Pontificia Universidad JAVERIANA Colombia

Tópicos Avanzados en Analítica

Maestría en Analítica para la Inteligencia de Negocios

Sergio Alberto Mora Pardo - H2 2024

Deep Computer Vision
Convolutional Neural Networks

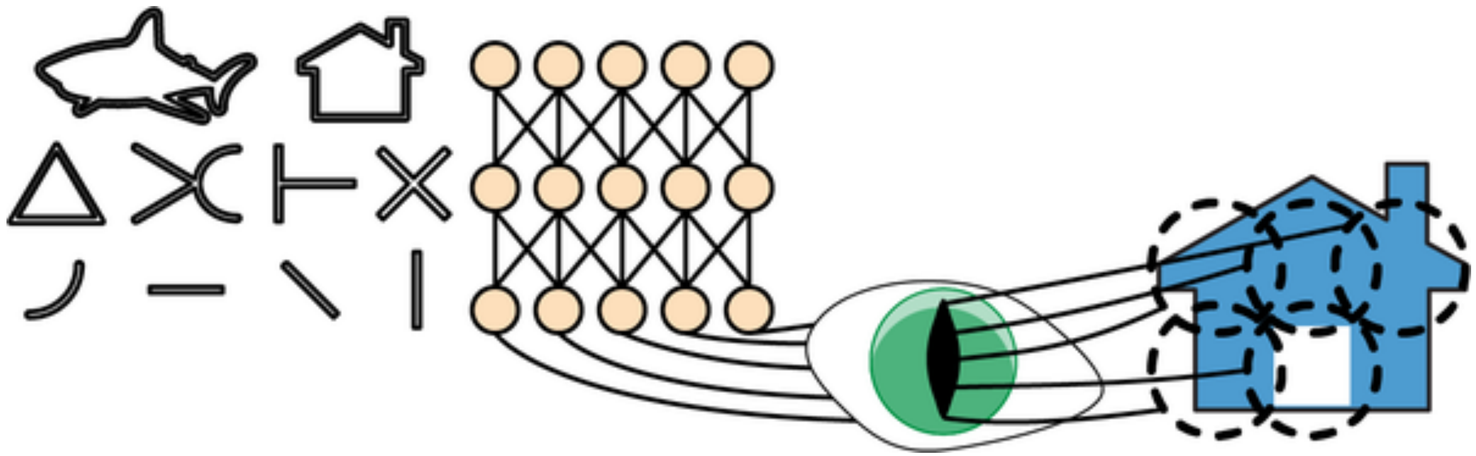
Pontificia Universidad
JAVERIANA
Colombia

Deep Computer Vision

Convolutional Neural Network

Visual Cortex

Biological neurons in the visual cortex respond to specific patterns in small regions of the visual field called receptive fields.



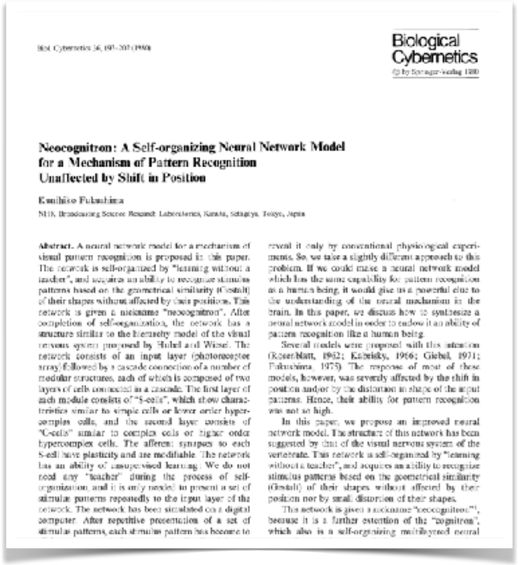
Pontificia Universidad JAVERIANA Colombia

Deep Computer Vision

Convolutional Neural Network

Visual Cortex

Biological neurons in the visual cortex respond to specific patterns in small regions of the visual field called receptive fields.



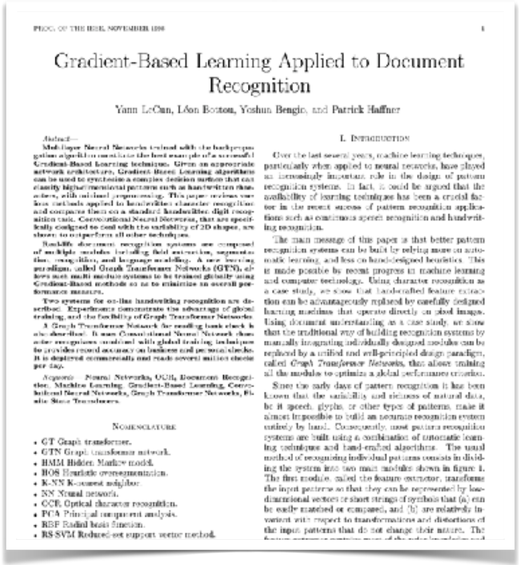
Biological Cybernetics

Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position

Kunihiko Fukushima

Kunihiko Fukushima

Spring 1980



Gradient-Based Learning Applied to Document Recognition

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

Yann Le Cun et al.

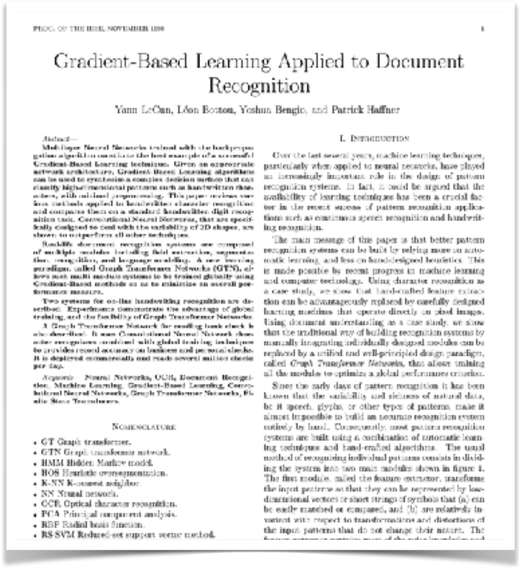
November 1998

Deep Computer Vision

Convolutional Neural Network

Visual Cortex

Biological neurons in the visual cortex respond to specific patterns in small regions of the visual field called receptive fields.



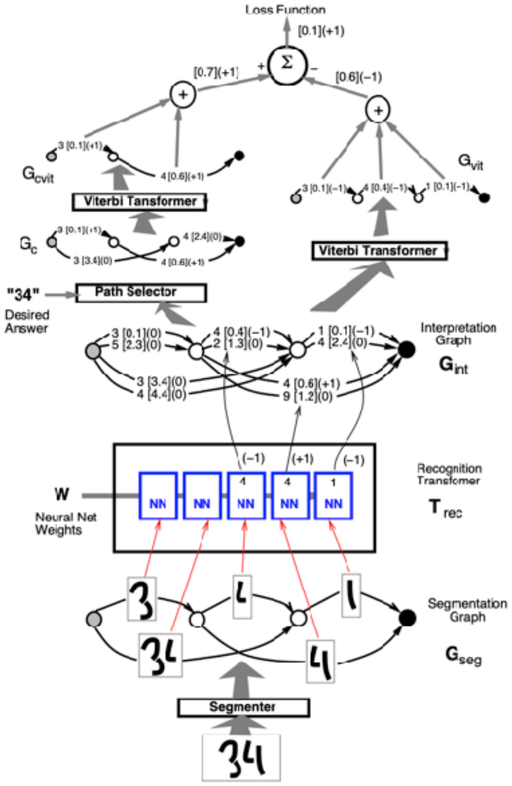
Gradient-Based Learning Applied to Document Recognition

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

Yann Le Cun et al.

November 1998

LeNet-5

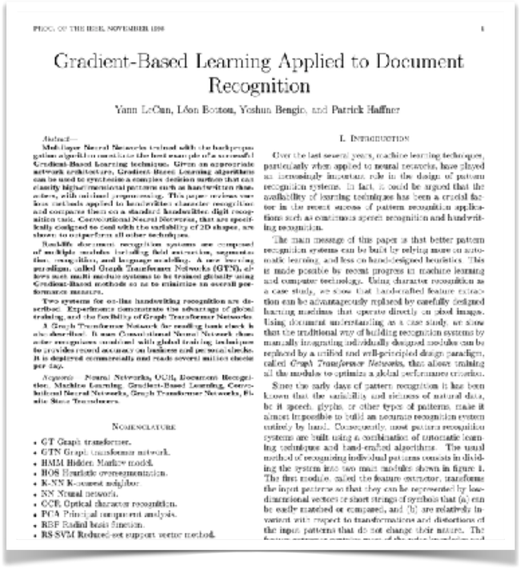


Deep Computer Vision

Convolutional Neural Network

Visual Cortex

Biological neurons in the visual cortex respond to specific patterns in small regions of the visual field called receptive fields.



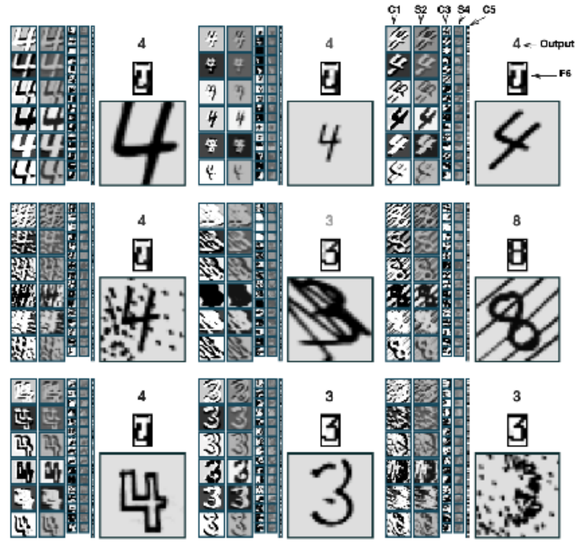
Gradient-Based Learning Applied to Document Recognition

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

Yann Le Cun et al.

November 1998

LeNet-5



Deep Computer Vision

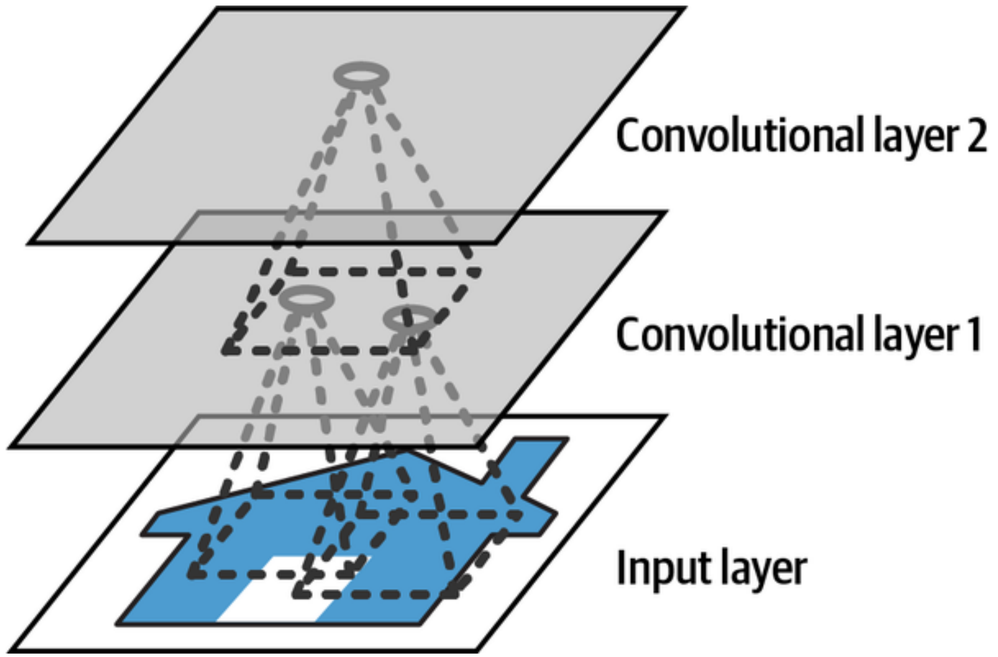
Convolutional Neural Network

Convolutional Layer

Each neuron in the second convolutional layer is connected only to neurons located within a small rectangle in the first layer.

Then assemble them into larger, higher-level features in the next hidden layer,

This architecture allows the network to focus on small, low-level features in the first hidden layer,



Deep Computer Vision

Convolutional Neural Network

Convolutional Layer

Each neuron in the second convolutional layer is connected only to neurons located within a small rectangle in the first layer.

