A_{34}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial z_{11}}{\partial A_{11}} & \frac{\partial z_{11}}{\partial A_{21}} & \cdots & \frac{\partial z_{11}}{\partial A_{34}} \\ \frac{\partial z_{21}}{\partial A_{11}} & \frac{\partial z_{21}}{\partial A_{21}} & \cdots & \frac{\partial z_{21}}{\partial A_{34}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_{43}}{\partial A_{11}} & \frac{\partial z_{43}}{\partial A_{21}} & \cdots & \frac{\partial z_{43}}{\partial A_{34}} \end{bmatrix} \quad \text{eq.a2-1} \end{aligned}$$

Let's compute the value of an arbitrary element of this matrix, say derivative of z_{32} w.r.t. A_{23} (you should try out with some other combination of entries)

$$\begin{aligned} \frac{\partial z_{32}}{\partial A_{23}} &= \frac{\partial(a_{22}\beta_{11} + a_{23}A_{12} + a_{32}\beta_{21} + a_{33}\beta_{22})}{\partial A_{23}} \quad \text{using rep.2} \\ &= \beta_{12} \\ &= \frac{\partial\left(\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} * \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix}\right)}{\partial A_{23}} \quad \text{using rep.1} \\ &= \frac{\partial\left(\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}\right)}{\partial A_{23}} * \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} + \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} * \frac{\partial\left(\begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix}\right)}{\partial A_{23}} \\ &= \frac{\partial\left(\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}\right)}{\partial A_{23}} * \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} * \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \quad \text{eq.a2-2} \end{aligned}$$

In the matrix in eq.a2-1 above, each column involves the derivative of ${}_v\mathbf{Z}$ w.r.t. a single entry of \mathbf{A} , hence, the value of the derivative $\frac{\partial\mathbf{A}}{\partial A_{23}}$ is equal for all entries in the column corresponding to A_{23} . Extending the derivative in eq.a2-2 to all the entries z_{pq} , $\forall 1 \leq p \leq 4$ & $\forall 1 \leq q \leq 3$ and using rep.2, we get the following

$$\frac{d\mathbf{Z}}{dA_{23}} = \frac{d\mathbf{A}}{dA_{23}} * \beta = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} * \beta$$

where, the convolution is performed with the same padding and stride as that used for computing \mathbf{Z} . And, vectorizing the above derivative we get

$$\frac{d_v\mathbf{Z}}{dA_{23}} = \text{vec}\left(\frac{d\mathbf{A}}{dA_{23}} * \beta\right) \quad \text{eq.a2-3}$$

Extending the derivative in eq.a2-3 to all the entries A_{rs} , $\forall 1 \leq r \leq 3$ & $\forall 1 \leq s \leq 4$, we get the following

$$\frac{d_v \mathbf{Z}}{d_v \beta} = \left[\text{vec} \left(\frac{d\mathbf{A}}{dA_{11}} * \beta \right) \quad \text{vec} \left(\frac{d\mathbf{A}}{dA_{21}} * \beta \right) \quad \dots \quad \text{vec} \left(\frac{d\mathbf{A}}{dA_{34}} * \beta \right) \right]$$

References

- [1] T. Minka, “Old and new matrix algebra useful for statistics,” Sep. 1997. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/old-new-matrix-algebra-useful-statistics/>.
- [2] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Second. John Wiley, 1999, ISBN: 0471986321 9780471986324 047198633X 9780471986331.
- [3] “CS231n: Convolutional neural networks for visual recognition.” (2021), [Online]. Available: <http://vision.stanford.edu/teaching/cs231n/> (visited on 02/06/2022).