Convolutional Neural Networks

Changyou Chen

Department of Computer Science and Engineering
University at Buffalo, SUNY
changyou@buffalo.edu

March 7, 2019

Some Properties of Convolution

Equivariance of Convolution to Translation

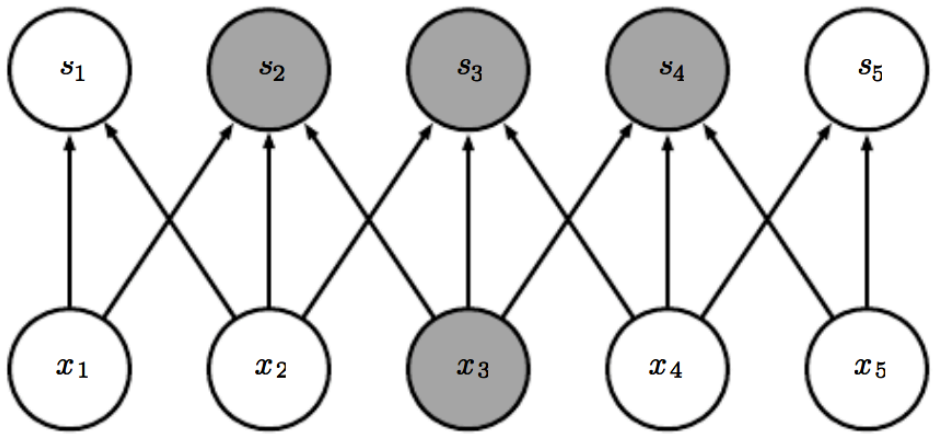
1. The particular form of parameter sharing leads to equivalence to translation:
   - meaning that if the input changes, the output changes in the same way.
2. If g is a function that translates the input, *i.e.*, that shifts it, then the convolution function is equivalent to g:
   - $\mathbf{I}(x, y)$ is image brightness at point (x, y).
   - $\mathbf{I}'(x, y) = g(\mathbf{I}) = \mathbf{I}(x - 1, y)$, *i.e.*, shifts every pixel of $\mathbf{I}$ one unit to the right.
   - If we apply g to $\mathbf{I}$ and then apply convolution, the output will be the same as if we applied convolution to $\mathbf{I}'$, then applied transformation g to the output.
3. Convolution is equivariance to translation.

Question

Is convolutional operator linear or nonlinear?

Convolution as Linear Operator

- Convolution can be viewed as multiplication by a matrix, with some constraints:
  - Construct a matrix $\mathbf{A}$, such that the following 1-D convolution satisfies: $\mathbf{s} = \mathbf{x} * \mathbf{w} = \mathbf{A}\,\mathbf{x}$.

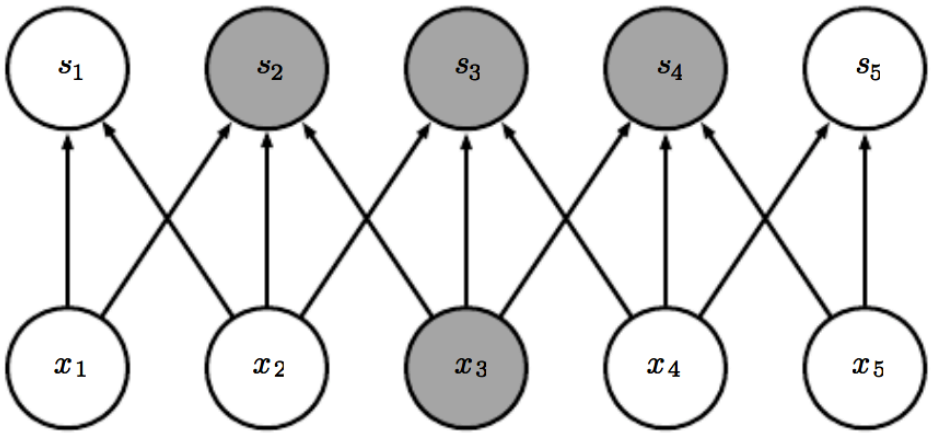


Convolution as Linear Operator

- Convolution can be viewed as multiplication by a matrix, with some constraints:
  - Construct a matrix $\mathbf{A}$, such that the following 1-D convolution satisfies: $\mathbf{s} = \mathbf{x} * \mathbf{w} = \mathbf{A}\,\mathbf{x}$.

