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

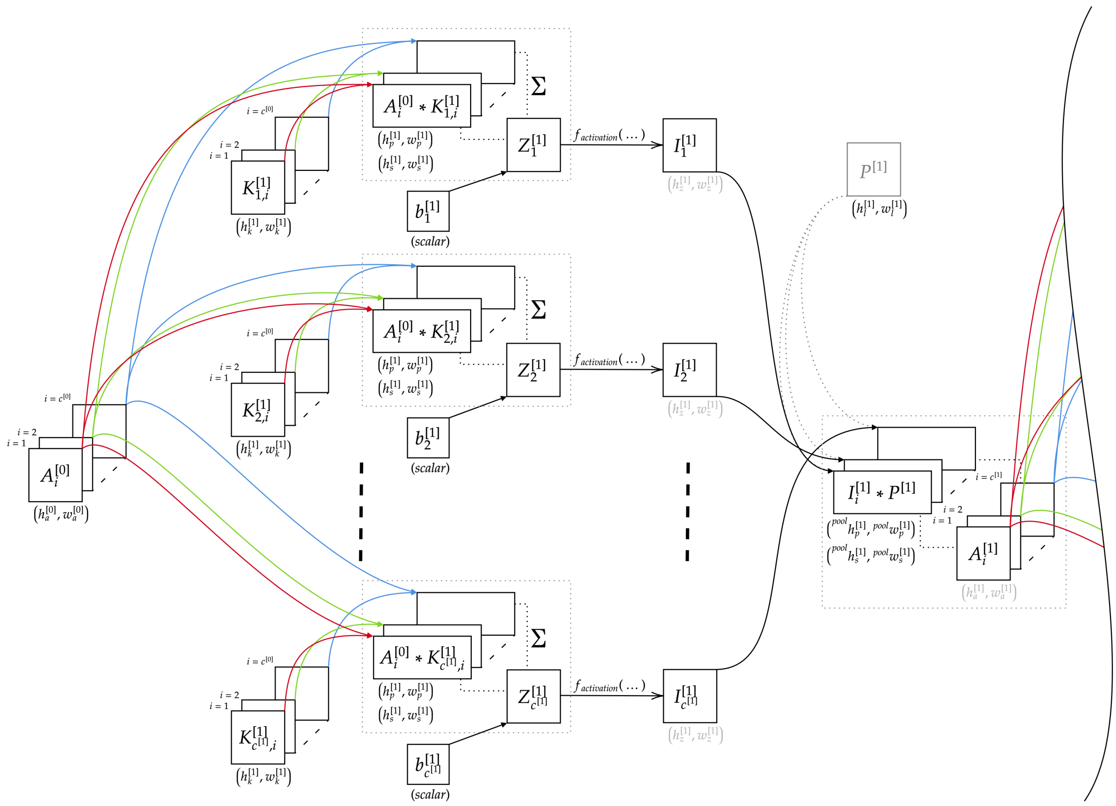
Harsha Vardhan

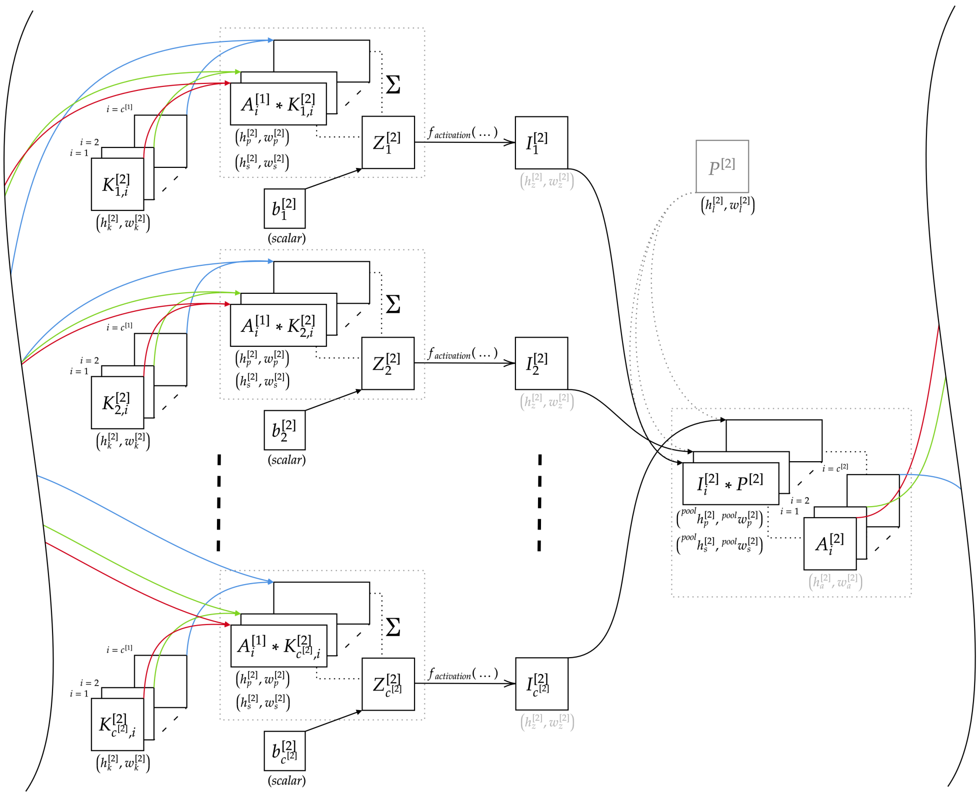
March 27, 2022

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





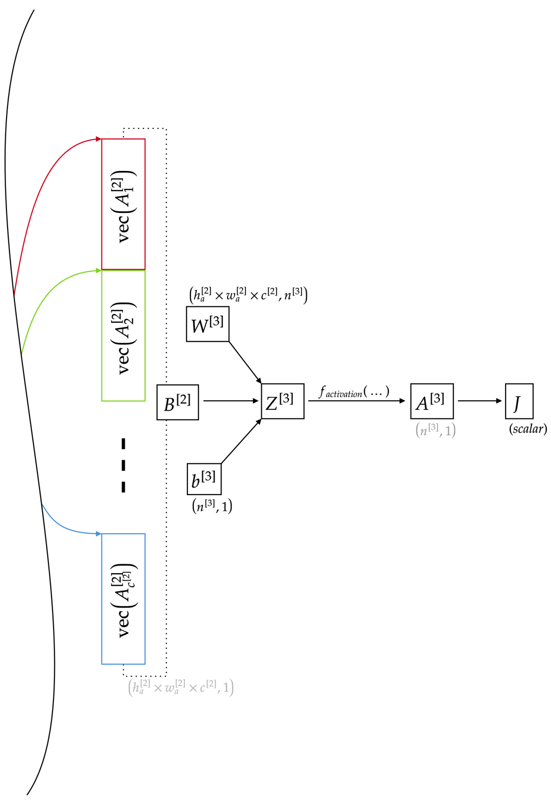


Figure 1: a three layer CNN with layers-1 & layer-2 being convolutional layers, and layer-3 a fully-connected layer

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

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

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 $J = \mathrm{tr}(J) = J^\intercal = \mathrm{tr}(J^\intercal)$. And the derivative can be computed as follows:
$$\begin{aligned} \mathrm{d}J &= \mathrm{d}(\mathrm{tr}(J)) \\ &= \mathrm{tr}(\mathrm{d}J) \end{aligned}$$
The objective while computing the above derivative, is to massage the expression to the following form:
$$\mathrm{d}y = \mathrm{tr}(\mathbf{A}\mathrm{d}\mathbf{X})$$
then,
$$\frac{\mathrm{d}y}{\mathrm{d}\mathbf{X}} = \mathbf{A}$$
- Let $\mathbf{A}$ and $\mathbf{B}$ be $m \times n$ and $p \times q$ matrices. Then the essence of the derivative $\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\mathbf{B}}$ 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 $\texttt{vec}(\mathbf{A})$; 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 $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{W}^{[3]}}$, $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{b}^{[3]}}$, and $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{B}^{[2]}}$

We have,
