Deep learning

4.4. Convolutions

François Fleuret

https://fleuret.org/dlc/
If they were handled as normal “unstructured” vectors, large-dimension signals such as sound samples or images would require models of intractable size.

For instance a linear layer taking a $256 \times 256$ RGB image as input, and producing an image of same size would require

$$\left(256 \times 256 \times 3\right)^2 \approx 3.87e+10$$

parameters, with the corresponding memory footprint ($\approx 150$Gb !), and excess of capacity.
Moreover, this requirement is inconsistent with the intuition that such large signals have some “invariance in translation”. A transformation meaningful at a certain location can / should be used everywhere.
Moreover, this requirement is inconsistent with the intuition that such large signals have some “invariance in translation”. A **transformation meaningful at a certain location can / should be used everywhere**.

A convolution layer embodies this idea. **It applies the same linear transformation locally, everywhere**.
Input

\[
\begin{array}{cccccccc}
1 & 4 & -1 & 0 & 2 & -2 & 1 & 3 & 3 & 1 \\
\end{array}
\]

\[W\]
\[ W - w + 1 \]

Input:

\[
\begin{bmatrix}
1 & 4 & -1 & 0 & 2 & -2 & 1 & 3 & 3 & 1
\end{bmatrix}
\]

Kernel:

\[
\begin{bmatrix}
1 & 2 & 0 & -1
\end{bmatrix}
\]

Output:

\[
\begin{bmatrix}
\end{bmatrix}
\]
$w - w + 1$
\[ W - w + 1 \]
Francois Fleuret

Deep learning / 4.4. Convolutions
The input is a 1D array: [1, 4, -1, 0, 2, -2, 1, 3, 3, 1].

The kernel is: [1, 2, 0, -1].

The output is calculated by sliding the kernel over the input array and computing the dot product at each position. The resulting output is: [9, 0, 1, 3].
Input

| 1 | 4 | -1 | 0 | 2 | -2 | 1 | 3 | 3 | 1 |

$W$

Kernel

| 1 | 2 | 0 | -1 |

$w$

Output

| 9 | 0 | 1 | 3 | -5 | -3 |

$W - w + 1$
The diagram illustrates the process of convolution in deep learning. The input tensor is shown on the left, followed by the kernel on the right. The output tensor is calculated by sliding the kernel across the input tensor and performing element-wise multiplication and sum. The formula for the output is given by \( W - w + 1 \), where \( W \) is the input size, and \( w \) is the kernel size. The output tensor size is then \( (W - w + 1) \times (W - w + 1) \).
Input

\[
\begin{bmatrix}
1 & 4 & -1 & 0 & 2 & -2 & 1 & 3 & 3 & 1
\end{bmatrix}
\]

Kernel

\[
\begin{bmatrix}
1 & 2 & 0 & -1
\end{bmatrix}
\]

Output

\[
\begin{bmatrix}
9 & 0 & 1 & 3 & -5 & -3 & 6
\end{bmatrix}
\]

\[W - w + 1\]
Formally, in 1d, given
\[ x = (x_1, \ldots, x_W) \]
and a “convolution kernel” (or “filter”) of width \( w \)
\[ u = (u_1, \ldots, u_w) \]
Formally, in 1d, given

\[ x = (x_1, \ldots, x_W) \]

and a “convolution kernel” (or “filter”) of width \( w \)

\[ u = (u_1, \ldots, u_w) \]

the convolution \( x \ast u \) is a vector of size \( W - w + 1 \), with

\[
(x \ast u)_i = \sum_{j=1}^{w} x_{i-1+j} u_j \\
= (x_i, \ldots, x_{i+w-1}) \cdot u
\]

for instance

\[
(1, 2, 3, 4) \ast (3, 2) = (3 + 4, 6 + 6, 9 + 8) = (7, 12, 17).
\]
Formally, in 1d, given
\[ x = (x_1, \ldots, x_W) \]
and a “convolution kernel” (or “filter”) of width \( w \)
\[ u = (u_1, \ldots, u_w) \]
the convolution \( x \otimes u \) is a vector of size \( W - w + 1 \), with
\[
(x \otimes u)_i = \sum_{j=1}^{w} x_{i-1+j} u_j \\
= (x_i, \ldots, x_{i+w-1}) \cdot u
\]
for instance
\[
(1, 2, 3, 4) \otimes (3, 2) = (3 + 4, 6 + 6, 9 + 8) = (7, 12, 17).
\]
This differs from the usual convolution since the kernel and the signal are both visited in increasing index order.
Convolution can implement in particular differential operators, e.g.

\[(0, 0, 0, 0, 1, 2, 3, 4, 4, 4) \odot (-1, 1) = (0, 0, 0, 1, 1, 1, 1, 0, 0, 0).\]
Convolution can implement in particular differential operators, e.g.