$$\mathbf{s} = \underbrace{\begin{bmatrix} w_2 & w_3 & 0 & 0 & 0 \\ w_1 & w_2 & w_3 & 0 & 0 \\ 0 & w_1 & w_2 & w_3 & 0 \\ 0 & 0 & w_1 & w_2 & w_3 \\ 0 & 0 & 0 & w_1 & w_2 \end{bmatrix}}_{\mathbf{A}}$$

  - $\mathbf{A}$ is called the univariate Toeplitz matrix.



Convolution as Linear Operator

- For a 2-D convolution $\mathbf{S} = \mathbf{X} * \mathbf{W} \stackrel{?}{=} \text{mat}\,(\mathbf{A}\,\text{vec}(\mathbf{X}))$
  - "vec" means vectorizing a matrix into a vector row by row; "mat" means reshaping a vector into a matrix of the original size.

$$\underbrace{\mathbf{S}}_{2\times 2} = \underbrace{\begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix}}_{\mathbf{X}:3\times 3} * \underbrace{\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}}_{\mathbf{W}:2\times 2}$$

Convolution as Linear Operator

- For 2-D convolution $\mathbf{S} = \mathbf{X} * \mathbf{W} = \text{mat}\left(\mathbf{A}\,\text{vec}(\mathbf{X})\right)$ ($\mathbf{A}$ is a doubly block circulant matrix):

$$\mathbf{A} = \begin{bmatrix} W_{11} & W_{12} & 0 & W_{21} & W_{22} & 0 & 0 & 0 & 0 \\ 0 & W_{11} & W_{12} & 0 & W_{21} & W_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & W_{11} & W_{12} & 0 & W_{21} & W_{22} & 0 \\ 0 & 0 & 0 & 0 & W_{11} & W_{12} & 0 & W_{21} & W_{22} \end{bmatrix}$$

$$\underbrace{\mathbf{S}}_{2\times 2} = \underbrace{\begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix}}_{\mathbf{X}:3\times 3} * \underbrace{\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}}_{\mathbf{W}:2\times 2}$$

Transposed Convolution

1. Recalled $\mathbf{S} = \mathbf{X} * \mathbf{W} = \text{mat}\left(\mathbf{A}\,\text{vec}(\mathbf{X})\right)$.
2. We can go from a smaller size matrix (*e.g.* $\mathbf{S}$) to a larger size matrix (*e.g.* $\mathbf{X}$) by using the transpose of $\mathbf{A}$:

$$\mathbf{X}' = \text{mat}\left(\mathbf{A}^T \text{vec}(\mathbf{S})\right)$$

3. Sometimes referred to "Deconvolution"[1].

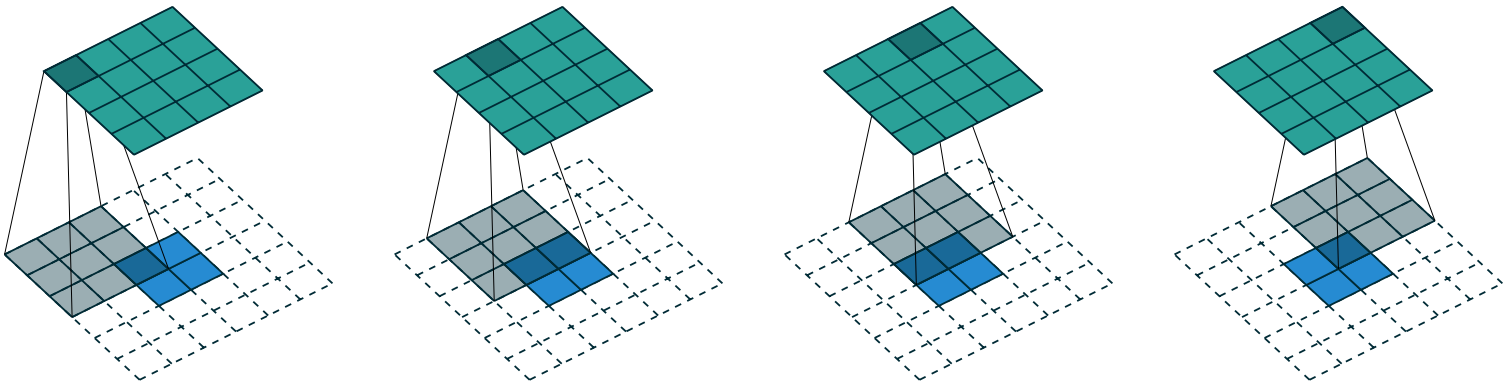
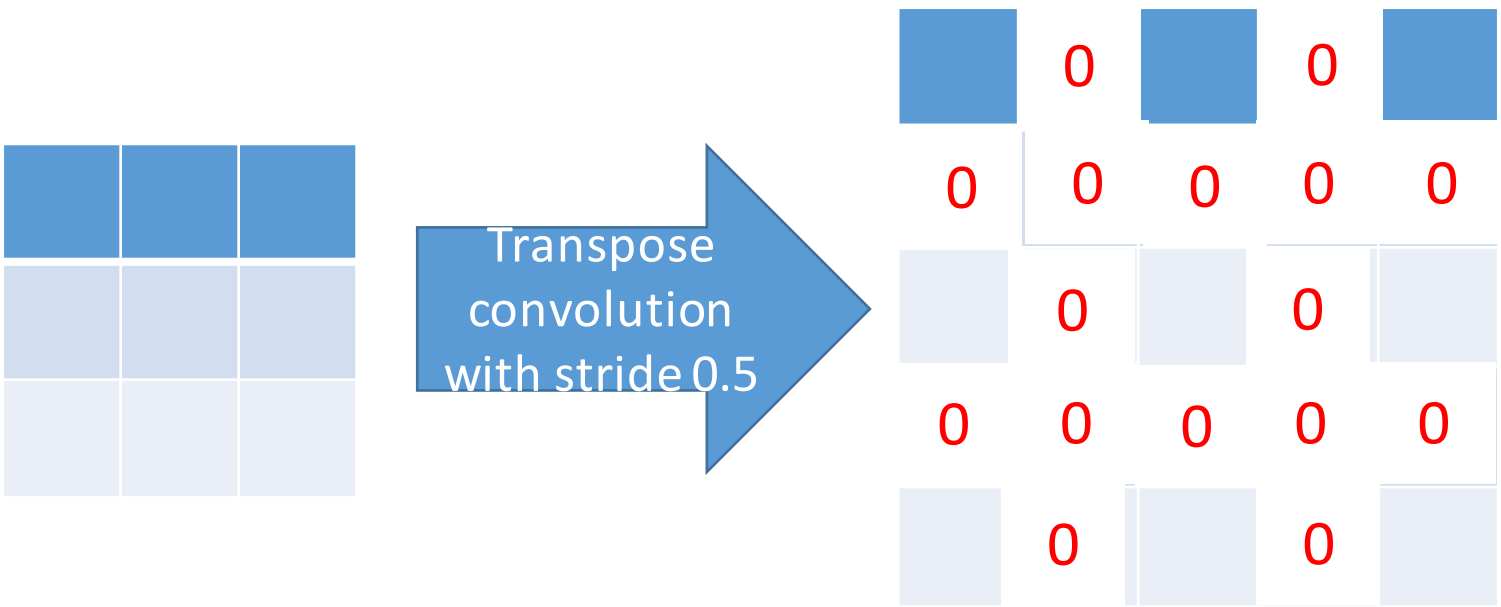


Figure: The transpose of convolving a 3×3 kernel over a 4×4 input using unit strides; Equivalent to convolving a 3×3 kernel over a 2×2 input padded with a 2×2 border of zeros using unit strides.

