

## Artificial Neural Networks (Gerstner). Solutions for week 8

### Convolutional neural networks

#### Exercise 1. Backprop in convnet

Consider a very simple convolutional neural network with 3 dimensional input (e.g. RGB image), one convolution layer with  $5 \times 5 \times 3$  filters, stride 1, non-linearity  $\sigma$  and one linear layer, i.e.

$$a_{ijk} = b_k + \sum_{x=1}^5 \sum_{y=1}^5 \sum_{c=1}^3 I_{i+x-1, j+y-1, c} w_{xyck}^{(1)} \quad (1)$$

$$x_{ijk}^{(1)} = \sigma(a_{ijk}) \quad (2)$$

$$\hat{y}_o = \sum_{ijk} w_{ijko}^{(2)} x_{ijk}^{(1)} \quad (3)$$

- With loss  $L = \frac{1}{2} \sum_o (t_o - \hat{y}_o)^2$ , compute  $\partial L / \partial w_{1111}$ .
- Compare your result to the one you obtain with a dense (fully-connected) layer, i.e.  $a_k = b_k + \sum_x \sum_y \sum_c I_{xyc} w_{xyck}^{(1)}$ ,  $x_k^{(1)} = \sigma(a_k)$  and  $\hat{y}_o = \sum_k w_{ok}^{(2)} x_k^{(1)}$ .
- How does the result for the convolutional network change, if you introduce a max pooling layer after the convolutional layer?

#### Solution:

a.

$$\partial L / \partial w_{1111}^{(1)} = - \sum_o (t_o - \hat{y}_o) \sum_{ij} w_{ij1o}^{(2)} \sigma'(a_{ij}) I_{ij1} \quad (4)$$

b.

$$\partial L / \partial w_{1111}^{(1)} = - \sum_o (t_o - \hat{y}_o) w_{o1}^{(2)} \sigma'(a_1) I_{111} \quad (5)$$

- The sum does then not run anymore over all  $i$  and  $j$  but only over those that were “winning”, i.e. had the maximal value in the group of hidden activations seen by one max-pooling filter.

#### Exercise 2. Computing volume sizes

Given a volume of width  $n$ , height  $n$  and depth  $c$ .

- Show that the convolution with  $k$  filters of size  $f \times f \times c$  with stride  $s$  and padding  $p$  leads to a new volume of size

$$\left( \frac{n + 2p - f}{s} + 1 \right) \times \left( \frac{n + 2p - f}{s} + 1 \right) \times k. \quad (6)$$

- What padding  $p$  do you have to choose such that the input and output volumes have the same width and depth for stride  $s = 1$ . Check your result for the special case of  $n = 4$  and  $f = 3$ .

#### Solution:

- With padding  $p$  on both sides, the width of the padded image is  $n + 2p$ . If we position the left edge of a filter at the left edge of the padded image – let us call this position  $x = 1$  – the right border of a filter of width  $f$  will be at  $x = f$ . If we position the right edge of the filter at the right

border of the padded image, i.e. at  $x = n + 2p$  the left edge of the filter is at  $x = n + 2p - f + 1$ . With stride 1 there are thus  $n + 2p - f + 1$  different possibilities to position the filter relative to the image. With larger stride  $s > 1$ , we count only every  $s$  possibility for the first  $n + 2p - f$  positions and the last, i.e.  $\frac{n+2p-f}{s} + 1$ . The same reasoning applies for the height and all  $k$  filters. Note that not all choices of  $n, p, f, s$  are reasonable, as there could be fractions.

b.

$$n + 2p - f + 1 = n \quad (7)$$

$$\Rightarrow p = \frac{f - 1}{2} \quad (8)$$

Thus, having odd filter width and height  $f$  allows to have the same padding on the left and the right side. This is one of the reasons, odd filter widths are more common than even ones.

### Exercise 3. Translation equivariance and translation invariance

On the right there is an image we would like to process with a convolutional network with one convolution layer with two 3x3 filters (depicted below the image), relu non-linearity, one 2x2 max-pooling layer (stride 2).

