**Convolution operations**

Convolution is an operation on two function of a real valued argument.

The conv operation is usually denoted with an asterisk

In the case of indiscrete value

s(t)=(x∗w)(t)=∫x(a)w(t−a)da

In the case of dsicrete value

s(t)=(x∗w)(t)=∑a=−∞∞x(a)w(t−a)

Note: w needs to be 0 for all negative arguments

* x: input
* w: kernel
* s: feature map

For 2D input and 2D kernel:

S(i,j)=(I∗K)(i,j)=∑m∑nI(m,n)K(i−m,j−n)�(�,�)=(�∗�)(�,�)=∑�∑��(�,�)�(�−�,�−�)

It is commutative, so we can also right

S(i,j)=(K∗I)(i,j)=∑m∑nI(i−m,j−n)K(m,n)�(�,�)=(�∗�)(�,�)=∑�∑��(�−�,�−�)�(�,�)

Cross-coorelation:

S(i,j)=(I∗K)(i,j)=∑m∑nI(i+m,j+n)K(m,n)

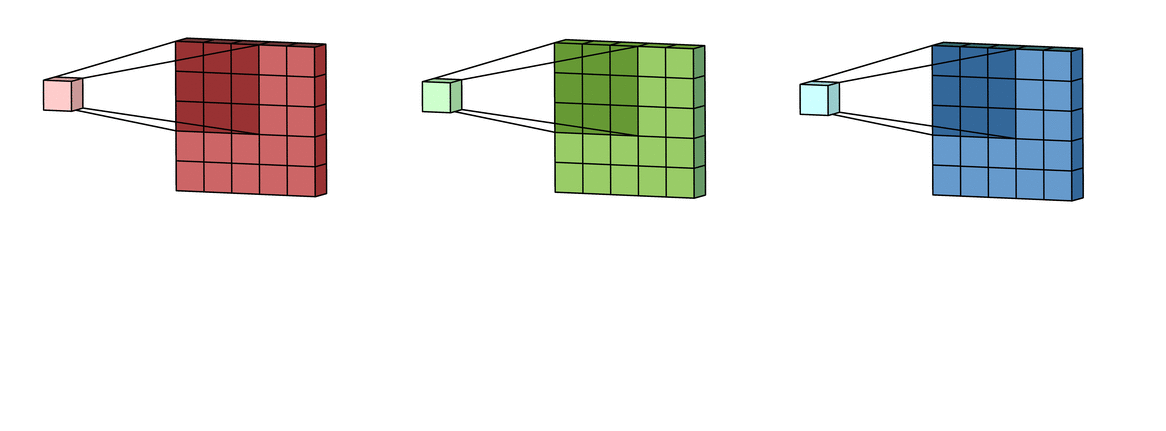
The name “Convolutional neural network” indicates that the network employs a mathematical operation called Convolution. Convolution is a specialized kind of linear operation. Convnets are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.

*Convolution between two functions in mathematics produces a third function expressing how the shape of one function is modified by other*

## Convolution Kernels

A kernel is a small 2D matrix whose contents are based upon the operations to be performed. A kernel maps on the input image by simple matrix multiplication and addition, the output obtained is of lower dimensions and therefore easier to work with.

**Stride**defines by what step does to kernel move, for example stride of 1 makes kernel slide by one row/column at a time and stride of 2 moves kernel by 2 rows/columns.



what is pooling and types its types

In machine learning and computer vision, pooling is a technique that reduces the spatial dimensions (width and height) of feature maps. Pooling is typically applied after convolutional layers in a neural network and is used to extract and preserve the most important features while reducing the number of parameters.