Then assemble them into larger, higher-level features in the next hidden layer,

Padding:

$f_h : height$

$i \quad to \quad i + f_h - 1$

Rows

$f_w : width$

$j \quad to \quad j + f_w - 1$

Columns

Deep Computer Vision

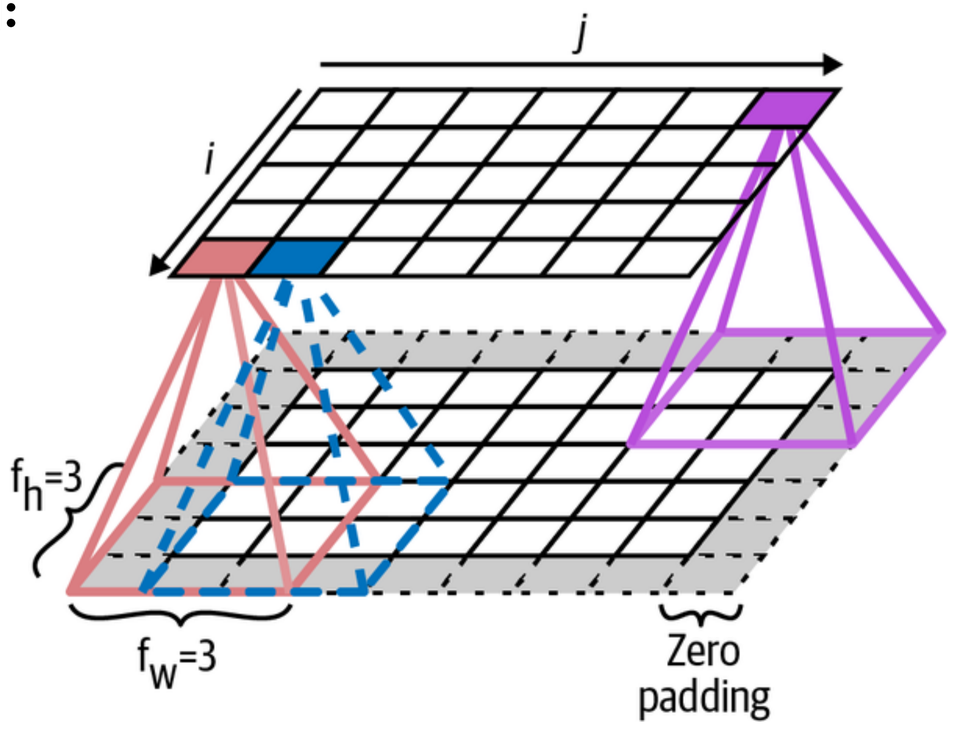
Convolutional Neural Network

Convolutional Layer

Each neuron in the second convolutional layer is connected only to neurons located within a small rectangle in the first layer.

Then assemble them into larger, higher-level features in the next hidden layer,

Padding:



Deep Computer Vision

Convolutional Neural Network

Convolutional Layer

Each neuron in the second convolutional layer is connected only to neurons located within a small rectangle in the first layer.

Then assemble them into larger, higher-level features in the next hidden layer,

Stride:

$f_h : height$

$s_h : vertical$

$f_w : width$

$s_w : horizontal$

$i \cdot s_h \quad to \quad i \cdot s_h + f_h - 1$

Rows

$j \cdot s_w \quad to \quad j \cdot s_w + f_w - 1$

Columns

Deep Computer Vision

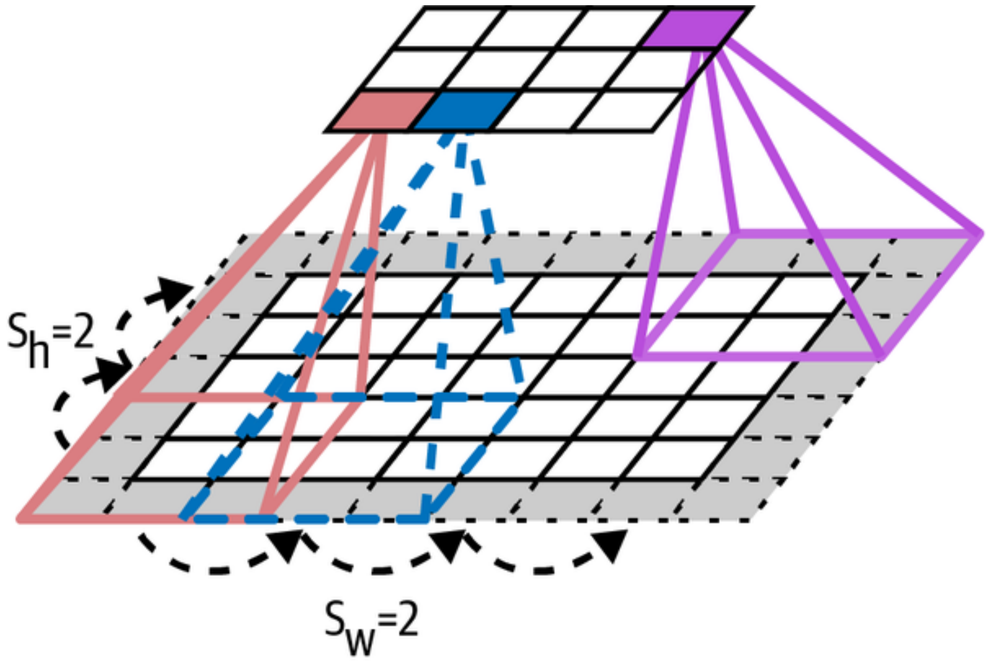
Convolutional Neural Network

Convolutional Layer

Each neuron in the second convolutional layer is connected only to neurons located within a small rectangle in the first layer.

Then assemble them into larger, higher-level features in the next hidden layer,

Stride (steps):



Deep Computer Vision

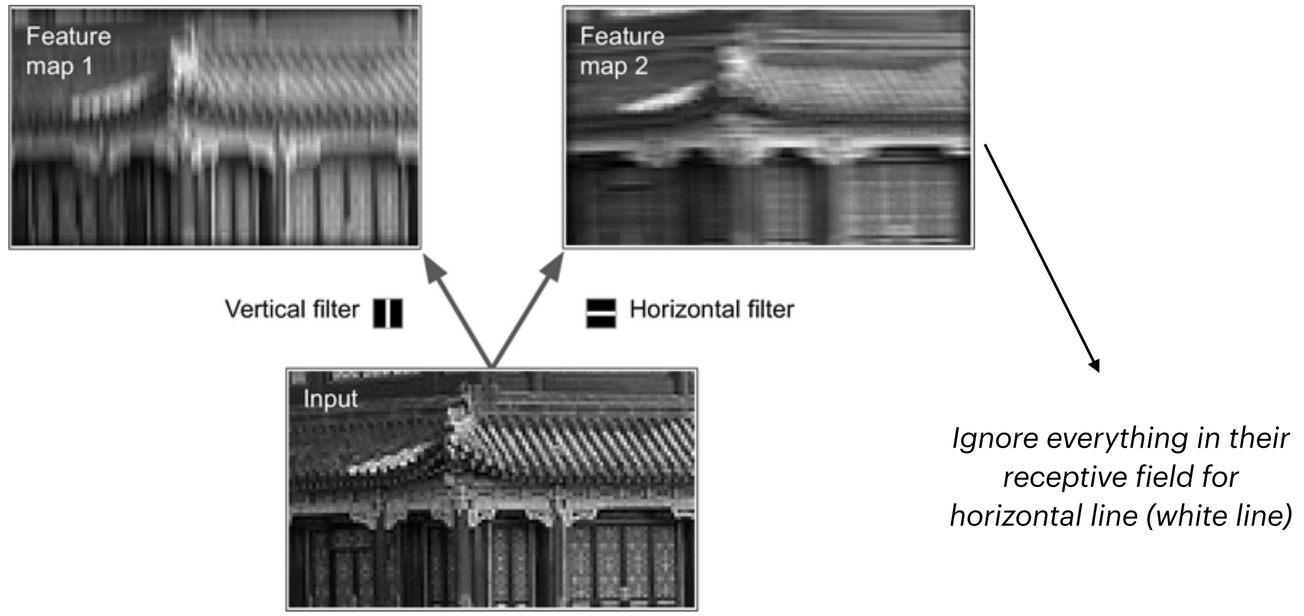
Convolutional Neural Network

Convolutional Layer

Each neuron in the second convolutional layer is connected only to neurons located within a small rectangle in the first layer.

Then assemble them into larger, higher-level features in the next hidden layer,

Filters: *"Convolution kernels" or "kernels"*



Ignore everything in their receptive field for horizontal line (white line)

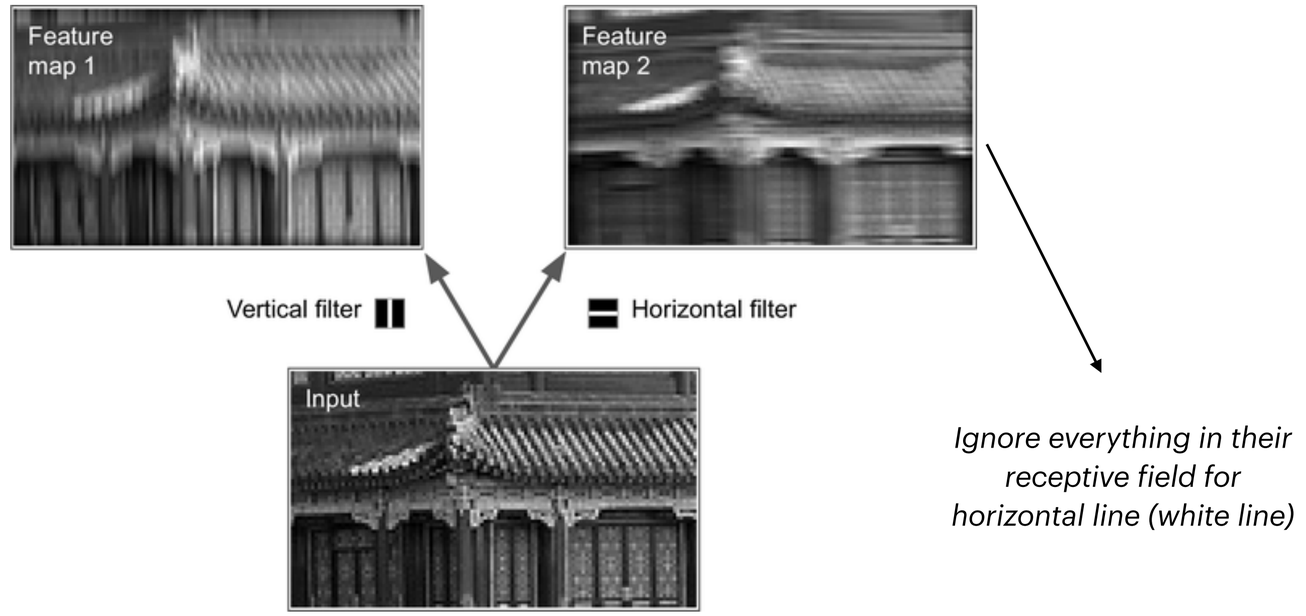
Deep Computer Vision

Convolutional Neural Network

Feature map

A layer full of neurons using the same filter outputs a "feature map"

Filters: *"Convolution kernels" or "kernels"*



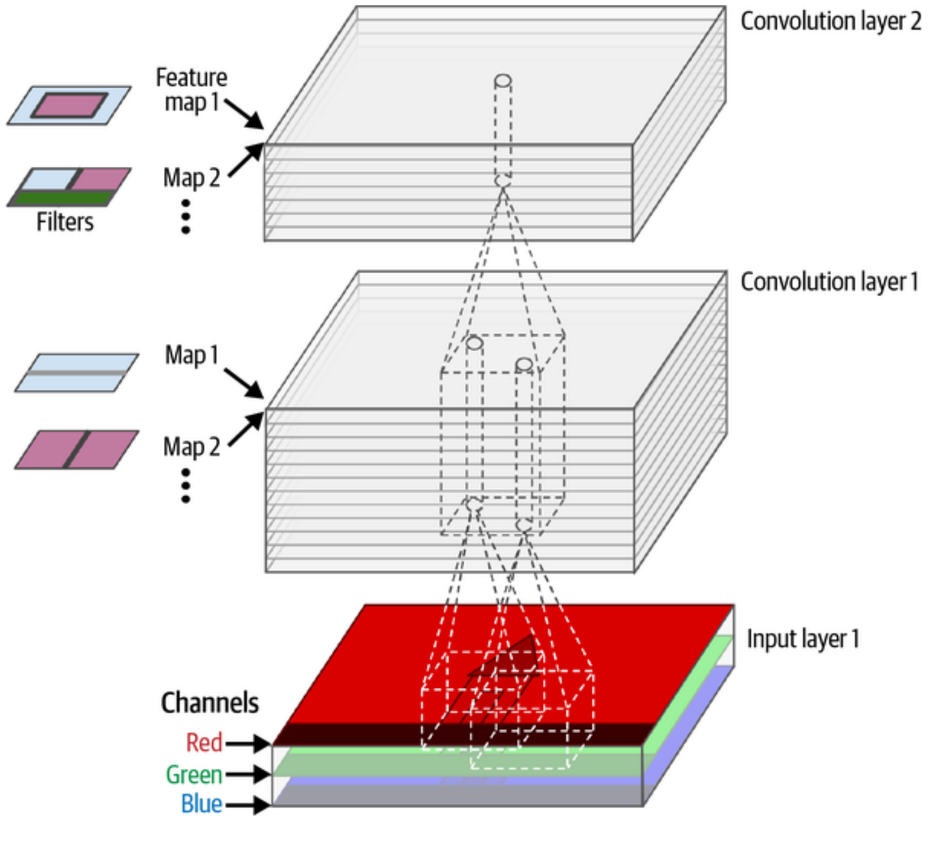
Ignore everything in their receptive field for horizontal line (white line)