0	0	0	0	0	0
0	1	2	0	3	0
0	1	2	0	1	0
0	4	1	1	2	0
0	0	0	2	1	0
0	0	0	0	0	0

image

0	1	0		0	1	1
1	0	1		0	1	0
0	1	0		1	1	0
filter 1				filter 2		

- Determine the width, height and depth of the volume after max-pooling.
- Compute the output of the max-pooling layer. Use stride 1 and all biases = 0.
- Move the image one pixel to the left (padding column with zeros). How does the output of the convolution layer change? How many output units of the max-pooling layer have changed?
- Is it possible to find for each output unit of the max-pooling layer a translation of the original image under which its value is invariant?

### Solution:

- The volume after convolution is 4x4x2, since there are 4 different positions for the  $x$  and  $y$  direction at which we can place the 3x3 filters and we have 2 filters. After max-pooling with 2x2 filters and stride 2 we are left with a volume of 2x2x2.
- We compute the output of the convolution with the filters and highlight in bold the maximum for each max-pooling filter:

3	3	5	1
<b>7</b>	4	4	<b>5</b>
2	<b>7</b>	<b>5</b>	3
4	3	2	4

output with filter 1

2	5	2	4
8	<b>9</b>	5	<b>7</b>
<b>7</b>	3	4	<b>6</b>
5	2	5	3

output with filter 2

- Because of the equivariance property of the convolution we can move columns 2 to 4 to the left and compute the new rightmost column. Three out of eight output units of the max pooling layer change their value under this transformation.

3	5	1	3
4	4	5	1
7	5	3	2
3	2	4	1

output with filter 1

5	2	4	1
9	5	7	2
3	4	6	1
2	5	3	0

output with filter 2

d. Yes.

#### Exercise 4. Reverse-mode automatic differentiation

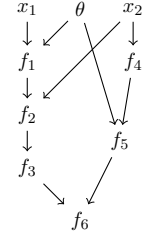
The backprop algorithm derived above is a special case of an algorithm called *Reverse-mode automatic differentiation*. Modern deep learning software packages rely on this algorithm, as it allows flexible exploration of different architectures and cost functions without having to adapt the standard BackProp algorithm for each special case. To understand reverse-mode automatic differentiation we look at the function

$$f(x_1, x_2, \theta) = \sin(\theta x_1 + x_2) + \theta x_2^2. \quad (9)$$

Using the simple functions  $f_1(x, y) = xy$ ,  $f_2(x, y) = x + y$ ,  $f_3(x) = \sin(x)$ ,  $f_4(x) = x^2$ ,  $f_5(x, y) = xy$  and  $f_6(x, y) = x + y$ , we can write the above function as

$$f(x_1, x_2, \theta) = f_6(f_3(f_2(f_1(x_1, \theta), x_2))), f_5(\theta, f_4(x_2))), \quad (10)$$

that has the Abstract Syntax Tree (AST) or Computation Graph depicted on the right.



The reverse-mode Automatic Differentiation algorithm proceeds now as follows:

#### AutoDiff

1. Determine all children of the variable(s) of interest. In the example, using  $\theta$ , this includes  $f_1, f_2, f_3, f_5, f_6$ .
2. Find a reverse ancestral (backwards) schedule of nodes. All of the children of a node should be scheduled before the node itself. In the previous example with  $\theta$ , the schedule could be  $f_6, f_3, f_2, f_1, f_5, \theta$ . For the full graph, this could be  $f_6, f_3, f_2, f_1, x_1, f_5, \theta, f_4, x_2$ .
3. Start with the first node  $n_1$  in the reverse schedule and define  $t_{n_1} = 1$ , e.g.  $t_{f_6} = 1$ .
4. For the next node  $n$  in the reverse schedule, find the child nodes  $\text{ch}(n)$ . Then define

$$t_n = \sum_{c \in \text{ch}(n)} \frac{\partial f_c}{\partial f_n} t_c. \quad (11)$$

5. The total derivative of  $f$  with respect to node  $n$  is given by  $t_n$ .

- a. Show that  $t_\theta = \partial f / \partial \theta$  by following the steps of the algorithm, i.e. by computing  $t_6, t_3, \dots$ , and comparing it to the result you get by differentiating Eq. 9 manually.