There are several types of pooling techniques, including:

1. Max pooling: This type of pooling takes the maximum value from a rectangular region of the feature map. It is the most commonly used type of pooling.
2. Average pooling: This type of pooling takes the average value from a rectangular region of the feature map. It is less commonly used than max pooling but can be effective in certain situations.
3. Global pooling: This type of pooling takes the maximum or average value over the entire feature map, resulting in a single value for each feature map channel. It is often used in the final layers of a neural network to aggregate the features.
4. Lp pooling: This type of pooling takes the Lp norm of a rectangular region of the feature map, where p is a positive integer. When p=2, it is called L2 pooling, which takes the square root of the sum of squares of the values in the region.
5. Fractional pooling: This type of pooling takes a fractional part of each value in the feature map. It is useful for reducing the spatial dimensions of feature maps with non-integer factors.

Variants of the Basic Convolution Function

The basic convolution function in deep learning is typically defined as a linear transformation that operates on a small patch of the input image at a time, followed by a non-linear activation function. However, there are several variants of this basic convolution function that have been developed to improve the performance of deep convolutional neural networks. Some of the most commonly used variants include:

1. Dilated convolution: In dilated convolution, the convolutional kernel is applied to the input with some spacing between the kernel elements, resulting in a larger receptive field. Dilated convolution can capture more global information in the input and is commonly used in tasks such as semantic segmentation.

2. Depthwise convolution: In depthwise convolution, the input is convolved with a separate kernel for each input channel, resulting in a set of output channels that represent the spatial response of each input channel. Depthwise convolution is computationally efficient and is commonly used in mobile and embedded applications.

3. Separable convolution: Separable convolution is a combination of depthwise convolution and pointwise convolution (convolution with a 1x1 kernel). The input is first convolved depthwise to capture spatial information, and then convolved with a 1x1 kernel to perform feature aggregation across channels. Separable convolution is computationally efficient and is commonly used in low-power devices.

4. Grouped convolution: In grouped convolution, the input is split into several groups, each of which is convolved with a separate kernel. The outputs of the groups are then concatenated to produce the final output. Grouped convolution can improve the expressiveness of deep convolutional neural networks by allowing different groups to learn different features.

5. Transposed convolution: Transposed convolution, also known as deconvolution, is used to perform upsampling of the feature maps by using a larger kernel size than the input size. Transposed convolution is commonly used in tasks such as image segmentation, where the output resolution needs to be higher than the input resolution.

Overall, these variants of the basic convolution function provide a flexible and powerful toolset for building deep convolutional neural networks that can capture complex spatial patterns in the input data.

what is LeNet

*LeNet is a convolutional neural network architecture that was proposed by Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner in 1998. It was one of the first convolutional neural network architectures to achieve good performance on handwritten digit recognition tasks and helped to establish convolutional neural networks as a powerful tool for computer vision tasks.*

*The LeNet architecture consists of seven layers, including two convolutional layers, two pooling layers, and three fully connected layers. The input to the network is a grayscale image of size 32x32 pixels.*

*The first layer in the network is a convolutional layer with 6 filters of size 5x5. The output of this layer is passed through a sigmoid activation function and then through a 2x2 max pooling layer.*

*The second layer is another convolutional layer with 16 filters of size 5x5. This layer is also followed by a sigmoid activation function and a 2x2 max pooling layer.*

*The third layer is a fully connected layer with 120 units, followed by a sigmoid activation function. The output of this layer is then passed through another fully connected layer with 84 units and a sigmoid activation function.*

*The final layer is a fully connected layer with 10 units, corresponding to the 10 possible classes of handwritten digits (0-9). The output of this layer is passed through a softmax activation function to obtain the class probabilities.*

*Overall, the LeNet architecture was a significant breakthrough in the field of computer vision and helped to establish the effectiveness of convolutional neural networks for image recognition tasks. It has since been surpassed by more complex architectures such as AlexNet, VGG, and ResNet, but it remains an important milestone in the history of deep learning.*

what is ZF-Net

ZF-Net (Zeiler and Fergus Net) is a convolutional neural network architecture that was proposed by Matthew D. Zeiler and Rob Fergus in 2013. It won the top prize in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2013, outperforming the previous year's winning architecture, AlexNet.