$$\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}} = \frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[3]}} \frac{\mathrm{d}\mathbf{A}^{[3]}}{\mathrm{d}\mathbf{Z}^{[3]}}$$
where,
$$\begin{aligned} \frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[3]}} &= \begin{bmatrix} \frac{\partial}{\partial a_1^{[3]}} & \frac{\partial}{\partial a_2^{[3]}} & \cdots & \frac{\partial}{\partial a_{n^{[3]}}^{[3]}} \end{bmatrix} \otimes J \\ &= \begin{bmatrix} \frac{\partial J}{\partial a_1^{[3]}} & \frac{\partial J}{\partial a_2^{[3]}} & \cdots & \frac{\partial J}{\partial a_{n^{[3]}}^{[3]}} \end{bmatrix} \end{aligned}$$

and,

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

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

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,

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

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

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

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

$$\begin{aligned} \mathrm{d}J &= \mathrm{tr}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}\mathrm{d}(\mathbf{W}^{[3]})^\intercal\mathbf{B}^{[2]}\right) \\ &= \mathrm{tr}\left((\mathbf{B}^{[2]})^\intercal\mathrm{d}\mathbf{W}^{[3]}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[4]}}\right)^\intercal\right) \\ &= \mathrm{tr}\left(\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}\right)^\intercal(\mathbf{B}^{[2]})^\intercal\mathrm{d}\mathbf{W}^{[3]}\right) \\ \implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{W}^{[3]}} &= \left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}\right)^\intercal(\mathbf{B}^{[2]})^\intercal \end{aligned}$$