- b. Draw the AST of a fully connected network with 2 hidden layers and a single output, and convince yourself that the backpropagation algorithm is a special case of reverse-mode automatic differentiation.
- c. Draw the AST of a network with one 1D convolutional layer (with filter length 3 and stride 1), followed by one max-pooling layer (with  $k = 2$ ), one dense layer, and a single output. For simplicity, assume only 1 filter is used in the convolutional layer.

**Solution:**

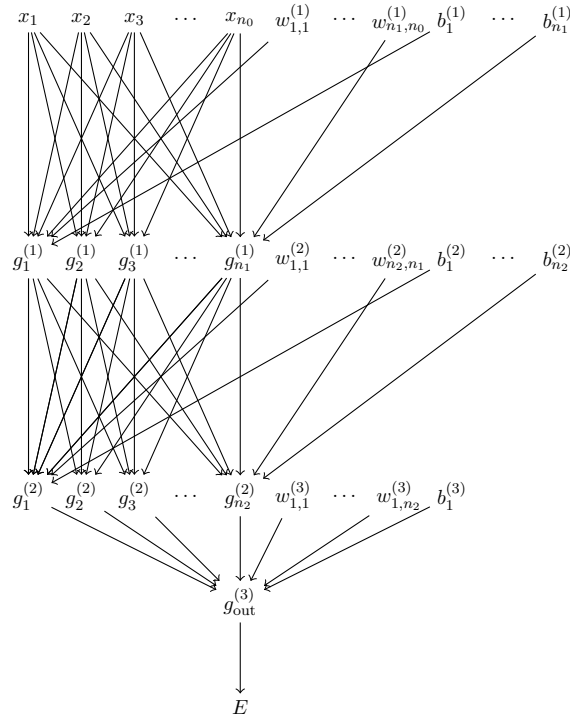
- a. Steps (1) and (2) are completed above. For Steps (3) and (4), we follow the nodes in the reverse schedule in order:

$$\begin{aligned}
 t_{f_6} &= 1 \\
 t_{f_3} &= \frac{\partial f_6}{\partial f_3} t_{f_6} = \frac{\partial(\sin(\theta x_1 + x_2) + \theta x_2^2)}{\partial(\sin(\theta x_1 + x_2))}(1) = 1 \\
 t_{f_2} &= \frac{\partial f_3}{\partial f_2} t_{f_3} = \frac{\partial(\sin(\theta x_1 + x_2))}{\partial(\theta x_1 + x_2)}(1) = \cos(\theta x_1 + x_2) \\
 t_{f_1} &= \frac{\partial f_2}{\partial f_1} t_{f_2} = \frac{\partial(\theta x_1 + x_2)}{\partial(\theta x_1)} \cos(\theta x_1 + x_2) = \cos(\theta x_1 + x_2) \\
 t_{f_5} &= \frac{\partial f_6}{\partial f_5} t_{f_6} = \frac{\partial(\sin(\theta x_1 + x_2) + \theta x_2^2)}{\partial(\theta x_2^2)}(1) = 1 \\
 t_\theta &= \frac{\partial f_1}{\partial \theta} t_{f_1} + \frac{\partial f_5}{\partial \theta} t_{f_5} = \frac{\partial(\theta x_1 + x_2)}{\partial \theta} \cos(\theta x_1 + x_2) + \frac{\partial(\theta x_2^2)}{\partial \theta}(1) \\
 &= x_1 \cos(\theta x_1 + x_2) + x_2^2
 \end{aligned}$$