Deep Computer Vision

Convolutional Neural Network

Stacking multiple Feature maps

A layer full of neurons using the same filter outputs a "feature map"

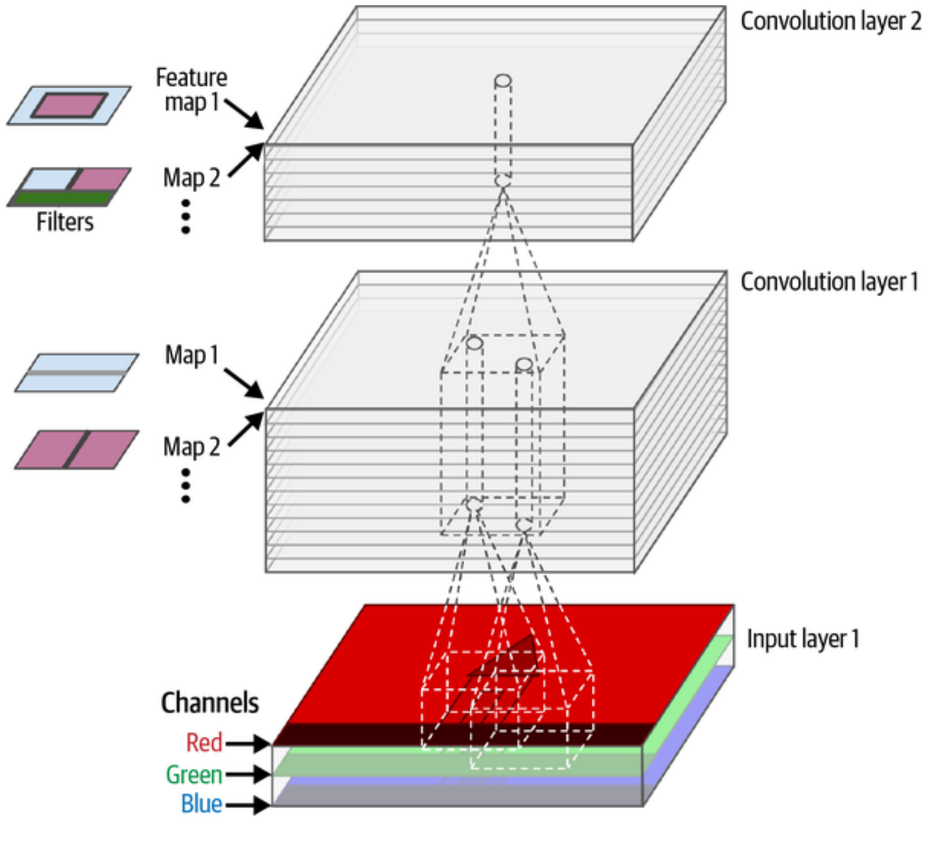


Deep Computer Vision

Convolutional Neural Network

Stacking multiple Feature maps

A layer full of neurons using the same filter outputs a "feature map"

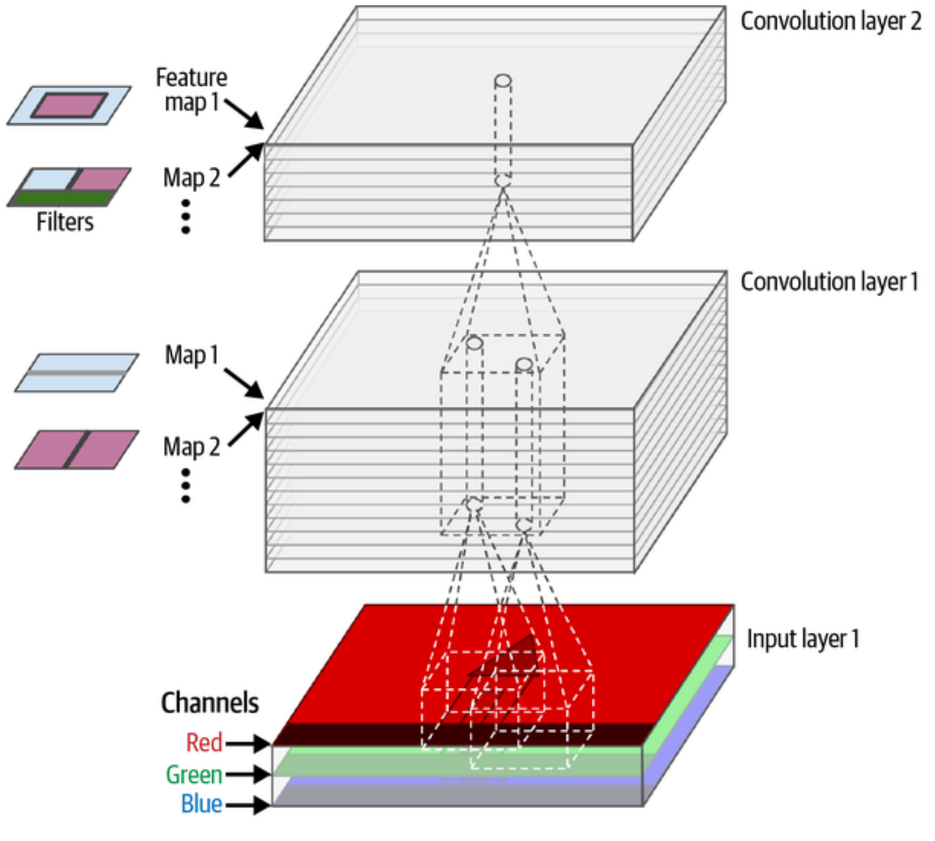


Deep Computer Vision

Convolutional Neural Network

Stacking multiple Feature maps

A layer full of neurons using the same filter outputs a "feature map"



- 1 neuron : 1 Pixel
- Between feature maps we have different params and bias

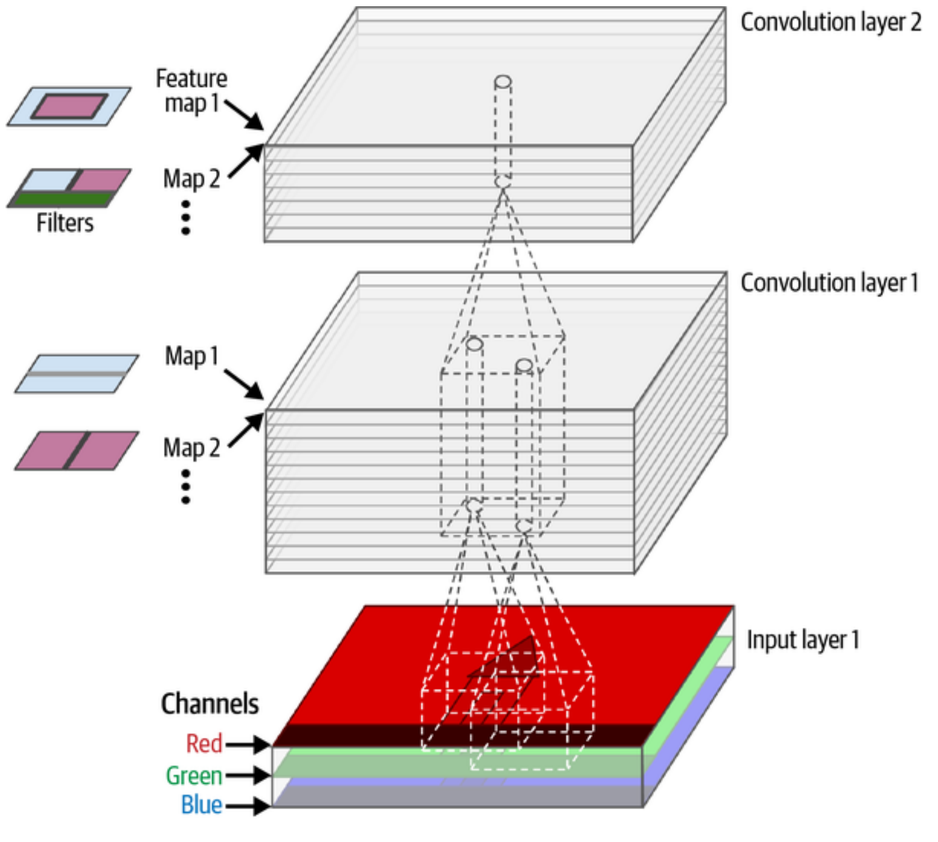
- The fact that all neurons in a feature map share the same parameters drastically reduces the number of parameters in the model

Deep Computer Vision

Convolutional Neural Network

Stacking multiple Feature maps

A layer full of neurons using the same filter outputs a "feature map"



- 1 neuron : 1 Pixel
- Between feature maps we have different params and bias

- The fact that all neurons in a feature map share the same parameters drastically reduces the number of parameters in the model

- Once CNN has learned to recognize a pattern in a location, it can recognize it in any other location.

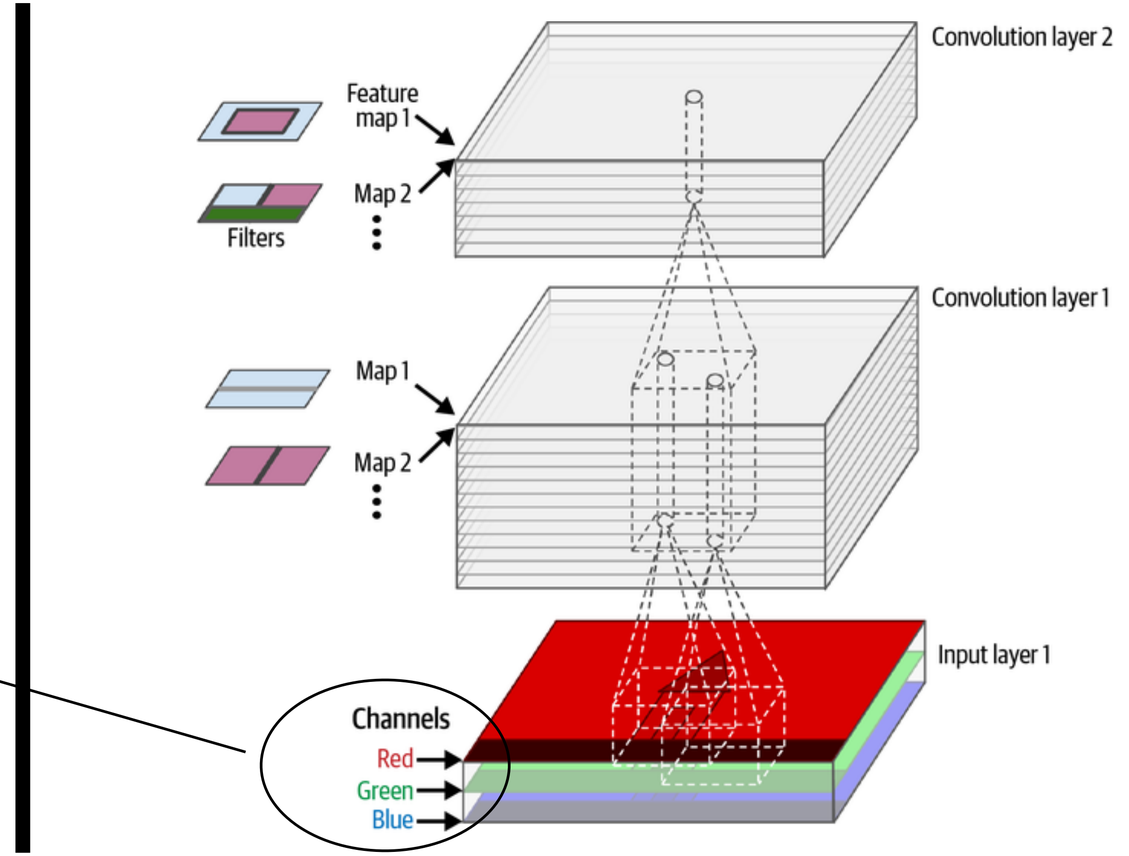
Deep Computer Vision

Convolutional Neural Network

Stacking multiple Feature maps

A layer full of neurons using the same filter outputs a "feature map"

RGB + Infrared



- 1 neuron : 1 Pixel
- Between feature maps we have different params and bias

- The fact that all neurons in a feature map share the same parameters drastically reduces the number of parameters in the model

- Once CNN has learned to recognize a pattern in a location, it can recognize it in any other location.

Deep Computer Vision

Convolutional Neural Network

Stacking multiple Feature maps

A layer full of neurons using the same filter outputs a “feature map”

Calculate the output of a neuron in a convolutional layer

$$z_{i,j,k} = b_k + \sum_{u=0}^{f_h-1} \sum_{v=0}^{f_w-1} \sum_{k'=0}^{f_{n'}-1} x_{i',j',k'} \times w_{u,v,k',k} \quad \text{with} \quad \begin{matrix} i' = i \times s_h + u \\ j' = j \times s_w + v \end{matrix}$$

$Z_{i,j,k}$ is the output of the neuron located in row i, column j in the map of characteristics k of the convolutional layer (layer l).