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

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

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

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

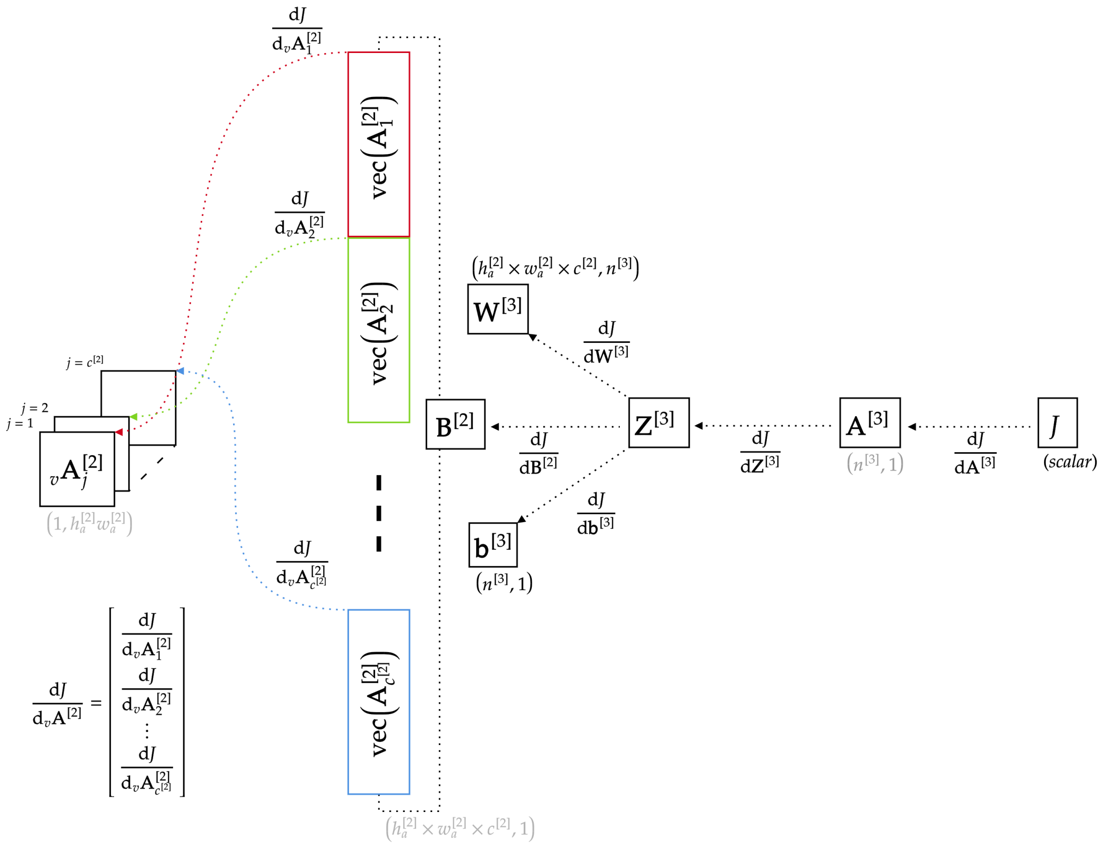


Figure 2: backward-propagation through the fully-connected layer, and the reshaping of the Jacobian $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{B}^{[2]}}$ into the combined (in terms of channels) Jacobian $\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{A}^{[2]}}$.

3.1.2 Computing $\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{A}^{[2]}}$

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

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

where,

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

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

$$
\begin{aligned}
\frac{\mathrm{d}J}{\mathrm{d}\mathbf{B}^{[2]}} &= \Bigg[\frac{\partial}{\partial a^{[2]}_{1,11}} \cdots \frac{\partial}{\partial a^{[2]}_{1,h^{[2]}_a w^{[2]}_a}} \frac{\partial}{\partial a^{[2]}_{2,11}} \cdots \\
&\qquad \frac{\partial}{\partial a^{[2]}_{2,h^{[2]}_a w^{[2]}_a}} \cdots \cdots \frac{\partial}{\partial a^{[2]}_{c^{[2]},11}} \cdots \frac{\partial}{\partial a^{[2]}_{c^{[2]},h^{[2]}_a w^{[2]}_a}}\Bigg] \otimes J \\
&= \Bigg[\frac{\partial J}{\partial a^{[2]}_{1,11}} \cdots \frac{\partial J}{\partial a^{[2]}_{1,h^{[2]}_a w^{[2]}_a}} \frac{\partial J}{\partial a^{[2]}_{2,11}} \cdots \\
&\qquad \frac{\partial J}{\partial a^{[2]}_{2,h^{[2]}_a w^{[2]}_a}} \cdots \cdots \frac{\partial J}{\partial a^{[2]}_{c^{[2]},11}} \cdots \frac{\partial J}{\partial a^{[2]}_{c^{[2]},h^{[2]}_a w^{[2]}_a}}\Bigg]
\end{aligned}
$$

We now reshape the vector $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{B}^{[2]}}$ into a $c^{[2]} \times h^{[2]}_a w^{[2]}_a$ 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. $\texttt{vec}(\mathbf{A}^{[2]}_j)$, as follows:

$$
\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{A}^{[2]}} = \begin{bmatrix}
\frac{\partial J}{\partial a^{[2]}_{1,11}} & \frac{\partial J}{\partial a^{[2]}_{1,21}} & \cdots & \frac{\partial J}{\partial a^{[2]}_{1,h^{[2]}_a w^{[2]}_a}} \\
\frac{\partial J}{\partial a^{[2]}_{2,11}} & \frac{\partial J}{\partial a^{[2]}_{2,21}} & \cdots & \frac{\partial J}{\partial a^{[2]}_{2,h^{[2]}_a w^{[2]}_a}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial J}{\partial a^{[2]}_{c^{[2]},11}} & \frac{\partial J}{\partial a^{[2]}_{c^{[2]},21}} & \cdots & \frac{\partial J}{\partial a^{[2]}_{c^{[2]},h^{[2]}_a w^{[2]}_a}}
\end{bmatrix}
$$

3.1.3 Computing $\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{I}^{[2]}}$

$$
\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{I}^{[2]}} = \frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{A}^{[2]}} \frac{\mathrm{d}_v\mathbf{A}^{[2]}}{\mathrm{d}_v\mathbf{I}^{[2]}}
$$

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

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

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

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

$$
\frac{\mathrm{d}\mathbf{A}^{[2]}_j}{\mathrm{d}I_{j,pq}} = \frac{\mathrm{d}\mathbf{I}^{[2]}_j}{\mathrm{d}I^{[2]}_{j,pq}} * P^{[2]} \implies \frac{\mathrm{d}_v\mathbf{A}^{[2]}_j}{\mathrm{d}I^{[2]}_{j,pq}} = \texttt{vec}\left(\frac{\mathrm{d}\mathbf{I}^{[2]}_j}{\mathrm{d}I^{[2]}_{j,pq}} * P^{[2]}\right)
$$

where, the convolution is performed with the same padding and stride as that used for computing $\mathbf{A}^{[2]}_j$ from $\mathbf{I}^{[2]}_j$, i.e. a padding of $({}^l h^{[2]}_p, {}^l w^{[2]}_p)$ and a stride of $({}^l h^{[2]}_s, {}^l w^{[2]}_s)$. Also, notice that the value of the derivative $\frac{\mathrm{d}\mathbf{I}^{[2]}_j}{\mathrm{d}I^{[2]}_{j,pq}}$ 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 $\frac{\mathrm{d}\mathbf{I}_j^{[2]}}{\mathrm{d}I_{j,pq}^{[2]}}$ 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

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

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]

We then, use the switches to compute a "mask", which is essentially the matrix $\frac{\mathrm{d}_v\mathbf{A}^{[l]}}{\mathrm{d}_v\mathbf{I}^{[l]}}$. For a detailed derivation of the mask, see section-5.