[1] There are several definitions of deconvolution. This is one of the mostly references.

Stride in Transposed Convolution

1. In convolution, the stride is defined as an integer $s \geq 1$.
2. In transposed convolution, the stride is a fractional number, *i.e.*, $s \in (0, 1]$.
3. Equivalent to adding $(\frac{1}{s} - 1)$ zeros between each two adjacent inputs.



Transposed Convolution

Relationship between Convolution and Transpose Convolution

A convolution with kernel size $k \times k$, stride s and zero-padding p and whose input size $i \times i$ is such that $i + 2p - k$ is a multiple of s has an associated transposed convolution described by $k' = k$, s' and p' and o'. Then the output size of deconvolution can be expressed as

Transposed Convolution

Relationship between Convolution and Transpose Convolution

A convolution with kernel size $k \times k$, stride s and zero-padding p and whose input size $i \times i$ is such that $i + 2p - k$ is a multiple of s has an associated transposed convolution described by $k' = k$, s' and p' and o'. Then the output size of deconvolution can be expressed as

$$o' = \frac{i + 2p - k}{ss'} + \frac{2p' - k + 1}{s'} + 1 \ .$$

Basic Gradient Computation

Notation

- $\mathcal{L}$: the loss function.
- $\mathbf{H}^{(1)}$: first hidden units. $H^{(1)}_{cij}$ indexes position c (typically a 3-dimensional index) within feature map i for example j.
- $\mathbf{H}^{(2)}$: first hidden units. $H^{(2)}_{cij}$ indexes position c (typically a 3-dimensional index) within feature map i for example j.
- $\mathbf{V}$: visible units with the same index-format as $\mathbf{H}$.
- $W^{(1)}$: weights defining the kernel for the first layer. $W^{(2)}_{cij}$ indexes the weight at position c within the kernel, connecting visible channel i to hidden channel j.
- $W^{(2)}$: weights defining the kernel for the second layer.
- Assume all strides to be one, and ignore the biases.

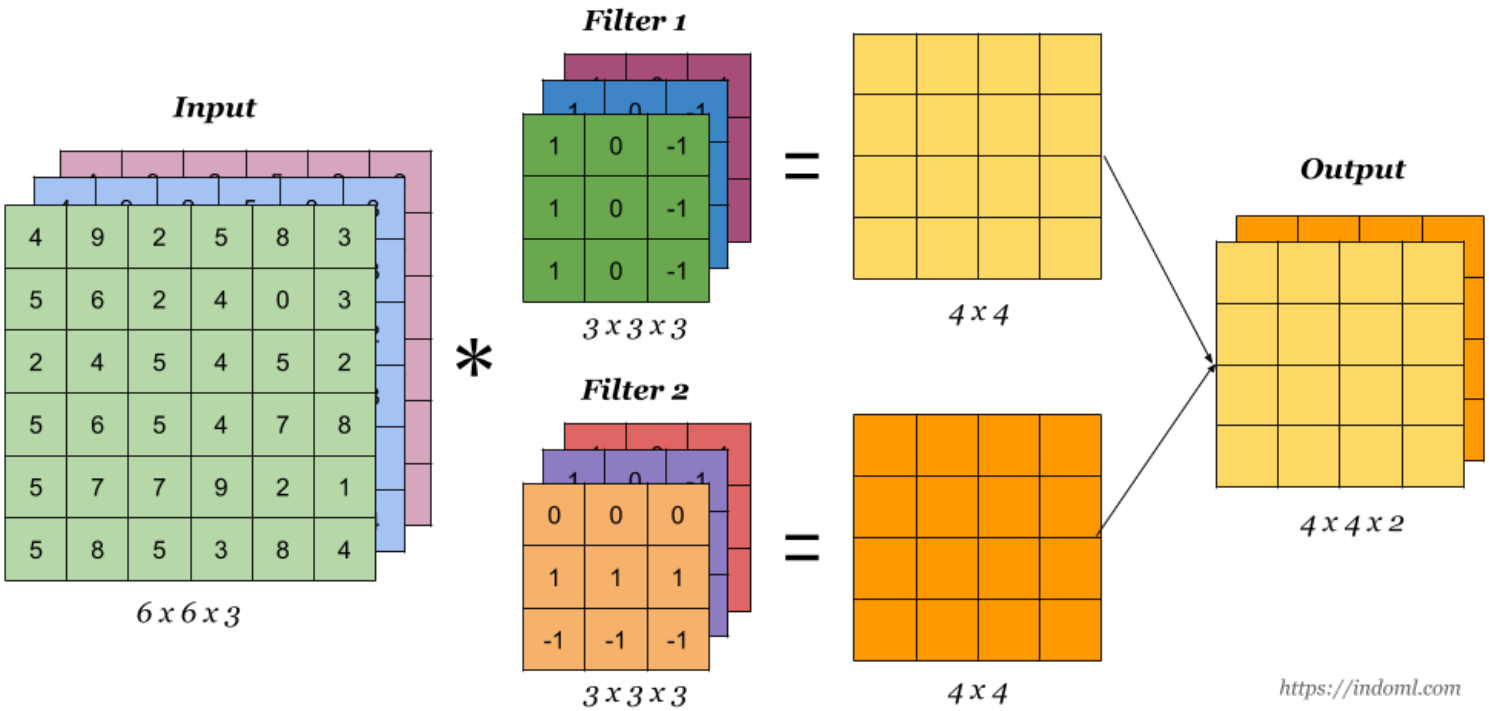
$$H^{(1)}_{cij} = \sum_{k,m} W^{(1)}_{kmi} V_{c+k,m,j}$$