$X_{i',j',k'}$ Is the output of the neuron located in layer l - 1, row i', column j', characteristic map k' (or channel k' if the previous layer is the input layer).

Deep Computer Vision

Convolutional Neural Network

Stacking multiple Feature maps

A layer full of neurons using the same filter outputs a "feature map"

Calculate the output of a neuron in a convolutional layer

$$z_{i,j,k} = b_k + \sum_{u=0}^{f_h-1} \sum_{v=0}^{f_w-1} \sum_{k'=0}^{f_{n'}-1} x_{i',j',k'} \times w_{u,v,k',k} \quad \text{with} \quad \begin{matrix} i' = i \times s_h + u \\ j' = j \times s_w + v \end{matrix}$$

b_k Is the bias term for the characteristic map k (in layer l). You can think of it as a knob that adjusts the general brightness of the k characteristic map.

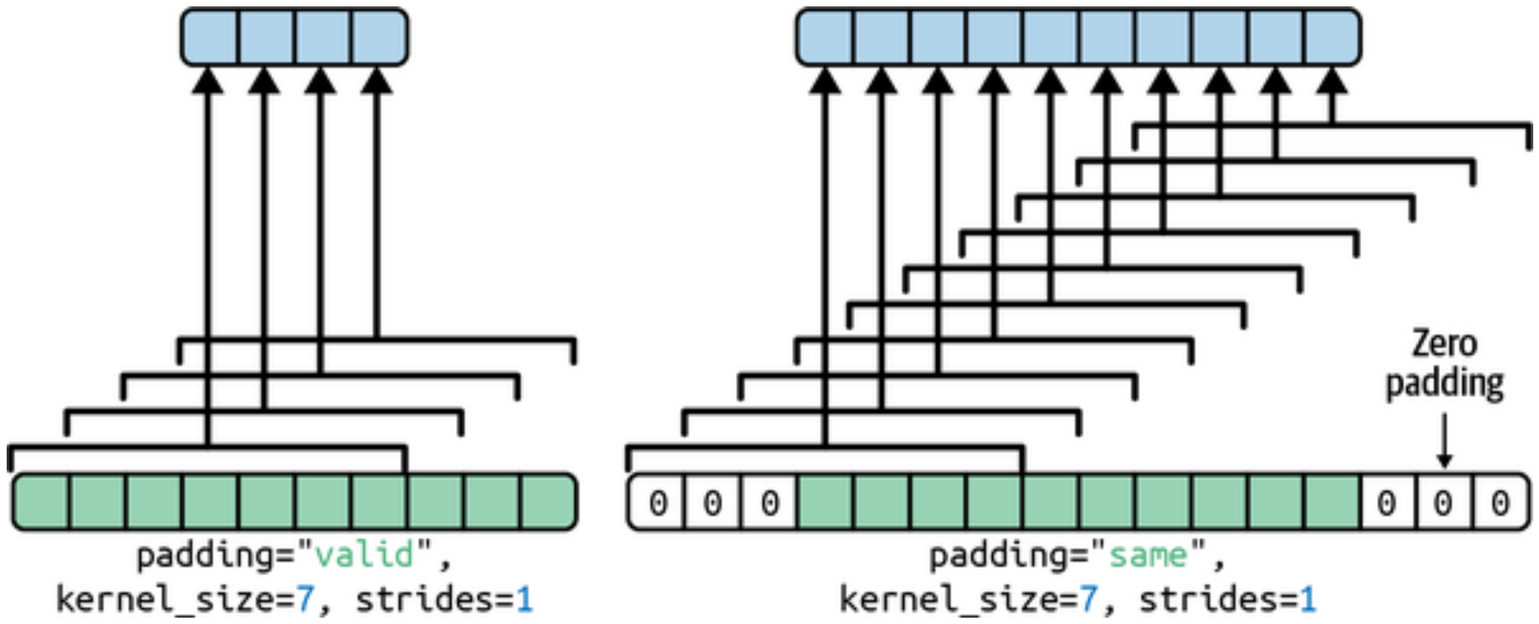
$W_{u,v,k',k}$ Is the connection weight between any neuron in the characteristic map k of the l layer and its input located in row u, column v (in relation to the receptive field of the neuron) and the map of characteristics k'.

Deep Computer Vision

Convolutional Neural Network

Stacking multiple Feature maps

A layer full of neurons using the same filter outputs a "feature map"

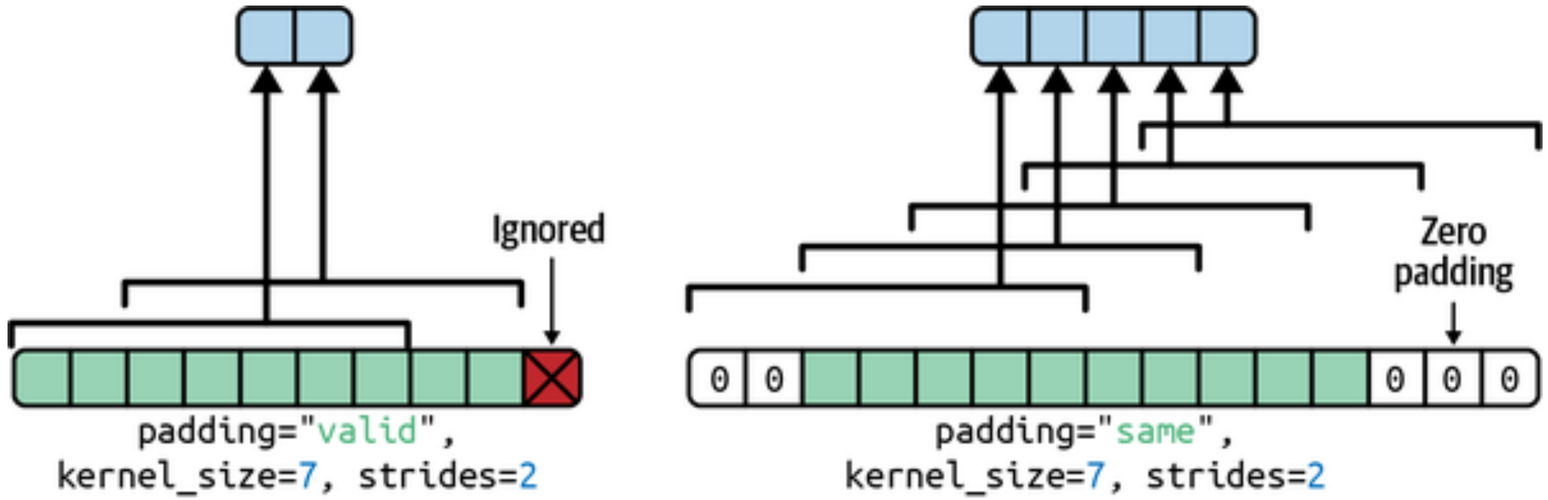


Deep Computer Vision

Convolutional Neural Network

Stacking multiple Feature maps

A layer full of neurons using the same filter outputs a "feature map"



Deep Computer Vision

Convolutional Neural Network

Stacking multiple Feature maps

A layer full of neurons using the same filter outputs a “feature map”

Hyperparams

- *Filters*
- *Strides*
- *Kernel Size*
- *Activation*
- *Padding*
- *Kernel Initializer*

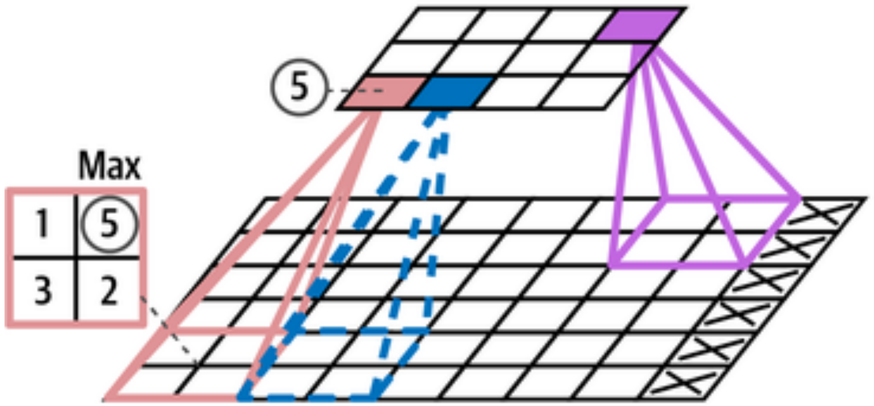
Deep Computer Vision

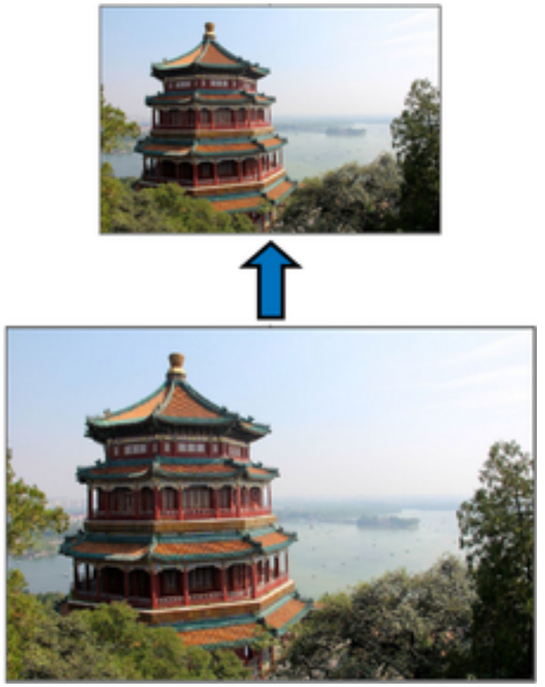
Convolutional Neural Network

Pooling or Grouping layers

Its objective is to subsampling (i.e., reduce) the input image to reduce the computational load, memory usage and the number of parameters (thus limiting the risk of overfitting).

Max pooling





Kernel size: 2 x 2

Stride: 2

Padding: “valid”

Deep Computer Vision

Convolutional Neural Network

Pooling or Grouping layers

Its objective is to subsampling (i.e., reduce) the input image to reduce the computational load, memory usage and the number of parameters (thus limiting the risk of overfitting).

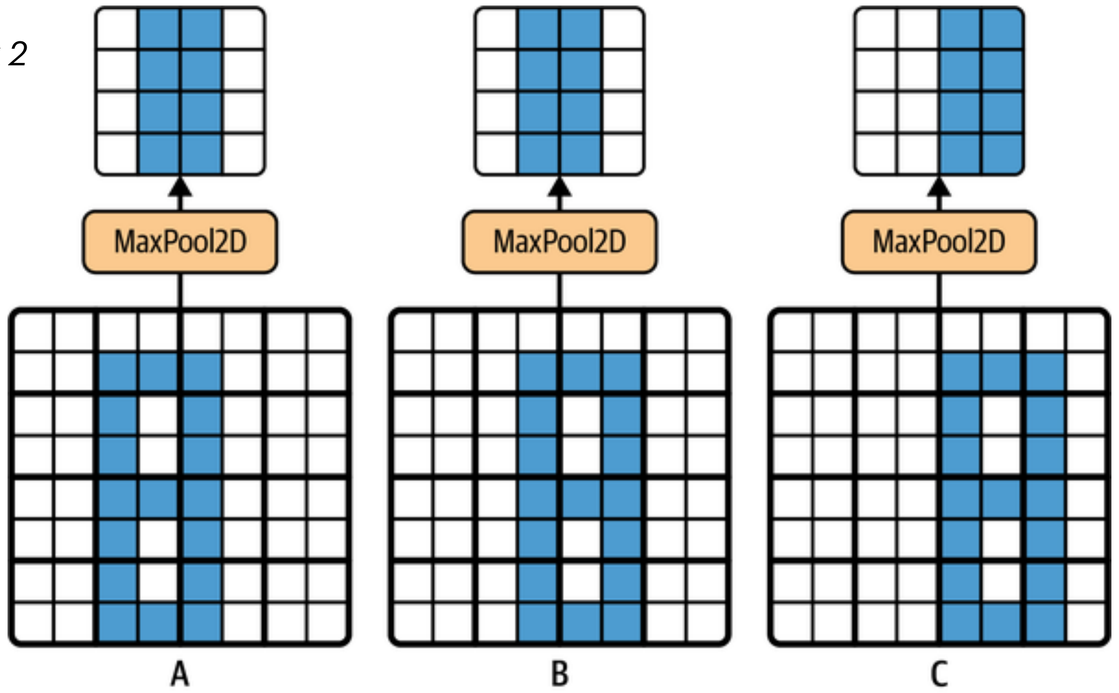
Max pooling

Kernel size: 2 x 2

Stride: 2

Invariancia a traducciones pequeñas

Output: 75% de los valores de entrada.



Deep Computer Vision

Convolutional Neural Network

Pooling or Grouping layers

Its objective is to subsampling (i.e., reduce) the input image to reduce the computational load, memory usage and the number of parameters (thus limiting the risk of overfitting).

Max pooling

- *Preserves only the strongest characteristics*
- *The following layers get a cleaner signal to work with*
- *A stronger translation invariance*
- *Requires a little less computing*

Avg. pooling

- *Loses less information*

Deep Computer Vision

Convolutional Neural Network

Pooling or Grouping layers

Its objective is to subsampling (i.e., reduce) the input image to reduce the computational load, memory usage and the number of parameters (thus limiting the risk of overfitting).