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

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

Here, the derivative $\frac{\mathrm{d}_v\mathbf{I}^{[2]}}{\mathrm{d}_v\mathbf{Z}^{[2]}}$ 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

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

and then stacked as follows

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

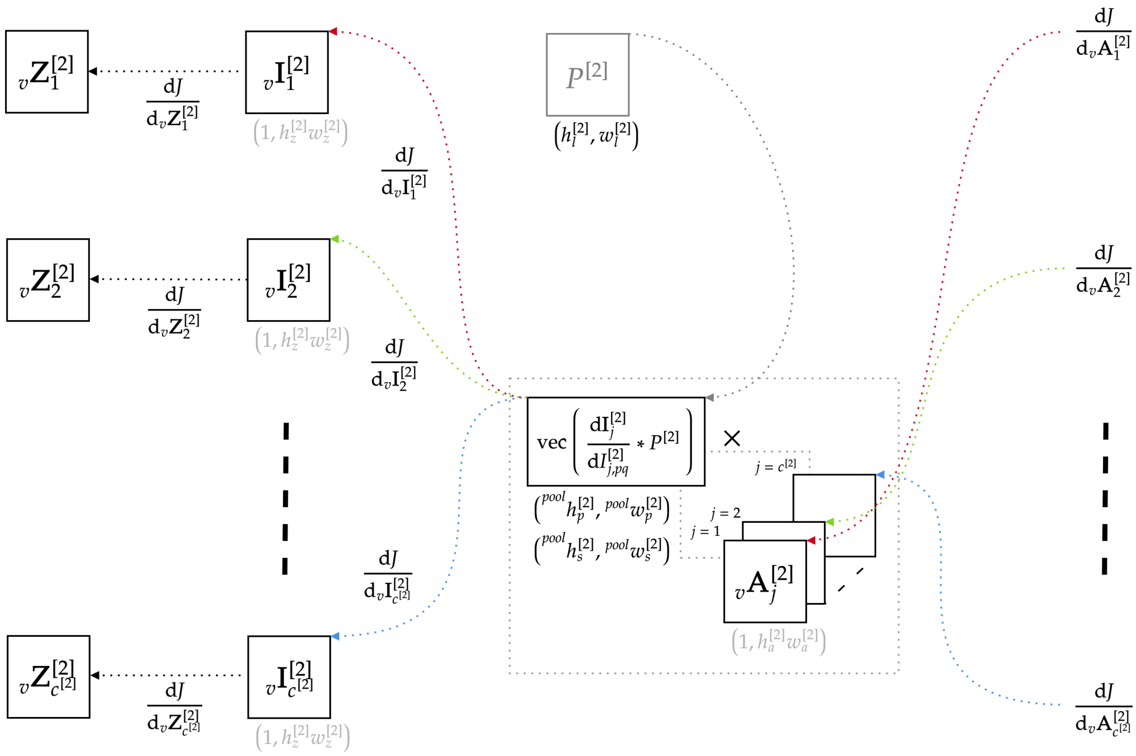


Figure 3: backward-propagation through a average-pooling layer. The derivative $\frac{d\mathbf{I}_j^{[2]}}{dI_{j,pq}^{[2]}}$ is computed $\forall 1 \leq p \leq h_z^{[2]}$ & $\forall 1 \leq q \leq w_z^{[2]}$. It is independent of the choice of the channel $1 \leq j \leq c^{[2]}$ for $\mathbf{I}_j^{[2]}$,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

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

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. $_v\mathbf{K}^{[2]}$, lets 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 $\frac{dJ}{d_v\mathbf{Z}^{[2]}}$