$$H^{(2)}_{cij} = \sum_{k,m} W^{(2)}_{kmi} H^{(1)}_{c+k,m,j}$$

Notation

$$H^{(1)}_{cij} = \sum_{k,m} W^{(1)}_{kmi} V_{c+k,m,j}$$

$$H^{(2)}_{cij} = \sum_{k,m} W^{(2)}_{kmi} H^{(1)}_{c+k,m,j}$$



Basic Gradients

In order to apply BP:

What is the gradients of $\frac{\partial \mathcal{L}}{\partial H^{(i)}_{cij}}$ and $\frac{\partial \mathcal{L}}{\partial W^{(i)}_{cij}}$?

Basic Gradients

The term $\frac{\partial \mathcal{L}}{\partial H^{(1)}_{kjn}}$ is calculated as

$$\frac{\partial \mathcal{L}}{\partial H^{(1)}_{cij}} = \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmn}} \frac{\partial H^{(2)}_{kmn}}{\partial H^{(1)}_{cij}}$$

Basic Gradients

The term $\frac{\partial \mathcal{L}}{\partial H^{(1)}_{kjn}}$ is calculated as

$$\frac{\partial \mathcal{L}}{\partial H^{(1)}_{cij}} = \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmn}} \frac{\partial H^{(2)}_{kmn}}{\partial H^{(1)}_{cij}}$$

Basic Gradients

The term $\frac{\partial \mathcal{L}}{\partial H^{(1)}_{kjn}}$ is calculated as

$$\frac{\partial \mathcal{L}}{\partial H^{(1)}_{cij}} = \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmn}} \frac{\partial H^{(2)}_{kmn}}{\partial H^{(1)}_{cij}}$$

$$= \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmn}} \frac{\partial \sum_{p,q} W^{(2)}_{pqm} H^{(1)}_{k+p,q,n}}{\partial H^{(1)}_{cij}}$$

Basic Gradients

The term $\frac{\partial \mathcal{L}}{\partial H^{(1)}_{kjn}}$ is calculated as

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial H^{(1)}_{cij}} &= \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmn}} \frac{\partial H^{(2)}_{kmn}}{\partial H^{(1)}_{cij}} \\
&= \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmn}} \frac{\partial \sum_{p,q} W^{(2)}_{pqm} H^{(1)}_{k+p,q,n}}{\partial H^{(1)}_{cij}} \\
&= \sum_{k,m} \frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmj}} \frac{\partial \sum_{p,q} W^{(2)}_{pqm} H^{(1)}_{k+p,q,j}}{\partial H^{(1)}_{cij}}
\end{aligned}$$