Max pooling

- *Preserves only the strongest characteristics*
- *The following layers get a cleaner signal to work with*
- *A stronger translation invariance*
- *Requires a little less computing*

Avg. pooling

- *Loses less information*

Global Avg. pooling

Only generates a single number per feature map.

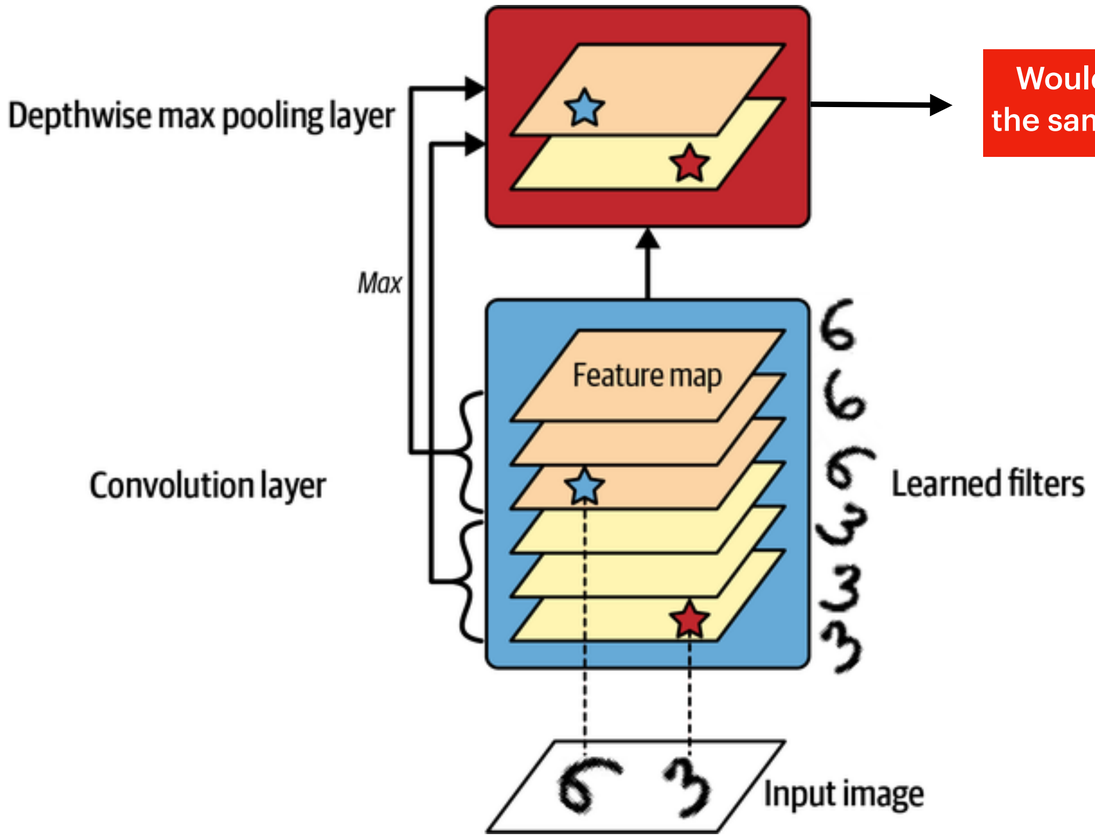
- ***Extremely destructive*** *(most of the information in the feature map is lost)*
- *It can be useful just before the output layer*

Deep Computer Vision

Convolutional Neural Network

Pooling or Grouping layers

Its objective is to subsampling (i.e., reduce) the input image to reduce the computational load, memory usage and the number of parameters (thus limiting the risk of overfitting).



Would guarantee that the exit is the same regardless of the rotation

Deep Co

Convolutio

Pooling or Grouping layers

Its objective is to subsampling (i.e., reduce) the input image to reduce the computational load, memory usage and the number of parameters (thus limiting the risk of overfitting).

Depthwise max poo

Convolution

uld guarantee that the exit is
ame regardless of the rotation

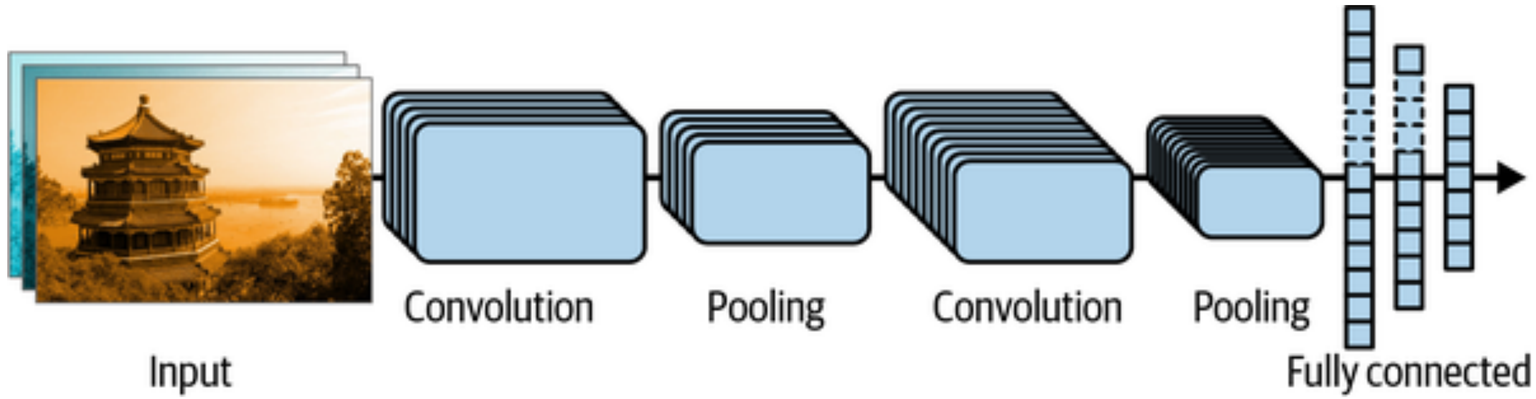
Deep Computer Vision

CNN Architectures

Typical CNN architecture

The typical architectures of CNN stack some convolutional layers.

The image becomes smaller and smaller as it advances through the network.



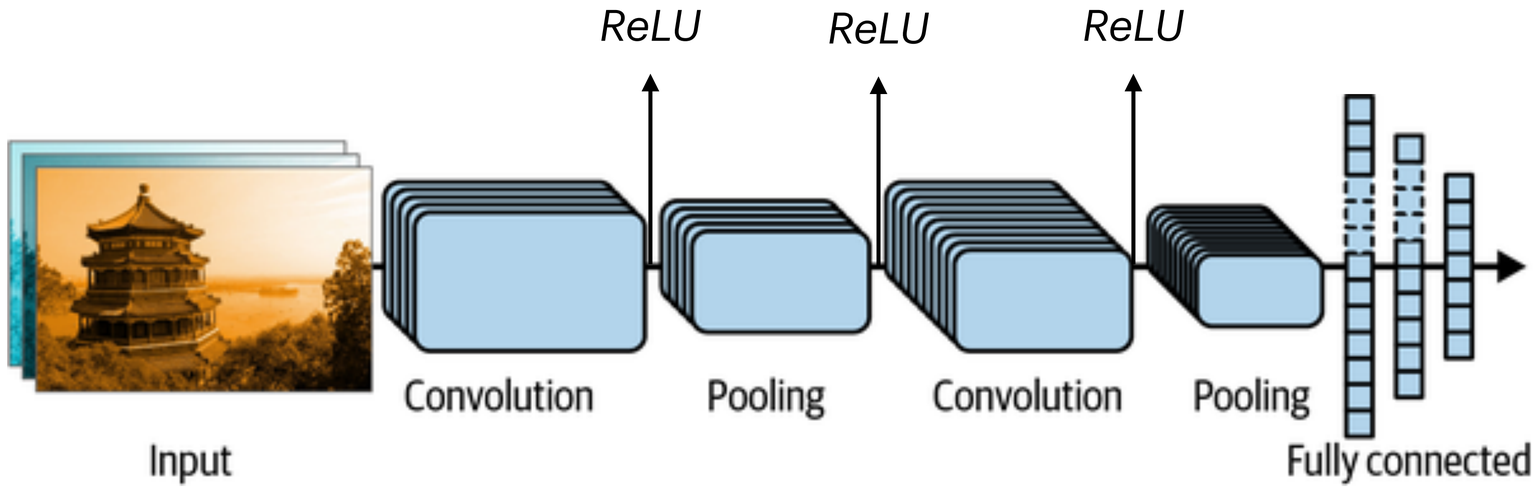
Deep Computer Vision

CNN Architectures

Typical CNN architecture

The typical architectures of CNN stack some convolutional layers.

The image becomes smaller and smaller as it advances through the network.



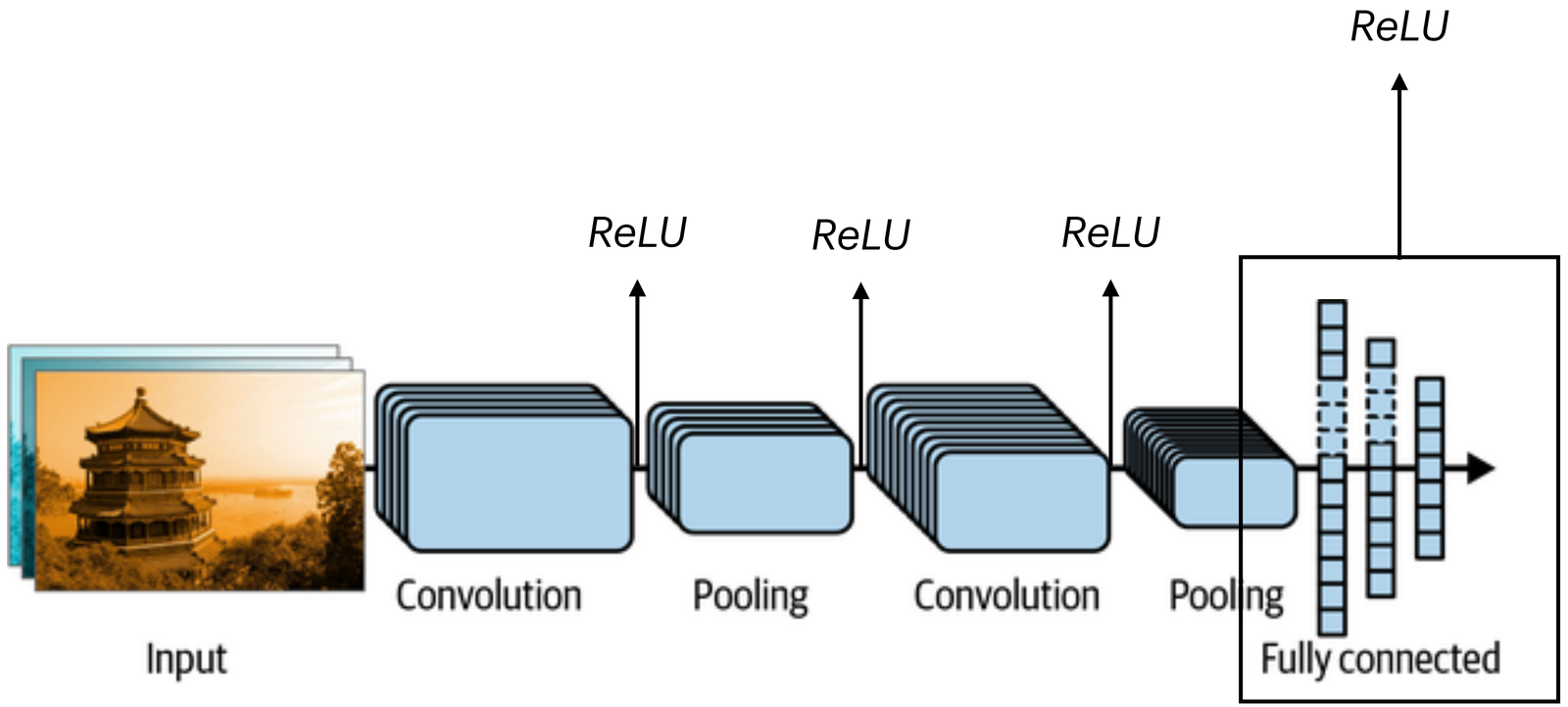
Deep Computer Vision

CNN Architectures

Typical CNN architecture

The typical architectures of CNN stack some convolutional layers.

The image becomes smaller and smaller as it advances through the network.