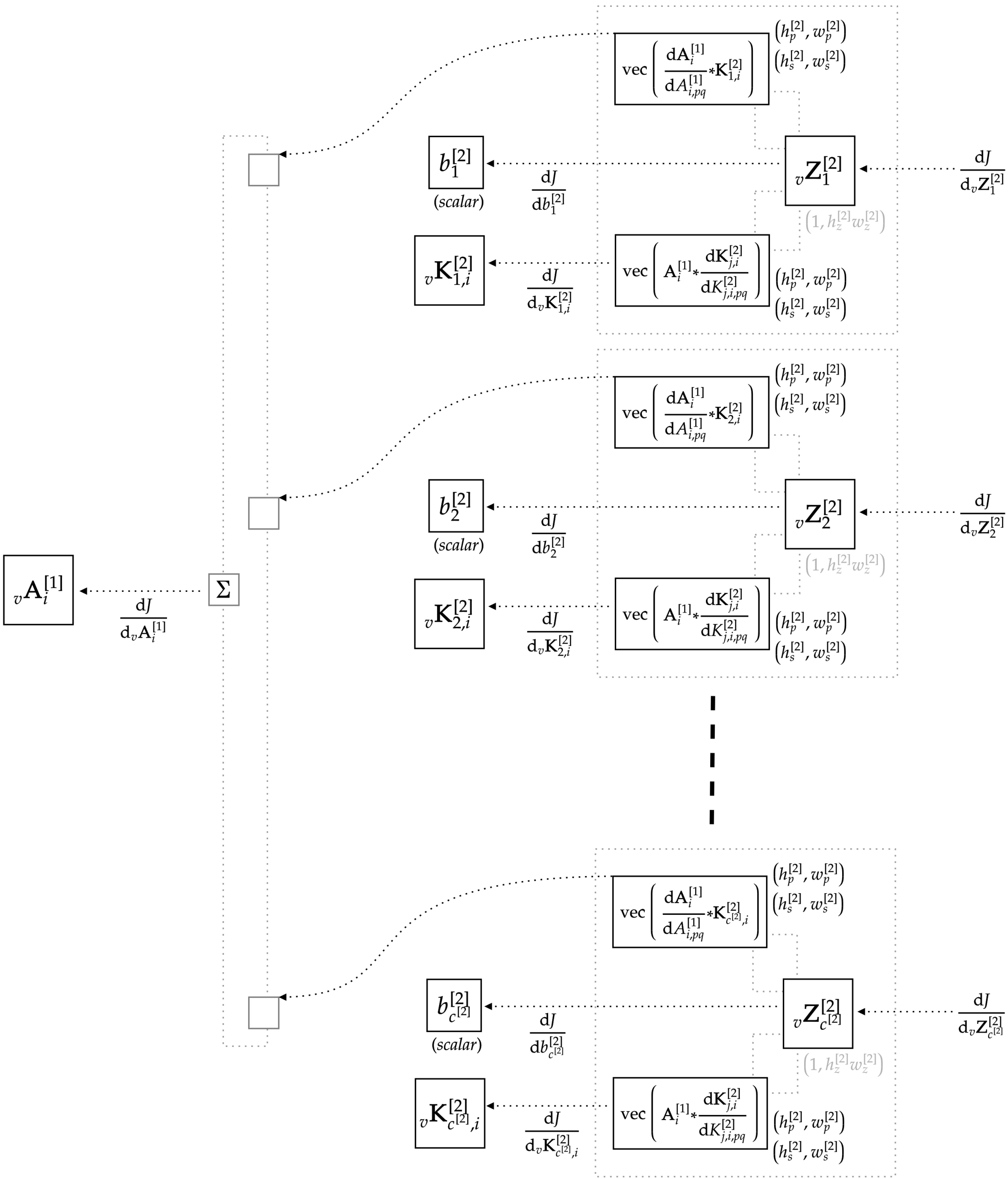


Figure 4: back-propagation through a convolutional for some channel-i, and must be performed for each $1 \leq i \leq c^{[1]}$. The derivative $\frac{\mathrm{d}\mathbf{K}_{j,i}^{[2]}}{\mathrm{d}K_{j,i,pq}^{[2]}}$ must be computed $\forall 1 \leq p \leq h_k^{[2]}$ & $\forall 1 \leq q \leq w_k^{[2]}$. Also, 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

$$\frac{\mathrm{d}\mathbf{Z}_j^{[2]}}{\mathrm{d}K_{j,i,pq}^{[2]}} = \mathbf{A}_i^{[1]} * \frac{\mathrm{d}\mathbf{K}_{j,i}^{[2]}}{\mathrm{d}K_{j,i,pq}^{[2]}} \implies \frac{\mathrm{d}_v\mathbf{Z}_j^{[2]}}{\mathrm{d}K_{j,i,pq}^{[2]}} = \mathtt{vec}\left(\mathbf{A}_i^{[1]} * \frac{\mathrm{d}\mathbf{K}_{j,i}^{[2]}}{\mathrm{d}K_{j,i,pq}^{[2]}}\right) \qquad \textit{See section-4}$$

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

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

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,

$$\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{K}_{:,i}^{[2]}} = \frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{Z}^{[2]}}\frac{\mathrm{d}_v\mathbf{Z}^{[2]}}{\mathrm{d}_v\mathbf{K}_{j,i}^{[2]}} \qquad \textbf{for any } 1 \leq j \leq c^{[2]}$$

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

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

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

$$
\begin{aligned}
\frac{\mathrm{d}J}{\mathrm{d}b_j^{[2]}} &= \frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{Z}_j^{[2]}}\frac{\mathrm{d}_v\mathbf{Z}_j^{[2]}}{\mathrm{d}b_j^{[2]}} \\
&= \frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{Z}_j^{[2]}}\mathbf{1}_{(h_z^{[2]}w_z^{[2]},1)}
\end{aligned}
$$

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

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

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 $1 \le i \le c^{[1]}$, is the sum of gradients flowing in from all the kernels, i.e.

$$
\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{A}_i^{[1]}} = \sum_{j=1}^{c^{[2]}} \frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{Z}_j^{[2]}}\frac{\mathrm{d}_v\mathbf{Z}_j^{[2]}}{\mathrm{d}_v\mathbf{A}_i^{[1]}}
$$

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

$$
\frac{\mathrm{d}\mathbf{Z}_j^{[2]}}{\mathrm{d}A_{i,pq}^{[1]}} = \frac{\mathrm{d}\mathbf{A}_i^{[1]}}{\mathrm{d}A_{i,pq}^{[1]}} * \mathbf{K}_{j,i}^{[2]} \implies \frac{\mathrm{d}_v\mathbf{Z}_j^{[2]}}{\mathrm{d}A_{i,pq}^{[1]}} = \mathrm{vec}\left(\frac{\mathrm{d}\mathbf{A}_i^{[1]}}{\mathrm{d}A_{i,pq}^{[1]}} * \mathbf{K}_{j,i}^{[2]}\right) \qquad \textit{See section-4}
$$

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

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

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

$$\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{A}^{[1]}} = \begin{bmatrix} \frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{A}_1^{[1]}} \\ \frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{A}_2^{[1]}} \\ \vdots \\ \frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{A}_{c^{[1]}}^{[1]}} \end{bmatrix}$$

3.1.6 Computing $\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{Z}^{[1]}}$, $\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{K}^{[1]}}$, and $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{b}^{[1]}}$

Given the derivative $\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{A}^{[1]}}$, we can compute the subsequent derivatives as follows

$$\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{I}^{[1]}} = \frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{A}^{[1]}} \frac{\mathrm{d}_v\mathbf{A}^{[1]}}{\mathrm{d}_v\mathbf{I}^{[1]}}$$

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

$$\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{Z}^{[1]}} = \frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{I}^{[1]}} \frac{\mathrm{d}_v\mathbf{I}^{[1]}}{\mathrm{d}_v\mathbf{Z}^{[1]}}$$

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

$$\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{K}_{:,i}^{[2]}} = \frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{Z}^{[2]}} \frac{\mathrm{d}_v\mathbf{Z}^{[2]}}{\mathrm{d}_v\mathbf{K}_{j,i}^{[2]}}$$

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

$$\frac{\mathrm{d}J}{\mathrm{d}\mathbf{b}^{[1]}} = \left(\frac{\mathrm{d}J}{\mathrm{d}_v\mathbf{Z}^{[1]}} \mathbf{1}_{(h_z^{[1]} w_z^{[1]},1)}\right)^{\mathsf{T}}$$