We see that  $t_\theta$  is equal to the result we would expect from differentiating taking  $t_\theta = \partial f / \partial \theta$  directly. For completeness, evaluating the rest of the graph gives

$$\begin{aligned}
 t_{x_1} &= \frac{\partial f_1}{\partial x_1} t_{f_1} = \frac{\partial(\theta x_1)}{\partial x_1} \cos(\theta x_1 + x_2) = \theta \cos(\theta x_1 + x_2) \\
 t_{f_4} &= \frac{\partial f_5}{\partial f_4} t_{f_5} = \frac{\partial(\theta x_2^2)}{\partial(x_2^2)}(1) = \theta \\
 t_{x_2} &= \frac{\partial f_2}{\partial x_2} t_{f_2} + \frac{\partial f_4}{\partial x_2} t_{f_4} = \frac{\partial(\theta x_1 + x_2)}{\partial x_2} \cos(\theta x_1 + x_2) + \frac{\partial(x_2^2)}{\partial x_2}(\theta) \\
 &= \cos(\theta x_1 + x_2) + 2\theta x_2
 \end{aligned}$$

- b. The AST is shown below.



We consider taking the derivative with respect to  $w_{1,1}^{(1)}$  as an example, which in AutoDiff corresponds to finding  $t_{w_{1,1}^{(1)}}$ . The backwards schedule of nodes to find this derivative is  $E, g_{\text{out}}^{(3)}, g_1^{(2)}, \dots, g_{n_2}^{(2)}, g_1^{(1)}, w_{1,1}^{(0)}$ . We assume quadratic error and denote the target as  $y$ . Where possible, we substitute the formula for  $\delta_i^{(n)}$  from BackProp in the lecture slides. Starting with  $t_E = 1$ ,

$$\begin{aligned}
 t_{g_{\text{out}}^{(3)}} &= \frac{\partial E}{\partial g_{\text{out}}^{(3)}} t_E = -(y - g_{\text{out}}^{(3)}) \\
 t_{g_i^{(2)}} &= \frac{\partial g_{\text{out}}^{(3)}}{\partial g_i^{(2)}} t_{g_{\text{out}}^{(3)}} = -w_{1i}^{(3)} g_{\text{out}}'^{(3)} [a_{\text{out}}^{(3)}] (y - g_{\text{out}}^{(3)}) \\
 &= -w_{1i}^{(3)} \delta_{\text{out}}^{(3)} \\
 t_{g_1^{(1)}} &= \sum_{i=1}^{n_2} \frac{\partial g_i^{(2)}}{\partial g_1^{(1)}} t_{g_i^{(2)}} = -\sum_{i=1}^{n_2} w_{i1}^{(2)} g_i'^{(2)} [a_i^{(2)}] w_{1i}^{(3)} \delta_{\text{out}}^{(3)} \\
 &= -\sum_{i=1}^{n_2} w_{i1}^{(2)} \delta_i^{(2)} \\
 t_{w_{11}^{(1)}} &= \frac{\partial g_1^{(1)}}{\partial w_{11}^{(0)}} t_{g_1^{(1)}} = -x_1 g_1'^{(1)} [a_1^{(1)}] \sum_{i=1}^{n_2} w_{i1}^{(2)} \delta_i^{(2)} \\
 &= -x_1 \delta_1^{(1)}
 \end{aligned}$$

which corresponds to the calculation of the weight update for  $w_{11}^{(1)}$  in BackProp, after multiplying by  $-\eta$  for stochastic gradient descent. By symmetry, the weight update has the same form for any weight in the first layer. Similarly, we can show that the AutoDiff update corresponds to BackProp in the other two layers as well (and, by induction, any layer in an arbitrarily deep network). We therefore conclude that BackProp is a special case of AutoDiff.

- c. The AST is shown below. Note that, in fully connected layers, unit-to-unit connections are dense and weight-to-unit connections are sparse. Conversely, in convolutional layers, unit-to-unit connections are sparse and weight-to-unit connections are dense.