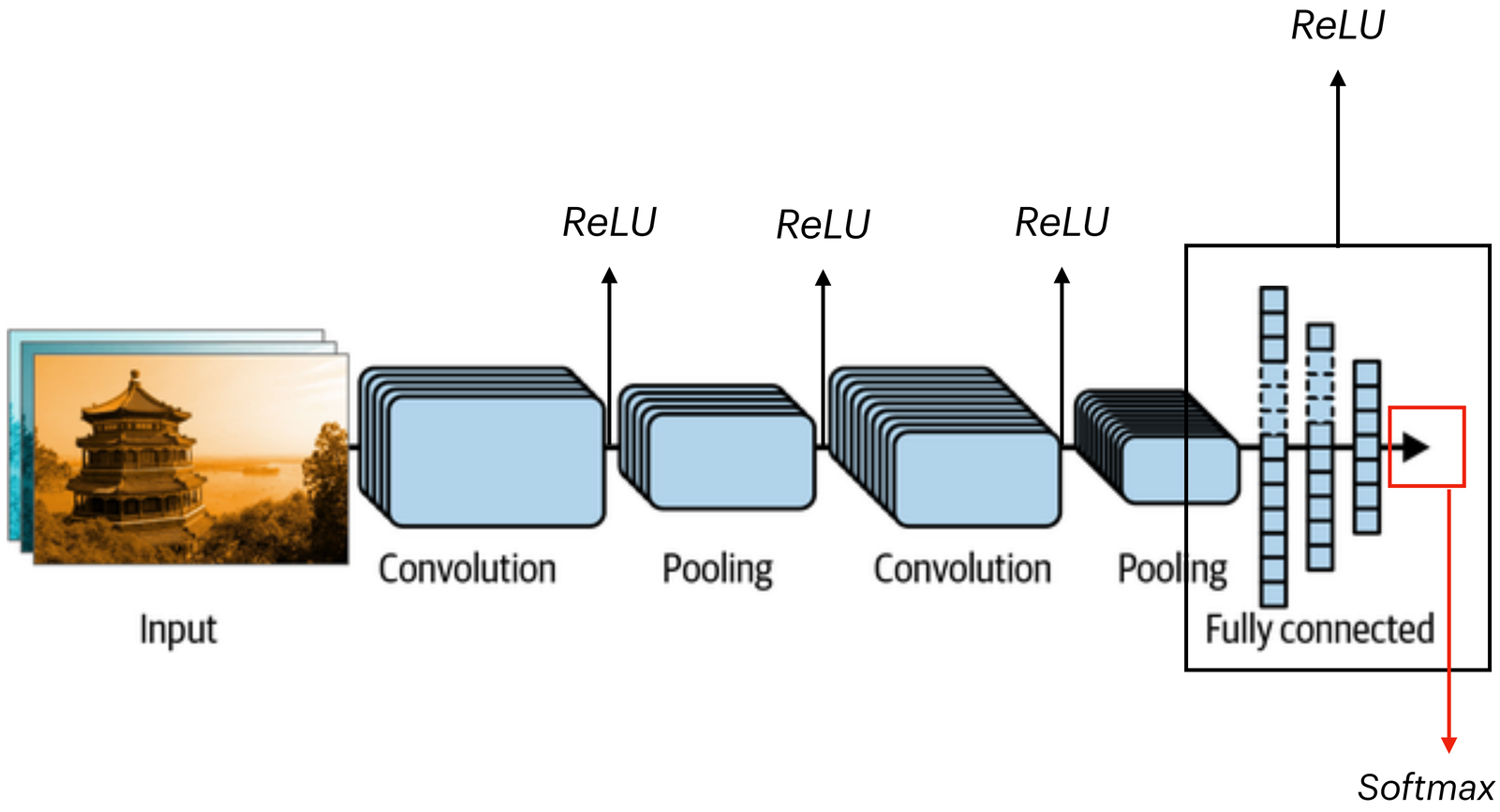
Deep Computer Vision

CNN Architectures

Typical CNN architecture

The typical architectures of CNN stack some convolutional layers.

The image becomes smaller and smaller as it advances through the network.



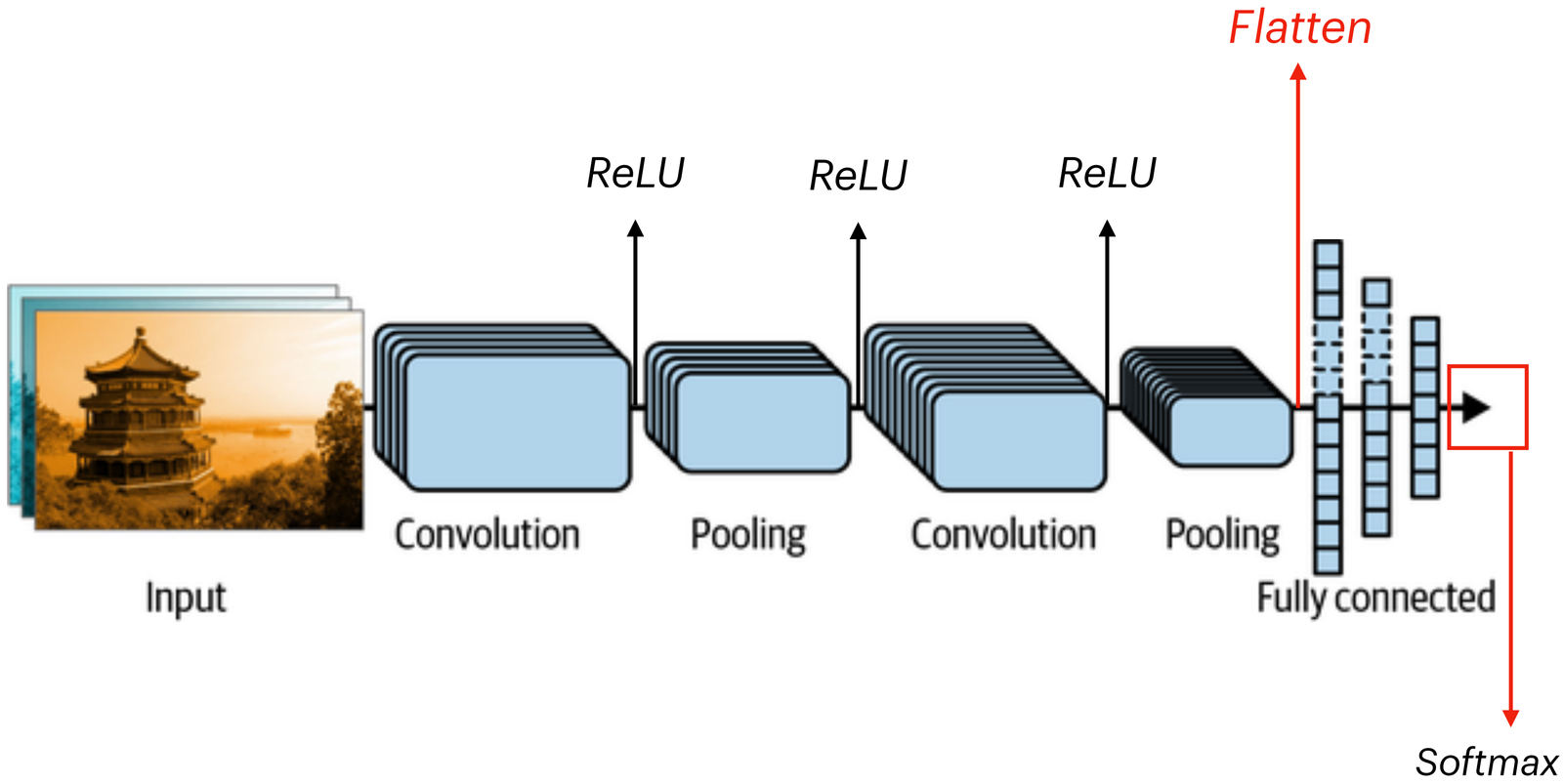
Deep Computer Vision

CNN Architectures

Typical CNN architecture

The typical architectures of CNN stack some convolutional layers.

The image becomes smaller and smaller as it advances through the network.



Deep Computer Vision

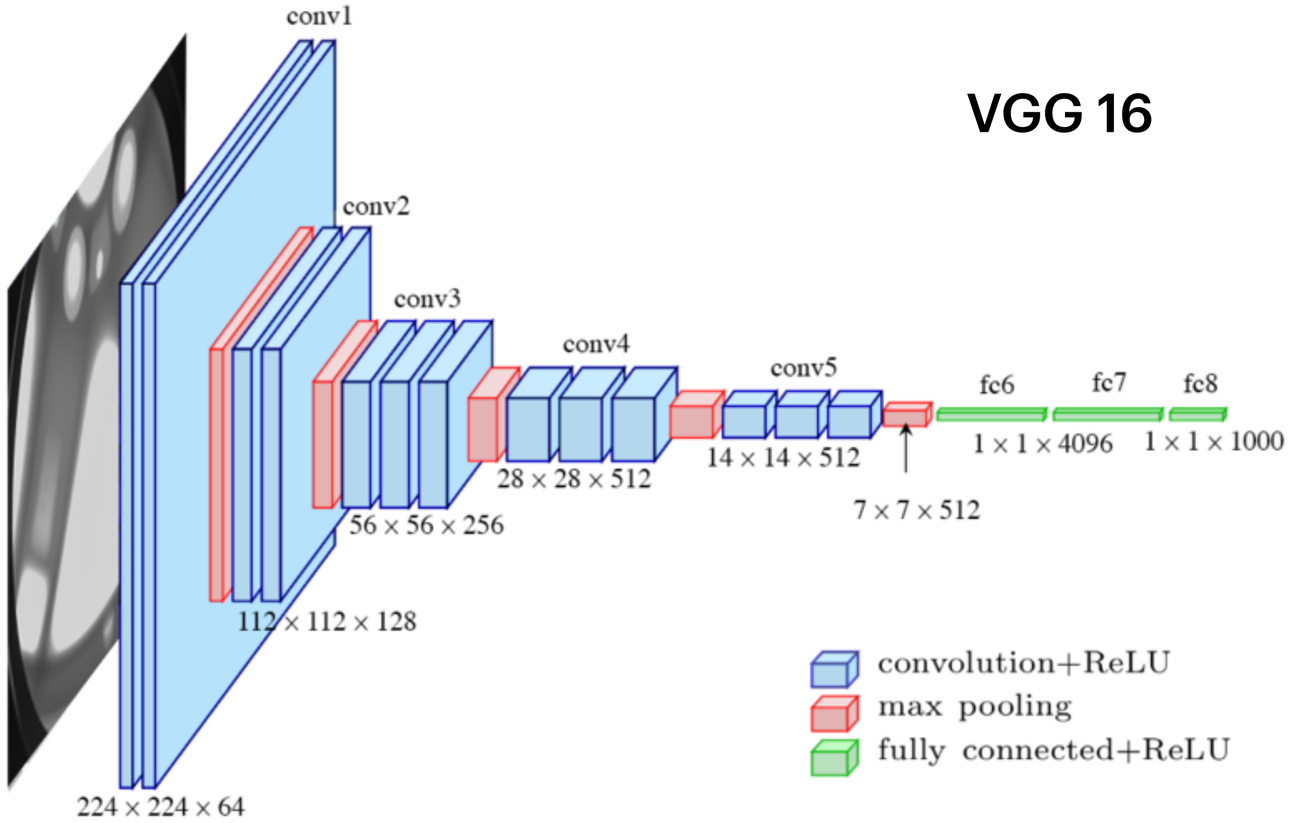
CNN Architectures

Typical CNN architecture

The typical architectures of CNN stack some convolutional layers.

The image becomes smaller and smaller as it advances through the network.

Destilation Law



VGG 16

Pontificia Universidad JAVERIANA Colombia

Deep Computer Vision

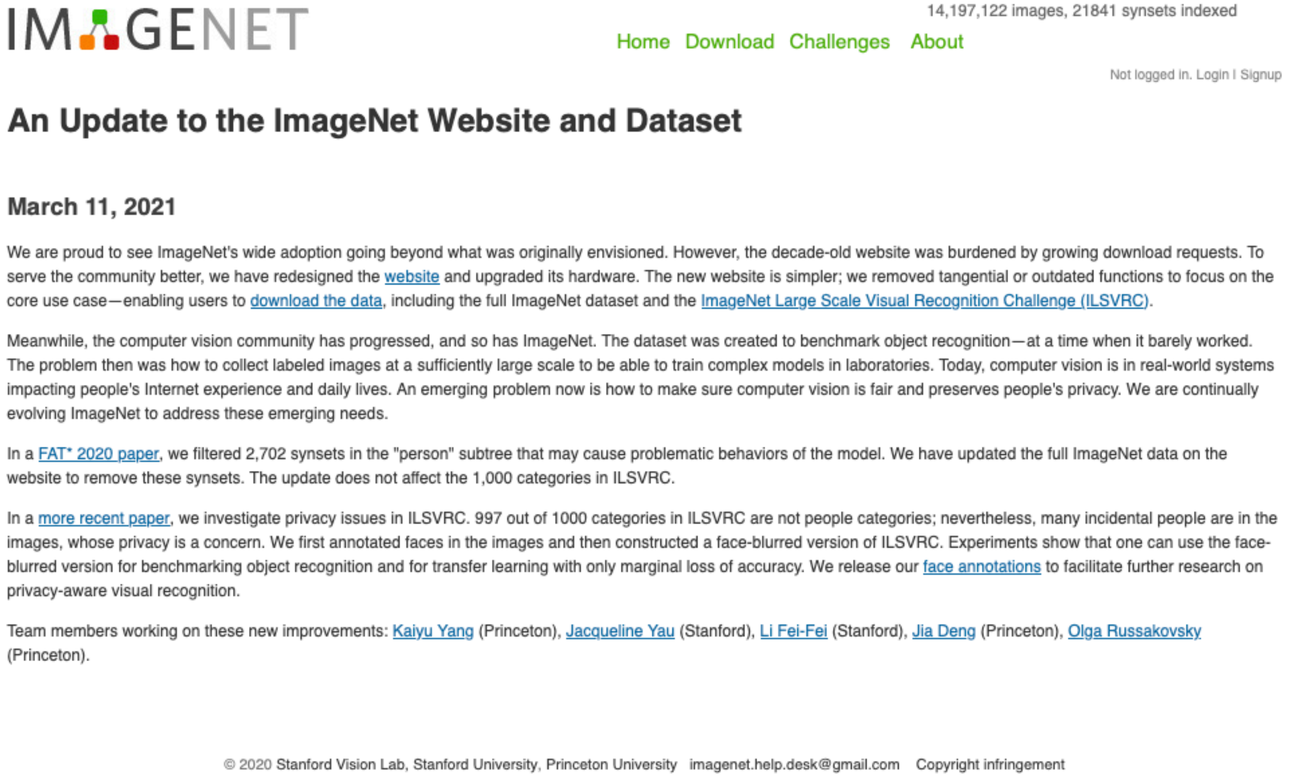
CNN Architectures

Typical CNN architecture

The typical architectures of CNN stack some convolutional layers.