$\langle$straightforward-to-compute$\rangle$

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:

$$\nabla_{K_{j,i}^{[l]}} J = \left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{K}_{j,i}^{[l]}}\right)^{\mathsf{T}}$$
$$\nabla_{b^{[l]}} J = \left(\frac{\mathrm{d}_v\mathbf{Z}^{[2]}}{\mathrm{d}_v\mathbf{b}^{[l]}}\right)^{\mathsf{T}}$$

4 Appendix A: 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.

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

Also, the vectorized forms of these matrices are as follows

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

Essence of matrix/vector derivative: Let $\mathbf{M}$ and $\mathbf{N}$ be any two vectors/matrices. Then, the objective of the derivative $\frac{\mathrm{d}\mathbf{M}}{\mathrm{d}\mathbf{N}}$ 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:

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

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

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)

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

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}, $\forall 1 \le p \le 4$ & $\forall 1 \le q \le 3$ and using rep.2, we get the following

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

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

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

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

$$
\frac{\mathrm{d}_v\mathbf{Z}}{\mathrm{d}_v\beta} = \left[\mathrm{vec}\left(\mathbf{A} * \tfrac{\mathrm{d}\beta}{\mathrm{d}\beta_1 1}\right) \quad \mathrm{vec}\left(\mathbf{A} * \tfrac{\mathrm{d}\beta}{\mathrm{d}\beta_2 1}\right) \quad \mathrm{vec}\left(\mathbf{A} * \tfrac{\mathrm{d}\beta}{\mathrm{d}\beta_1 2}\right) \quad \mathrm{vec}\left(\mathbf{A} * \tfrac{\mathrm{d}\beta}{\mathrm{d}\beta_2 2}\right)\right]
$$

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

$$\frac{\mathrm{d}_v\mathbf{Z}}{\mathrm{d}_v\mathbf{A}} = \begin{bmatrix} \frac{\partial}{\partial A_{11}} & \frac{\partial}{\partial A_{21}} & \dots & \frac{\partial}{\partial A_{34}} \end{bmatrix} \otimes {}_v\mathbf{Z}$$
$$= \begin{bmatrix} \frac{\partial_v\mathbf{Z}}{\partial A_{11}} & \frac{\partial_v\mathbf{Z}}{\partial A_{21}} & \dots & \frac{\partial_v\mathbf{Z}}{\partial A_{34}} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\partial z_{11}}{\partial A_{11}} & \frac{\partial z_{11}}{\partial A_{21}} & \dots & \frac{\partial z_{11}}{\partial A_{34}} \\ \frac{\partial z_{21}}{\partial A_{11}} & \frac{\partial z_{21}}{\partial A_{21}} & \dots & \frac{\partial z_{21}}{\partial A_{34}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_{43}}{\partial A_{11}} & \frac{\partial z_{43}}{\partial A_{21}} & \dots & \frac{\partial z_{43}}{\partial A_{34}} \end{bmatrix} \qquad \text{eq.a2-1}$$

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)

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

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}, $\forall 1 \le p \le 4$ & $\forall 1 \le q \le 3$ and using rep.2, we get the following

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

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

$$\frac{\mathrm{d}_v\mathbf{Z}}{\mathrm{d}A_{23}} = \text{vec}\left(\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}A_{23}} * \beta\right) \qquad \text{eq.a2-3}$$

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

$$\frac{\mathrm{d}_v \mathbf{Z}}{\mathrm{d}_v \beta} = \left[\mathrm{vec}\left(\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}A_{11}} * \beta \right) \quad \mathrm{vec}\left(\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}A_{21}} * \beta \right) \quad \ldots \quad \mathrm{vec}\left(\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}A_{34}} * \beta \right) \right]$$

5 Appendix B: max-pooling back-propagation mask

Here, we will compute the mask for routing the gradients during back-propagation through a max-pooling layer. Let $\mathbf{I}$ be a (h_i, w_i) dimensional matrix, Ω be a (h_l, w_l) dimensional pooling-window, and $\mathbf{A}$ be a (h_a, w_a) dimensional matrix obtained by pooling $\mathbf{I}$ by Ω with a padding of $({}^l h_p, {}^l w_p)$ and a stride of $({}^l h_s, {}^l w_s)$.Then,

$$\mathbf{I} = \begin{bmatrix} I_{11} & I_{12} & \dots & I_{1w_i} \\ I_{21} & I_{22} & \dots & I_{2w_i} \\ \vdots & \vdots & \ddots & \vdots \\ I_{h_i 1} & I_{h_i 2} & \dots & I_{h_i w_i} \end{bmatrix}$$

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} & \dots & \Omega_{1w_a} \\ \Omega_{21} & \Omega_{22} & \dots & \Omega_{2w_a} \\ \vdots & \vdots & \ddots & \vdots \\ \Omega_{h_a 1} & \Omega_{h_a 2} & \dots & \Omega_{h_a w_a} \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1w_a} \\ A_{21} & A_{22} & \dots & A_{2w_a} \\ \vdots & \vdots & \ddots & \vdots \\ A_{h_a 1} & A_{h_a 2} & \dots & A_{h_a w_a} \end{bmatrix}$$

For example, let $(h_i, w_i) = (3, 4)$, $(h_l, w_l) = (3, 2)$, $({}^l h_p, {}^l w_p) = (1, 1)$, and $({}^l h_s, {}^l w_s) = (1, 2)$, then pooling results in a $(h_a, w_a) = (3, 3)$ dimensional matrix, i.e.