ZF-Net architecture is very similar to AlexNet, but with some minor differences. The architecture consists of eight layers, including five convolutional layers, two fully connected layers, and one output layer. The input to the network is a color image of size 224x224 pixels.

The first layer in the network is a convolutional layer with 96 filters of size 7x7, followed by a rectified linear unit (ReLU) activation function and a 2x2 max pooling layer.

The second layer is another convolutional layer with 256 filters of size 5x5, followed by a ReLU activation function and a 2x2 max pooling layer.

The third, fourth, and fifth layers are all convolutional layers with 384, 384, and 256 filters, respectively. Each layer is followed by a ReLU activation function.

The sixth layer is a fully connected layer with 4096 units, followed by a ReLU activation function and a 50% dropout layer.

The seventh layer is another fully connected layer with 4096 units, followed by a ReLU activation function and a 50% dropout layer.

The final layer is a fully connected layer with 1000 units, corresponding to the 1000 possible classes in the ImageNet dataset. The output of this layer is passed through a softmax activation function to obtain the class probabilities.

One of the key differences between ZF-Net and AlexNet is that ZF-Net uses a smaller stride in the first convolutional layer (2 instead of 4), which helps to preserve more spatial information in the input image. Additionally, ZF-Net uses a smaller filter size in the first convolutional layer (7x7 instead of 11x11), which helps to reduce the number of parameters in the network.

Overall, ZF-Net was an important milestone in the development of convolutional neural networks for image classification tasks, as it demonstrated that small changes to the architecture can have a significant impact on performance.

what is GooLeNet

GoogLeNet, also known as Inception v1, is a convolutional neural network architecture that was proposed by researchers at Google in 2014. It was designed to improve upon the previous state-of-the-art networks for image classification, including VGGNet and AlexNet, by using a more efficient architecture with lower computational cost.

GoogLeNet architecture consists of a series of "Inception modules", which are designed to capture information at different scales and abstraction levels. Each Inception module consists of multiple convolutional layers with different filter sizes, followed by max pooling and concatenation of the output feature maps. This allows the network to capture both local and global information in the input image.

The architecture also uses 1x1 convolutions, which help to reduce the dimensionality of the input feature maps and improve computational efficiency. Additionally, it uses an auxiliary classifier at intermediate layers, which encourages the network to learn features that are useful for classification even before reaching the final output layer.

The final layer of the network is a fully connected layer with 1000 units corresponding to the 1000 possible classes in the ImageNet dataset. The output of this layer is passed through a softmax activation function to obtain the class probabilities.

GoogLeNet achieved state-of-the-art performance on the ImageNet dataset, with an error rate of 6.67% in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014, which was the best performance at the time. Its efficient architecture also paved the way for the development of even more complex and efficient networks, such as ResNet and MobileNet.

What is ResNet

ResNet (short for "Residual Network") is a convolutional neural network architecture that was proposed by researchers at Microsoft Research in 2015. It was designed to improve upon the previous state-of-the-art networks for image classification, including GoogLeNet and VGGNet, by allowing for deeper architectures without the problem of vanishing gradients.

The key idea behind ResNet is the use of residual blocks, which allow for the training of very deep neural networks. A residual block consists of multiple convolutional layers with skip connections that allow the input of a layer to be added directly to the output of a later layer. This effectively creates a "shortcut" path that allows the gradient to flow more easily through the network and avoids the problem of vanishing gradients.

The ResNet architecture consists of a series of residual blocks with increasing numbers of filters, followed by global average pooling and a fully connected layer with 1000 units corresponding to the 1000 possible classes in the ImageNet dataset.

ResNet achieved state-of-the-art performance on the ImageNet dataset, with an error rate of 3.57% in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015, which was significantly better than the previous state-of-the-art. It also paved the way for the development of even deeper and more complex architectures, such as ResNet-101 and ResNet-152, which have been shown to achieve even better performance on a variety of computer vision tasks.