The image becomes smaller and smaller as it advances through the network.

Destilation Law



IMAGENET

14,197,122 images, 21841 synsets indexed

Home Download Challenges About

Not logged in. Login | Signup

An Update to the ImageNet Website and Dataset

March 11, 2021

We are proud to see ImageNet's wide adoption going beyond what was originally envisioned. However, the decade-old website was burdened by growing download requests. To serve the community better, we have redesigned the website and upgraded its hardware. The new website is simpler; we removed tangential or outdated functions to focus on the core use case—enabling users to download the data, including the full ImageNet dataset and the ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

Meanwhile, the computer vision community has progressed, and so has ImageNet. The dataset was created to benchmark object recognition—at a time when it barely worked. The problem then was how to collect labeled images at a sufficiently large scale to be able to train complex models in laboratories. Today, computer vision is in real-world systems impacting people's Internet experience and daily lives. An emerging problem now is how to make sure computer vision is fair and preserves people's privacy. We are continually evolving ImageNet to address these emerging needs.

In a FAT* 2020 paper, we filtered 2,702 synsets in the "person" subtree that may cause problematic behaviors of the model. We have updated the full ImageNet data on the website to remove these synsets. The update does not affect the 1,000 categories in ILSVRC.

In a more recent paper, we investigate privacy issues in ILSVRC. 997 out of 1000 categories in ILSVRC are not people categories; nevertheless, many incidental people are in the images, whose privacy is a concern. We first annotated faces in the images and then constructed a face-blurred version of ILSVRC. Experiments show that one can use the face-blurred version for benchmarking object recognition and for transfer learning with only marginal loss of accuracy. We release our face annotations to facilitate further research on privacy-aware visual recognition.

Team members working on these new improvements: Kaiyu Yang (Princeton), Jacqueline Yau (Stanford), Li Fei-Fei (Stanford), Jia Deng (Princeton), Olga Russakovsky (Princeton).

Deep Computer Vision

CNN Architectures

Typical CNN architecture

The typical architectures of CNN stack some convolutional layers.

The image becomes smaller and smaller as it advances through the network.

Destilation Law

Error rate improvement: ILSVRC ImageNet

Deep Computer Vision

CNN Architectures

Typical CNN architecture

The typical architectures of CNN stack some convolutional layers.

The image becomes smaller and smaller as it advances through the network.

Destilation Law

Error rate improvement: ILSVRC ImageNet

Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

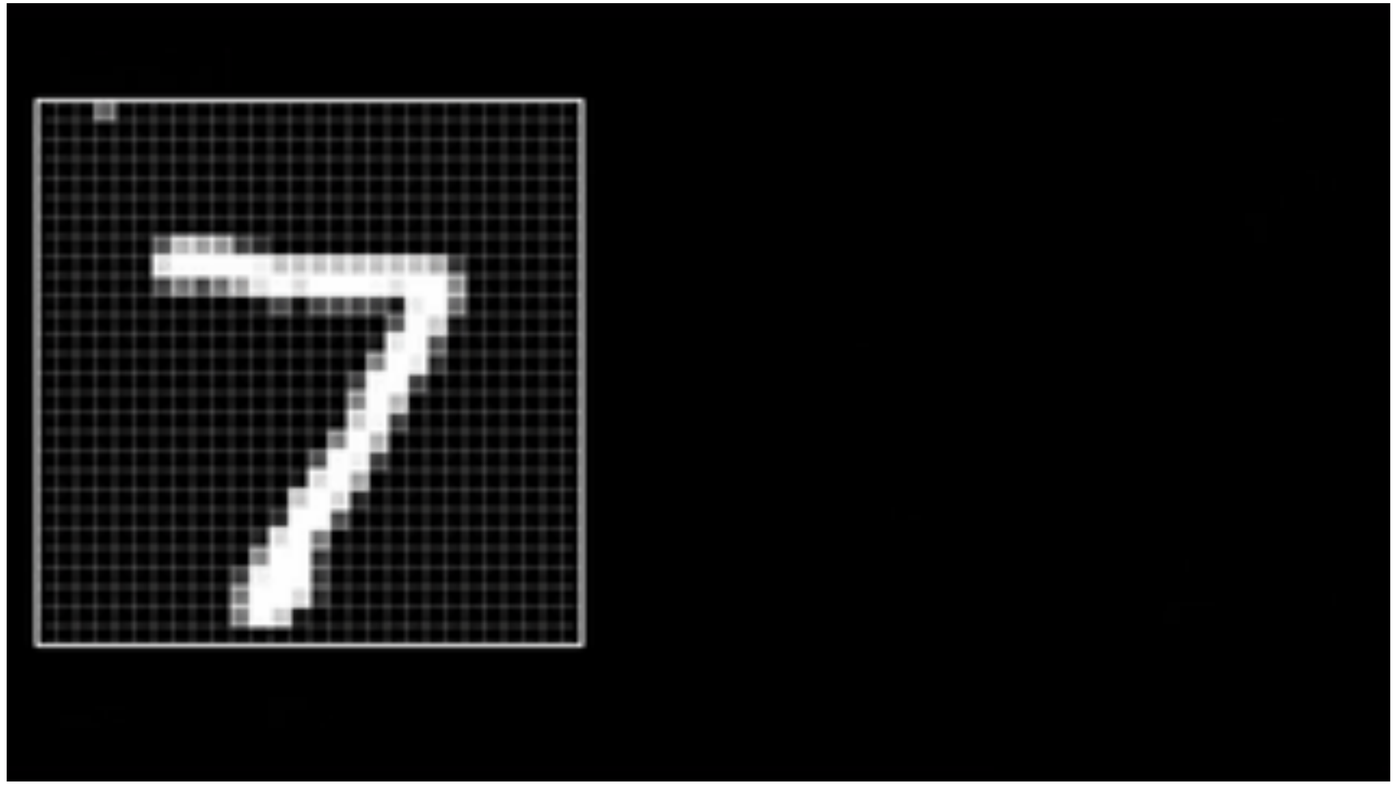
Deep Computer Vision

Convolutional Neural Network

Convolutional Layer

Each neuron in the second convolutional layer is connected only to neurons located within a small rectangle in the first layer.

Then assemble them into larger, higher-level features in the next hidden layer,



Pontificia Universidad JAVERIANA Colombia

Deep Computer Vision

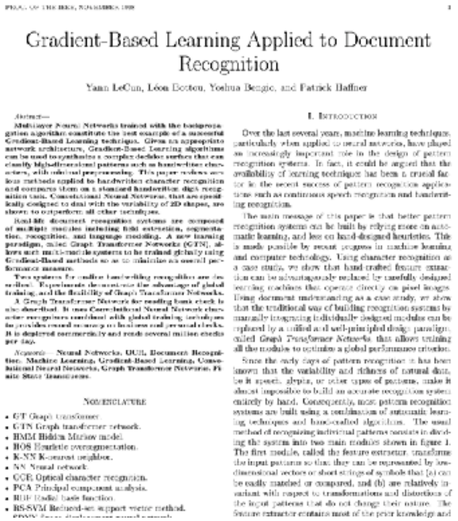
CNN Architectures

ILSVRC ImageNet

Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

LeNet-5 (1998)



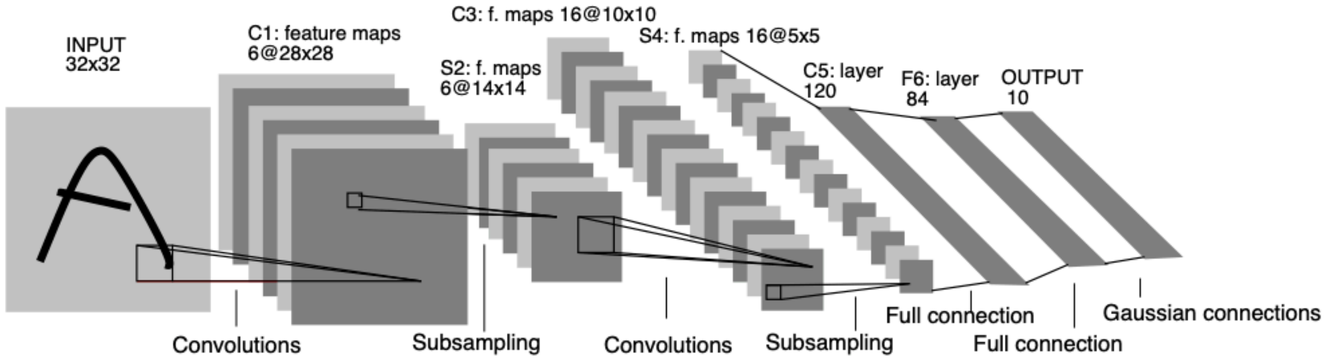
Gradient-Based Learning Applied to Document Recognition

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

I. Introduction

Over the last several years, machine learning techniques, particularly when applied to neural networks, have played an increasingly important role in the design of pattern recognition systems. In fact, it could be argued that the availability of learning techniques has been a crucial factor in the recent success of pattern recognition applications such as continuous speech recognition and handwriting recognition.

The main message of this paper is that better pattern recognition systems can be built by relying more on automatic learning, and less on hand-designed heuristics. This is made possible by recent progress in machine learning and computer technology. Using character recognition as a case study, we show that hand-crafted feature extraction can be advantageously replaced by carefully designed learning machines that operate directly on pixel images. Using document understanding as a case study, we show that the traditional way of building recognition systems by manually integrating individually designed modules can be replaced by a cleanly formalized paradigm, called Graph Transformer Networks, that allows training all the modules to optimize a global performance criterion.



Pontificia Universidad JAVERIANA Colombia

Deep Computer Vision

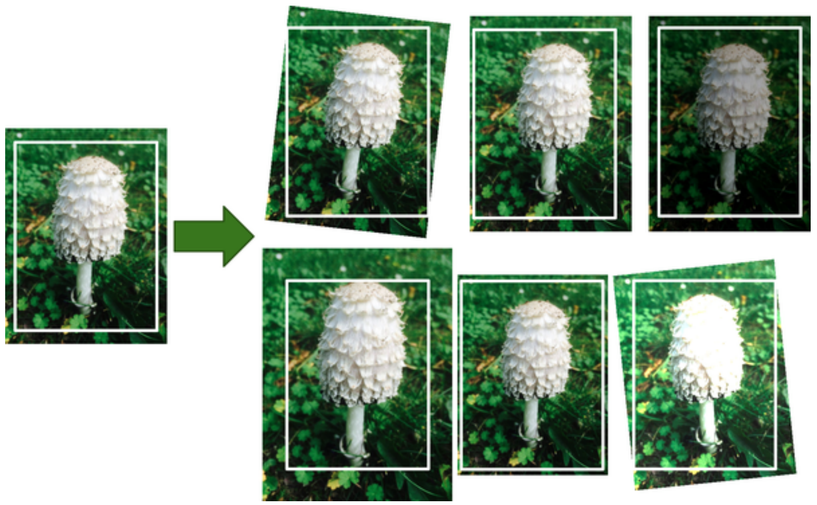
CNN Architectures

ILSVRC ImageNet

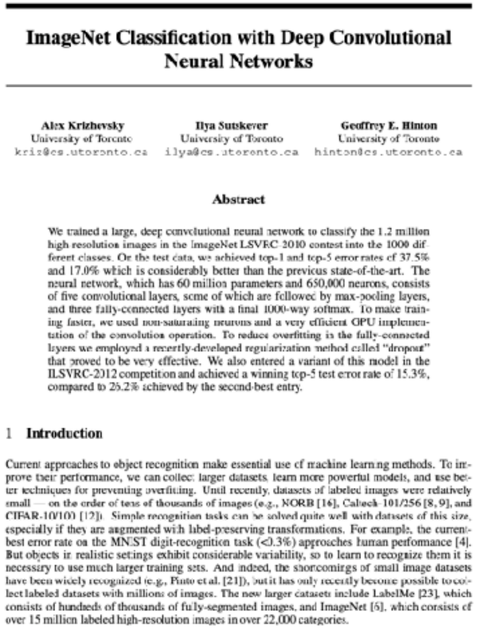
Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

AlexNet CNN (ILSVRC winner 2012)



INCREASE IN DATA



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky — University of Toronto — kriz@cs.utoronto.ca
Ilya Sutskever — University of Toronto — ilya@cs.utoronto.ca
Geoffrey E. Hinton — University of Toronto — hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high resolution images in the ImageNet LSVRC 2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

1 Introduction

Current approaches to object recognition make essential use of machine learning methods. To improve their performance, we can collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting. Until recently, datasets of labeled images were relatively small — on the order of tens of thousands of images (e.g., NORB [16], Caltech-101/256 [8, 9], and CIFAR-10/100 [12]). Simple recognition tasks can be solved quite well with datasets of this size, especially if they are augmented with label-preserving transformations. For example, the current-best error rate on the MNIST digit-recognition task (<0.3%) approaches human performance [4]. But objects in realistic settings exhibit considerable variability, so to learn to recognize them it is necessary to use much larger training sets. And indeed, the shortcomings of small image datasets have been widely recognized (e.g., Pinto et al. [21]), but it has only recently become possible to collect labeled datasets with millions of images. The new larger datasets include LabelMe [23], which consists of hundreds of thousands of fully-segmented images, and ImageNet [6], which consists of over 15 million labeled high-resolution images in over 22,000 categories.