$$\begin{aligned} \mathbf{A} &= \mathbf{I} * \Omega \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{11} & I_{12} & I_{13} & I_{14} & 0 \\ 0 & I_{21} & I_{22} & I_{23} & I_{24} & 0 \\ 0 & I_{31} & I_{32} & I_{33} & I_{34} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} * \max \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \\ \Omega_{31} & \Omega_{32} \end{bmatrix} \end{aligned}$$

$$
= \begin{bmatrix}
\max \begin{bmatrix} 0 & 0 \\ 0 & I_{11} \\ 0 & I_{21} \end{bmatrix} & \max \begin{bmatrix} 0 & 0 \\ I_{12} & I_{13} \\ I_{22} & I_{23} \end{bmatrix} & \max \begin{bmatrix} 0 & 0 \\ I_{14} & 0 \\ I_{24} & 0 \end{bmatrix} \\
\max \begin{bmatrix} 0 & I_{11} \\ 0 & I_{21} \\ 0 & I_{31} \end{bmatrix} & \max \begin{bmatrix} I_{12} & I_{13} \\ I_{22} & I_{23} \\ I_{32} & I_{33} \end{bmatrix} & \max \begin{bmatrix} I_{14} & 0 \\ I_{24} & 0 \\ I_{34} & 0 \end{bmatrix} \\
\max \begin{bmatrix} 0 & I_{21} \\ 0 & I_{31} \\ 0 & 0 \end{bmatrix} & \max \begin{bmatrix} I_{22} & I_{23} \\ I_{32} & I_{33} \\ 0 & 0 \end{bmatrix} & \max \begin{bmatrix} I_{24} & 0 \\ I_{34} & 0 \\ 0 & 0 \end{bmatrix}
\end{bmatrix} \qquad \text{(rep.1)}
$$

$$
= \begin{bmatrix}
\max\{0, I_{11}, I_{21}\} & \max\{0, I_{12}, I_{22}, I_{13}, I_{23}\} & \max\{0, I_{14}, I_{24}\} \\
\max\{0, I_{11}, I_{21}, I_{31}\} & \max\{I_{12}, I_{22}, I_{32}, I_{13}, I_{23}, I_{33}\} & \max\{0, I_{14}, I_{24}, I_{34}\} \\
\max\{0, I_{21}, I_{31}\} & \max\{0, I_{22}, I_{23}, I_{32}, I_{33}\} & \max\{0, I_{24}, I_{34}\}
\end{bmatrix} \qquad \text{(rep.2)}
$$

$$
= \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \qquad \text{(rep.3)}
$$

In the equations above, Ω and its entries Ω_{ij} ($\forall 1 \leq i \leq h_l$ & $\forall 1 \leq j \leq w_l$) are placeholders for the entries in each of the windows (see rep.1) of matrix $\mathbf{I}$. In general, the matrix $\mathbf{A}$ will have dimensions given by

$$
h_a = \left\lfloor \frac{h_i + 2 \times {}^l h_p - h_l}{{}^l h_s} + 1 \right\rfloor ; \quad w_a = \left\lfloor \frac{w_i + 2 \times {}^l w_p - w_l}{{}^l w_s} + 1 \right\rfloor
$$

Also, the vectorized forms of the matrices $\mathbf{I}$, $\mathbf{A}$, and Ω, are as follows

$$
{}_v\mathbf{I} = \begin{bmatrix} I_{11} \\ I_{21} \\ \vdots \\ I_{h_i 1} \\ I_{12} \\ I_{22} \\ \vdots \\ I_{h_i 2} \\ \vdots \\ \vdots \\ I_{1 w_i} \\ I_{2 w_i} \\ \vdots \\ I_{h_i w_i} \end{bmatrix}; \qquad {}_v\mathbf{A} = \begin{bmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{h_a 1} \\ A_{12} \\ A_{22} \\ \vdots \\ A_{h_a 2} \\ \vdots \\ \vdots \\ A_{1 w_a} \\ A_{2 w_a} \\ \vdots \\ A_{h_a w_a} \end{bmatrix}; \qquad {}_v\Omega = \begin{bmatrix} \Omega_{11} \\ \Omega_{21} \\ \vdots \\ \Omega_{h_l 1} \\ \Omega_{12} \\ \Omega_{22} \\ \vdots \\ \Omega_{h_l 2} \\ \vdots \\ \vdots \\ \Omega_{1 w_l} \\ \Omega_{2 w_l} \\ \vdots \\ \Omega_{h_l w_l} \end{bmatrix}
$$

Each entry A_{pq}, $\forall 1 \leq p \leq h_a$ & $\forall 1 \leq q \leq w_a$, has an integer index $1 \leq i_{pq} \leq$ associated with it; this is the index of the entry in ${}^{pq}_{v}\Omega$ (i.e., the window of $\mathbf{I}$ corresponding to A_{pq}) whose maximum-value was assigned to A_{pq}. This index is usually stored during forward-propagation. In the example above, let $A_{21} = \max\{0, a_{11}, a_{21}, a_{31}\} = a_{21}$, then the index $i_{21} = 5$, i.e. the 5^{th} entry of ${}_v\Omega = [0, 0, 0, a_{11}, a_{21}, a_{31}]^\intercal$ is the maximum.

Now, we have

$$
\begin{aligned}
\frac{\mathrm{d}_v\mathbf{A}}{\mathrm{d}_v\mathbf{I}} &= \begin{bmatrix} \frac{\partial}{\partial I_{11}} & \frac{\partial}{\partial I_{21}} & \cdots & \frac{\partial}{\partial I_{34}} \end{bmatrix} \otimes {}_v\mathbf{A} \\
&= \begin{bmatrix} \frac{\partial A_{11}}{\partial I_{11}} & \frac{\partial A_{11}}{\partial I_{21}} & \cdots & \frac{\partial A_{11}}{\partial I_{34}} \\ \frac{\partial A_{21}}{\partial I_{11}} & \frac{\partial A_{21}}{\partial I_{21}} & \cdots & \frac{\partial A_{21}}{\partial I_{34}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial A_{33}}{\partial I_{11}} & \frac{\partial A_{33}}{\partial I_{21}} & \cdots & \frac{\partial A_{33}}{\partial I_{34}} \end{bmatrix} \qquad \text{eq.b-1}
\end{aligned}
$$

since, $A_{pq} = I_{kl}$ (for any $1 \leq p \leq h_a$ & $1 \leq q \leq w_a$ and for some $1 \leq k \leq h_i$ & $1 \leq l \leq w_l$), each row of the matrix (in eq.b-1) has all of its entries equal to 0 except for one entry corresponding to the derivative $\frac{\mathrm{d}A_{pq}}{\mathrm{d}I_{kl}}$, which will be equal to 1.