Basic Gradients

The term $\frac{\partial \mathcal{L}}{\partial H^{(1)}_{kjn}}$ is calculated as

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial H^{(1)}_{cij}} &= \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmn}} \frac{\partial H^{(2)}_{kmn}}{\partial H^{(1)}_{cij}} \\
&= \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmn}} \frac{\partial \sum_{p,q} W^{(2)}_{pqm} H^{(1)}_{k+p,q,n}}{\partial H^{(1)}_{cij}} \\
&= \sum_{k,m} \frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmj}} \frac{\partial \sum_{p,q} W^{(2)}_{pqm} H^{(1)}_{k+p,q,j}}{\partial H^{(1)}_{cij}} \\
&= \sum_{k,m} \frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmj}} W^{(2)}_{c-k,i,j}
\end{aligned}$$

- The term $\frac{\partial \mathcal{L}}{\partial H^{(2)}_{kmn}}$ is backpropagated from last layer.

Basic Gradients

The gradient of the loss function with respect to the weight $W_{cij}^{(1)}$ is given by

Basic Gradients

The gradient of the loss function with respect to the weight $W_{cij}^{(1)}$ is given by

$$\frac{\partial \mathcal{L}}{\partial W_{cij}^{(1)}} = \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H_{kmn}^{(1)}} \frac{\partial H_{kmn}^{(1)}}{\partial W_{cij}^{(1)}}$$

Basic Gradients

The gradient of the loss function with respect to the weight $W^{(1)}_{cij}$ is given by

$$\frac{\partial \mathcal{L}}{\partial W^{(1)}_{cij}} = \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(1)}_{kmn}} \frac{\partial H^{(1)}_{kmn}}{\partial W^{(1)}_{cij}}$$

$$= \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(1)}_{kmn}} \frac{\partial \sum_{p,q} W^{(1)}_{pqm} V_{k+p,q,n}}{\partial W^{(1)}_{cij}}$$

Basic Gradients

The gradient of the loss function with respect to the weight $W^{(1)}_{cij}$ is given by

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial W^{(1)}_{cij}} &= \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(1)}_{kmn}} \frac{\partial H^{(1)}_{kmn}}{\partial W^{(1)}_{cij}} \\
&= \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(1)}_{kmn}} \frac{\partial \sum_{p,q} W^{(1)}_{pqm} V_{k+p,q,n}}{\partial W^{(1)}_{cij}} \\
&= \sum_{k,n} \frac{\partial \mathcal{L}}{\partial H^{(1)}_{kjn}} \frac{\partial \sum_{p,q} W^{(1)}_{pqj} V_{k+p,q,n}}{\partial W^{(1)}_{cij}}
\end{aligned}$$

Basic Gradients

The gradient of the loss function with respect to the weight $W^{(1)}_{cij}$ is given by

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial W^{(1)}_{cij}} &= \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(1)}_{kmn}} \frac{\partial H^{(1)}_{kmn}}{\partial W^{(1)}_{cij}} \\
&= \sum_{k,m,n} \frac{\partial \mathcal{L}}{\partial H^{(1)}_{kmn}} \frac{\partial \sum_{p,q} W^{(1)}_{pqm} V_{k+p,q,n}}{\partial W^{(1)}_{cij}} \\
&= \sum_{k,n} \frac{\partial \mathcal{L}}{\partial H^{(1)}_{kjn}} \frac{\partial \sum_{p,q} W^{(1)}_{pqj} V_{k+p,q,n}}{\partial W^{(1)}_{cij}} \\
&= \sum_{k,n} \frac{\partial \mathcal{L}}{\partial H^{(1)}_{kjn}} V_{k+c,i,n}
\end{aligned}$$