\[(0, 0, 0, 0, 1, 2, 3, 4, 4, 4) \ast (-1, 1) = (0, 0, 0, 1, 1, 1, 0, 0, 0).\]
Convolution can implement in particular differential operators, e.g.

$$(0, 0, 0, 0, 1, 2, 3, 4, 4, 4) \ast (-1, 1) = (0, 0, 0, 1, 1, 1, 0, 0, 0).$$

or crude “template matcher”, e.g.

$$
\begin{array}{ccc}
& & \\
& & \\
& & \\
& & \\
& & \\
& & \\
& & \\
& & \\
\end{array}
\ast

\begin{array}{ccc}
& & \\
& & \\
& & \\
& & \\
& & \\
& & \\
& & \\
\end{array}
=

\begin{array}{ccc}
& & \\
& & \\
& & \\
& & \\
& & \\
& & \\
& & \\
\end{array}
$$
It generalizes naturally to a multi-dimensional input, although specification can become complicated.
It generalizes naturally to a multi-dimensional input, although specification can become complicated.

Its most usual form for “convolutional networks” processes a 3d tensor as input (i.e. a multi-channel 2d signal) to output a 2d tensor. The kernel is not swiped across channels, just across rows and columns.

In this case, if the input tensor is of size $C \times H \times W$, and the kernel is $C \times h \times w$, the output is $(H - h + 1) \times (W - w + 1)$. 
It generalizes naturally to a multi-dimensional input, although specification can become complicated.

Its most usual form for “convolutional networks” processes a 3d tensor as input (i.e. a multi-channel 2d signal) to output a 2d tensor. The kernel is not swiped across channels, just across rows and columns.

In this case, if the input tensor is of size $C \times H \times W$, and the kernel is $C \times h \times w$, the output is $(H - h + 1) \times (W - w + 1)$.

We say “2d signal” even though it has $C$ channels, since it is a feature vector indexed by a 2d location without structure on the feature indexes.
It generalizes naturally to a multi-dimensional input, although specification can become complicated.

Its most usual form for “convolutional networks” processes a 3d tensor as input (i.e. a multi-channel 2d signal) to output a 2d tensor. The kernel is not swiped across channels, just across rows and columns.

In this case, if the input tensor is of size $C \times H \times W$, and the kernel is $C \times h \times w$, the output is $(H - h + 1) \times (W - w + 1)$.

⚠️ We say “2d signal” even though it has $C$ channels, since it is a feature vector indexed by a 2d location without structure on the feature indexes.

In a standard convolution layer, $D$ such convolutions are combined to generate a $D \times (H - h + 1) \times (W - w + 1)$ output.
Kernel $K$ of dimension $D 	imes H 	imes W$ and $C$ channels.
Kernel

\[ D \times H - h + 1 \times W - w + 1 \times 1 \]
\[ D - h + 1 \quad W - w + 1 \]
Kernel

\[ D - h + 1 \]

\[ W - w + 1 \]

Input

Output

Kernel

\[ h \]

\[ w \]

\[ C \]
\[ \text{Input} \]

\[ \begin{align*}
H & \quad w \\
W & \quad h \\
C & \quad w \\
C & \quad h
\end{align*} \]

\[ \text{Kernel} \]

\[ \begin{align*}
D & \quad h + 1 \\
W & \quad w + 1
\end{align*} \]

\[ \text{Output} \]
Kernel

\[ D - h + 1 \]

\[ W - w + 1 \]

François Fleuret

Deep learning / 4.4. Convolutions
Kernel $K$:

$$D - h + 1 \times W - w + 1$$
Kernel

\[ \begin{align*}
    D & - h + 1 \\
    W & - w + 1 \\
    C & - h + 1
\end{align*} \]
Kernels

\[ D - h + 1 \]

\[ W - w + 1 \]

\[ 1 \]
Kernel

Input

Output

\[ \text{H} - \text{h} + 1 \]

\[ \text{W} - \text{w} + 1 \]

\[ \text{C} \]

\[ \text{C} \]

\[ \text{W} \]

\[ \text{H} \]
Francois Fleuret

Deep learning / 4.4. Convolutions

Input

\[ D \times H \times W \]

Kernels

\[ D \times h \times w \]

Output

\[ W - w + 1 \times H - h + 1 \]
A convolution preserves the signal support structure: a 1d signal is converted into a 1d signal, a 2d signal into a 2d, and neighboring parts of the input signal influence neighboring parts of the output signal.
A convolution *preserves the signal support structure*: a 1d signal is converted into a 1d signal, a 2d signal into a 2d, and neighboring parts of the input signal influence neighboring parts of the output signal.