So, given i_{pq} $\forall 1 \leq p \leq h_a$ & $\forall 1 \leq q \leq w_a$, we can compute each row of the above matrix as follows

1. For each index i_{pq}, compute the position (k, l) of the corresponding entry in the matrix $\mathbf{I}$. Let $({}^{pq}r_b, {}^{pq}c_b)$ be the position of the entry from ${}^{pq}\Omega$ that is assigned to A_{pq} during max-pooling (see figure-5). [*Note that k (and l) need not be equal to ${}^{pq}r_b$ (and ${}^{pq}c_b$)*].

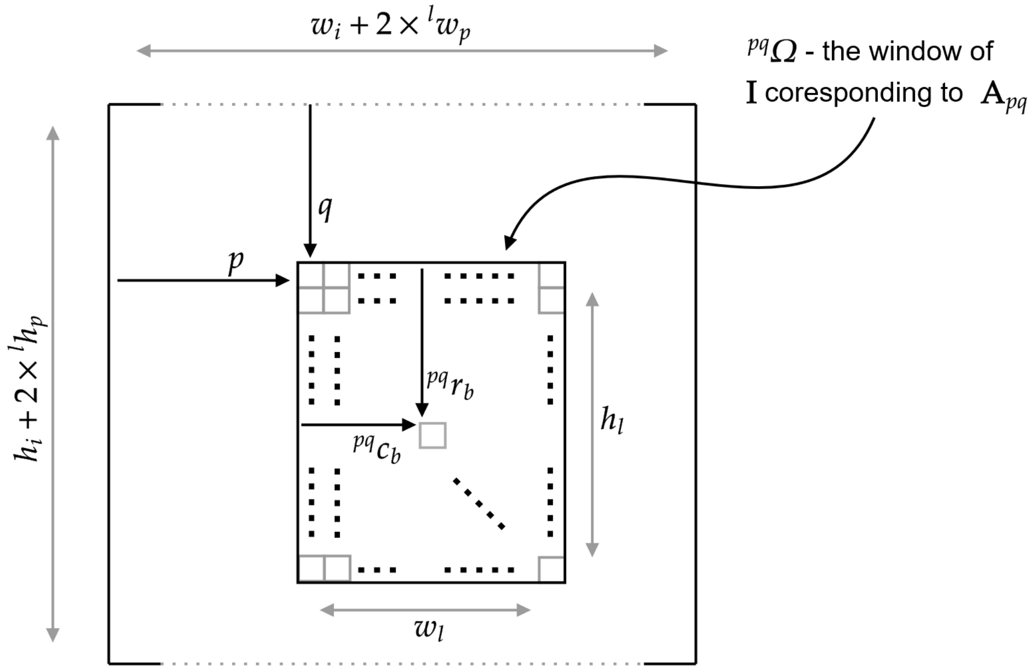


Figure 5: an abstract representation of the subset of entries of matrix $\mathbf{I}$, denoted as ${}^{pq}\Omega$, whose maximum is assigned to A_{pq} during max-pooling.

because ${}^{pq}_{v}\Omega$ is obtained by stacking the columns of the block, we have [*Note that the sub-script 'b' denotes that the position is relative to the block*]

$$ {}^{pq}r_b = \left\lceil \frac{i_{pq}}{h_l} \right\rceil; \qquad {}^{pq}c_b = i_{pq} - ({}^{pq}r_b - 1)h_l \qquad \text{eq.b-2.1} $$

based on the input-size, pooling-window size, padding, and stride, we can derive the position $({}^{pq}_{t}r_i, {}^{pq}_{t}c_i)$ of the entry of matrix $\mathbf{I}$ that occupies the top-left position in the block corresponding to A_{pq}, i.e. [*Note that the left-sub-script 't' denotes top-left, and the right-sub-script 'i' denotes relative to matrix* I].

$$ {}^{pq}_{t}r_i = 1 + {}^{l}h_s(p-1) - {}^{l}h_p; \qquad {}^{pq}_{t}c_i = 1 + {}^{l}w_s(q-1) - {}^{l}w_p \qquad \text{eq.b-2.2} $$

from eq.b-2.1 and eq.b-2.2, we compute the position $({}^{pq}r_i, {}^{pq}c_i)$ of the entry of the matrix $\mathbf{I}$ that is assigned to Z_{pq}, as follows

$$ {}^{pq}r_i = {}^{pq}r_b + {}^{pq}_{t}r_i - 1; \qquad {}^{pq}c_i = {}^{pq}c_b + {}^{pq}_{t}c_i - 1 \qquad \text{eq.b-2.3} $$

from eq.b-2.3, we then compute the index (or more precisely, the column) of the entry in the row, corresponding to A_{pq} in eq.b-1, which must be set equal to 1, i.e.

$$ \frac{\mathrm{d}A_{pq}}{\mathrm{d}_v\mathbf{I}} = \begin{bmatrix} 0 & 0 & \ldots & 1 & \ldots & 0 \end{bmatrix} $$

where, the 1 in the above row vector is at the position $j = ({}^{pq}c_i - 1)h_i + {}^{pq}r_i$

2. In the row corresponding to A_{pq} (in eq.b-1), set all entries equal to 0 except for the one corresponding to the derivative $\frac{\mathrm{d}A_{pq}}{\mathrm{d}I_{kl}}$, which must be set equal to 1.

The above steps can be vectorized for efficient computation. For more details, see `.\hand-sign-recognizer.ipynb`.

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).