And a convolution is *equivariant* to a translation of the input signal, since its output is translated similarly.
A convolution **preserves the signal support structure**: a 1d signal is converted into a 1d signal, a 2d signal into a 2d, and neighboring parts of the input signal influence neighboring parts of the output signal.

And a convolution is **equivariant** to a translation of the input signal, since its output is translated similarly.

A 3d convolution can be used if the channel index has some metric meaning, such as time for a series of grayscale video frames. Otherwise swiping across channels makes no sense.
We usually refer to one of the channels generated by a convolution layer as an activation map.

The sub-area of an input map that influences a component of the output as the receptive field of the latter.
We usually refer to one of the channels generated by a convolution layer as an activation map.

The sub-area of an input map that influences a component of the output as the receptive field of the latter.

In the context of convolutional networks, a standard linear layer is called a fully connected layer, or a dense layer, since every input influences every output.
The autograd-compliant function

\[
F.\text{conv2d}(\text{input}, \text{weight}, \text{bias}=\text{None}, \text{stride}=1, \text{padding}=0, \text{dilation}=1, \text{groups}=1)
\]

Implements a 2d convolution, where \text{weight} is of dimension \( D \times C \times h \times w \) and contains the kernels, \text{bias} is of dimension \( D \), \text{input} is of dimension

\[
N \times C \times H \times W
\]

and the result is of dimension

\[
N \times D \times (H - h + 1) \times (W - w + 1).
\]
The autograd-compliant function

F.conv2d(input, weight, bias=None, stride=1, padding=0, dilation=1, groups=1)

Implements a 2d convolution, where weight is of dimension $D \times C \times h \times w$
and contains the kernels, bias is of dimension $D$,
input is of dimension $N \times C \times H \times W$

and the result is of dimension

$N \times D \times (H - h + 1) \times (W - w + 1)$.

>>> weight = torch.randn(5, 4, 2, 3)
>>> bias = torch.randn(5)
>>> input = torch.randn(117, 4, 10, 3)
>>> output = F.conv2d(input, weight, bias)
>>> output.size()
torch.Size([117, 5, 9, 1])
The autograd-compliant function

\[ F.\text{conv2d}(\text{input}, \text{weight}, \text{bias}=\text{None}, \text{stride}=1, \text{padding}=0, \text{dilation}=1, \text{groups}=1) \]

Implements a 2d convolution, where \( \text{weight} \) is of dimension \( D \times C \times h \times w \) and contains the kernels, \( \text{bias} \) is of dimension \( D \), \( \text{input} \) is of dimension \( N \times C \times H \times W \) and the result is of dimension

\[ N \times D \times (H - h + 1) \times (W - w + 1). \]

```python
>>> weight = torch.randn(5, 4, 2, 3)
>>> bias = torch.randn(5)
>>> input = torch.randn(117, 4, 10, 3)
>>> output = F.conv2d(input, weight, bias)
>>> output.size()
torch.Size([117, 5, 9, 1])
```

Similar functions implement 1d and 3d convolutions.
```python
x = mnist_train.data[12].float().view(1, 1, 28, 28)

weight = torch.empty(5, 1, 3, 3)

weight[0, 0] = torch.tensor(
    [[0., 0., 0.],
     [0., 1., 0.],
     [0., 0., 0.]]
)

weight[1, 0] = torch.tensor(
    [[1., 1., 1.],
     [1., 1., 1.],
     [1., 1., 1.]]
)

weight[2, 0] = torch.tensor(
    [[-1., 0., 1.],
     [-1., 0., 1.],
     [-1., 0., 1.]]
)

weight[3, 0] = torch.tensor(
    [[-1., -1., -1.],
     [0., 0., 0.],
     [1., 1., 1.]]
)

weight[4, 0] = torch.tensor(
    [[0., -1., 0.],
     [-1., 4., -1.],
     [0., -1., 0.]]
)

y = F.conv2d(x, weight)
```
François Fleuret

Deep learning / 4.4. Convolutions
class torch.nn.Conv2d(in_channels, out_channels,
    kernel_size, stride=1, padding=0, dilation=1,
    groups=1, bias=True)

Wraps the convolution into a Module, with the kernels and biases as Parameter properly randomized at creation.

The kernel size is either a pair \((h, w)\) or a single value \(k\) interpreted as \((k, k)\).
class torch.nn.Conv2d(in_channels, out_channels,
                 kernel_size, stride=1, padding=0, dilation=1,
                 groups=1, bias=True)

Wraps the convolution into a Module, with the kernels and biases as Parameter properly randomized at creation.

The kernel size is either a pair \((h, w)\) or a single value \(k\) interpreted as \((k, k)\).

```python
>>> f = nn.Conv2d(in_channels = 4, out_channels = 5, kernel_size = (2, 3))
>>> for n, p in f.named_parameters(): print(n, p.size())
... weight torch.Size([5, 4, 2, 3])
   bias torch.Size([5])
>>> x = torch.randn(117, 4, 10, 3)
>>> y = f(x)
>>> y.size()
torch.Size([117, 5, 9, 1])
```
Padding, stride, and dilation
Convolutions have three additional parameters:

- The **padding** specifies the size of a zeroed frame added around the input,
- the **stride** specifies a step size when moving the kernel across the signal,
- the **dilation** modulates the expansion of the filter without adding weights.
Here with $C \times 3 \times 5$ as input
Here with $C \times 3 \times 5$ as input, a padding of $(2,1)$
Here with $C \times 3 \times 5$ as input, a padding of $(2, 1)$, a stride of $(2, 2)$
Here with $C \times 3 \times 5$ as input, a padding of $(2, 1)$, a stride of $(2, 2)$, and a kernel of size $C \times 3 \times 3$.
Here with $C \times 3 \times 5$ as input, a padding of $(2, 1)$, a stride of $(2, 2)$, and a kernel of size $C \times 3 \times 3$. 
Here with $C \times 3 \times 5$ as input, a padding of $(2, 1)$, a stride of $(2, 2)$, and a kernel of size $C \times 3 \times 3$.
Here with $C \times 3 \times 5$ as input, a padding of $(2, 1)$, a stride of $(2, 2)$, and a kernel of size $C \times 3 \times 3$. 

Input

```
  . . .
0 0 0
0 0 0
0 0 0
```

Output

```
  . .
0 0
```

Input

```
  . . .
  . . .
  . . .
```

Output

```
  .
```

François Fleuret

Deep learning / 4.4. Convolutions
Here with $C \times 3 \times 5$ as input, a padding of $(2,1)$, a stride of $(2,2)$, and a kernel of size $C \times 3 \times 3$.
Here with $C \times 3 \times 5$ as input, a padding of $(2,1)$, a stride of $(2,2)$, and a kernel of size $C \times 3 \times 3$. 

![Diagram of convolution process]

Input

Output
Here with $C \times 3 \times 5$ as input, a padding of $(2, 1)$, a stride of $(2, 2)$, and a kernel of size $C \times 3 \times 3$. 

Input

Output
Here with $C \times 3 \times 5$ as input, a padding of $(2,1)$, a stride of $(2,2)$, and a kernel of size $C \times 3 \times 3$
Here with $C \times 3 \times 5$ as input, a padding of $(2, 1)$, a stride of $(2, 2)$, and a kernel of size $C \times 3 \times 3$, the output is $1 \times 3 \times 3$. 

Input

Output
A convolution with a stride greater than 1 may not cover the input map entirely, hence may ignore some of the input values.
The dilation modulates the expansion of the filter support by adding rows and columns of zeros between coefficients (Yu and Koltun, 2015).

It is 1 for standard convolutions, but can be greater, in which case the resulting operation can be envisioned as a convolution with a regularly sparsified filter.
The dilation modulates the expansion of the filter support by adding rows and columns of zeros between coefficients (Yu and Koltun, 2015).

It is 1 for standard convolutions, but can be greater, in which case the resulting operation can be envisioned as a convolution with a regularly sparsified filter.

This notion comes from signal processing, where it is referred to as algorithme à trous, hence the term sometime used of “convolution à trous”.
Dilation = 1
Dilation = 1
Dilation = 1

Francois Fleuret Deep learning / 4.4. Convolutions
Dilation = 1

Francois Fleuret

Deep learning / 4.4. Convolutions
Dilation = 1
Dilation = 1
Dilation = 1
Dilation = 1

Input

Output
Dilation = 2

Input

Output
Dilation = 2
Dilation = 2

Input

Output
Dilation = 2

Input

Output
Dilation = 2
Dilation = 2
Dilation = 2

Input

Output
Dilation = 2

Francois Fleuret

Deep learning / 4.4. Convolutions

Input

Output
Dilation = 2

François Fleuret

Input

Output

Deep learning / 4.4. Convolutions
A 1d convolution with a kernel of size $k$ and dilation $d$ can be interpreted as a convolution with a filter of size $1 + (k - 1)d$ with only $k$ non-zero coefficients.

For example with $k = 3$ and $d = 4$, the difference between the input map size and the output map size is $1 + (3 - 1)4 - 1 = 8$.

```python
>>> x = torch.randn(1, 1, 20, 30)
>>> l = nn.Conv2d(1, 1, kernel_size = 3, dilation = 4)
>>> l(x).size()
torch.Size([1, 1, 12, 22])
```
Having a dilation greater than one increases the units’ receptive field size without increasing the number of parameters.

**Convolutions with stride or dilation strictly greater than one reduce the activation map size, for instance to make a final classification decision.**
The End
References