Local Response Normalization (LRN)

$$b_i = a_i k + \alpha \left(\sum_{j=j_{\text{low}}}^{j_{\text{high}}} a_j^2 \right)^{-\beta}$$

with

$$j_{\text{high}} = \min\left(i + \frac{r}{2}, f_n - 1\right)$$

$$j_{\text{low}} = \max\left(0, i - \frac{r}{2}\right)$$

Deep Computer Vision

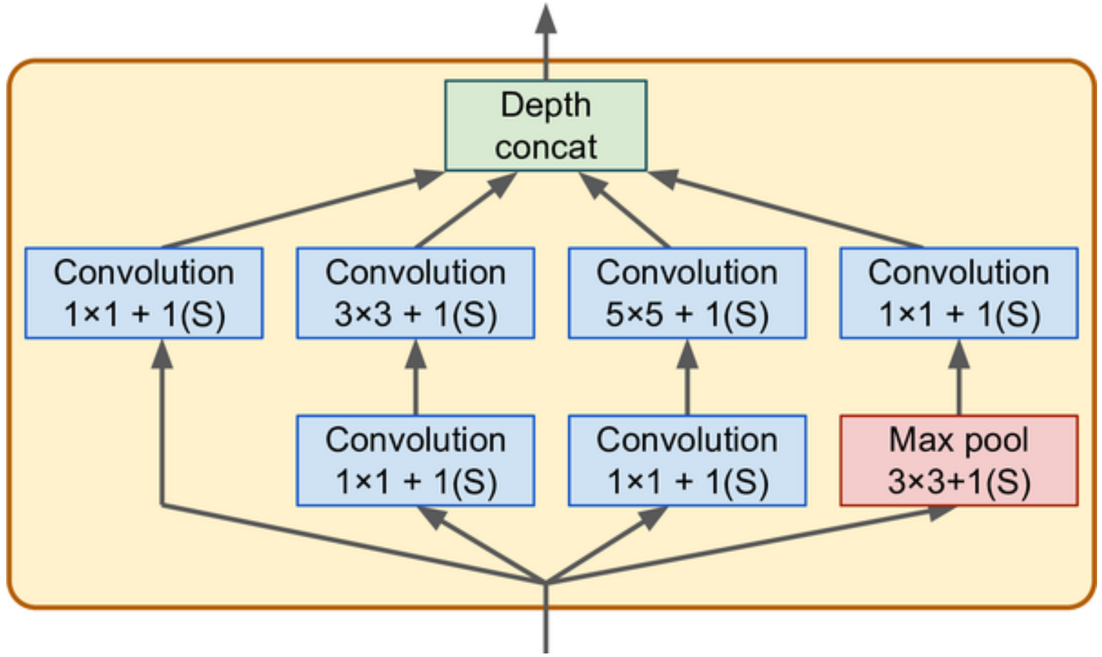
CNN Architectures

ILSVRC ImageNet

Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

GoogleNet (ILSVRC winner 2014)



Inception module

Going Deeper with Convolutions

Christian Szegedy[1], Wei Liu[2], Yangqing Jia[1], Pierre Sermanet[1], Scott Reed[3], Dragomir Anguelov[1], Dumitru Erhan[1], Vincent Vanhoucke[1], Andrew Rabinovich[4]

[1]Google Inc. [2]University of North Carolina, Chapel Hill

[3]University of Michigan, Ann Arbor [4]Magic Leap Inc.

Abstract

We propose a deep convolutional neural network architecture codenamed Inception that achieves the new state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14). The main hallmark of this architecture is the improved utilization of the computing resources inside the network. By a carefully crafted design, we increased the depth and width of the network while keeping the computational budget constant. To optimize quality, the architectural decisions were based on the Hebbian principle and the intuition of multi-scale processing. One particular incarnation used in our submission for ILSVRC14 is called GoogLeNet, a 22 layers deep network, the quality of which is assessed in the context of classification and detection.

1. Introduction

- *Has 10 times fewer parameters than AlexNet*
- *Approximately 6 million instead of 60 million*

Deep Computer Vision

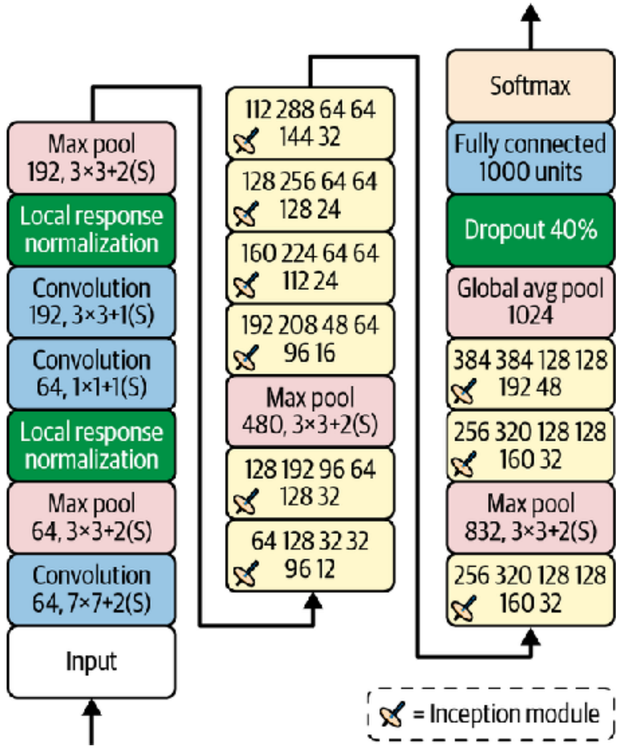
CNN Architectures

ILSVRC ImageNet

Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

GoogleNet (ILSVRC winner 2014)



Going Deeper with Convolutions

Christian Szegedy[1], Wei Liu[2], Yangqing Jia[1], Pierre Sermanet[1], Scott Reed[3], Dragomir Anguelov[1], Dumitru Erhan[1], Vincent Vanhoucke[1], Andrew Rabinovich[4]

[1]Google Inc. [2]University of North Carolina, Chapel Hill

[3]University of Michigan, Ann Arbor [4]Magic Leap Inc.

Abstract

- *Has 10 times fewer parameters than AlexNet*
- *Approximately 6 million instead of 60 million*

Pontificia Universidad JAVERIANA Colombia

Deep Computer Vision

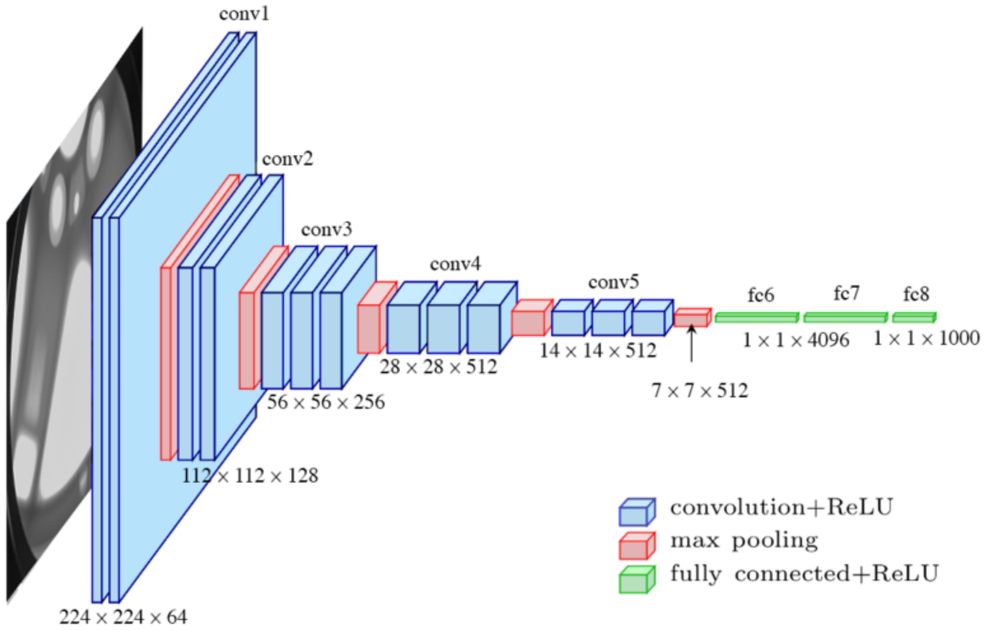
CNN Architectures

ILSVRC ImageNet

Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

VGGNet - Visual Geometry Group - Oxford University (ILSVRC sub-champion 2014)



Published as a conference paper at ICLR 2015

VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

Karen Simonyan* & Andrew Zisserman*
Visual Geometry Group, Department of Engineering Science, University of Oxford
{karen,az}@robots.ox.ac.uk

ABSTRACT

In this work we investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Our main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small (3 × 3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers. These findings were the basis of our ImageNet Challenge 2014 submission, where our team secured the first and the second places in the localisation and classification tracks respectively. We also show that our representations generalise well to other datasets, where they achieve state-of-the-art results. We have made our two best-performing ConvNet models publicly available to facilitate further research on the use of deep visual representations in computer vision.

1 INTRODUCTION

Convolutional networks (ConvNets) have recently enjoyed a great success in large-scale image and video recognition (Krizhevsky et al., 2012; Zeiler & Fergus, 2013; Sermanet et al., 2014; Simonyan & Zisserman, 2014) which has become possible due to the large public image repositories, such as ImageNet (Deng et al., 2009), and high-performance computing systems, such as GPUs or large-scale distributed clusters (Dean et al., 2012). In particular, an important role in the advance of deep visual recognition architectures has been played by the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2014), which has served as a testbed for a few generations of large-scale image classification systems, from high-dimensional shallow feature encodings (Perronnin et al., 2010) (the winner of ILSVRC-2011) to deep ConvNets (Krizhevsky et al., 2012) (the winner of ILSVRC-2012).

With ConvNets becoming more of a commodity in the computer vision field, a number of attempts have been made to improve the original architecture of Krizhevsky et al. (2012) in a bid to achieve better accuracy. For instance, the best-performing submissions to the ILSVRC-2013 (Zeiler & Fergus, 2013; Sermanet et al., 2014) utilised smaller receptive window size and smaller stride of the first convolutional layer. Another line of improvements dealt with training and testing the networks densely over the whole image and over multiple scales (Sermanet et al., 2014; Howard, 2014). In this paper, we address another important aspect of ConvNet architecture design – its depth. To this end, we fix other parameters of the architecture, and steadily increase the depth of the network by adding more convolutional layers, which is feasible due to the use of very small (3 × 3) convolution filters in all layers.

Deep Computer Vision

CNN Architectures

ILSVRC ImageNet

Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

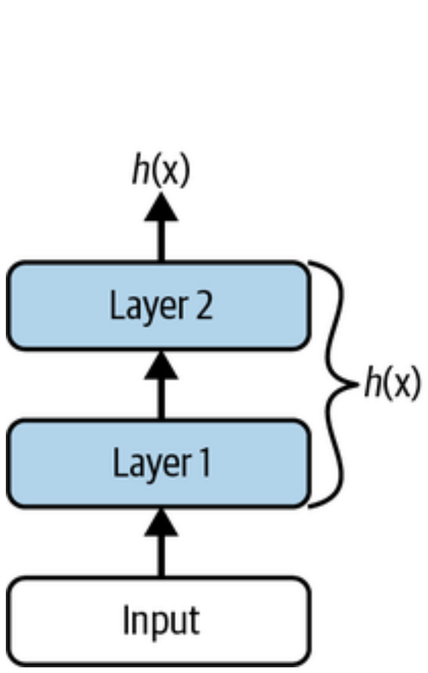
ResNet [Residual Network - Microsoft]
ILSVRC winner 2015

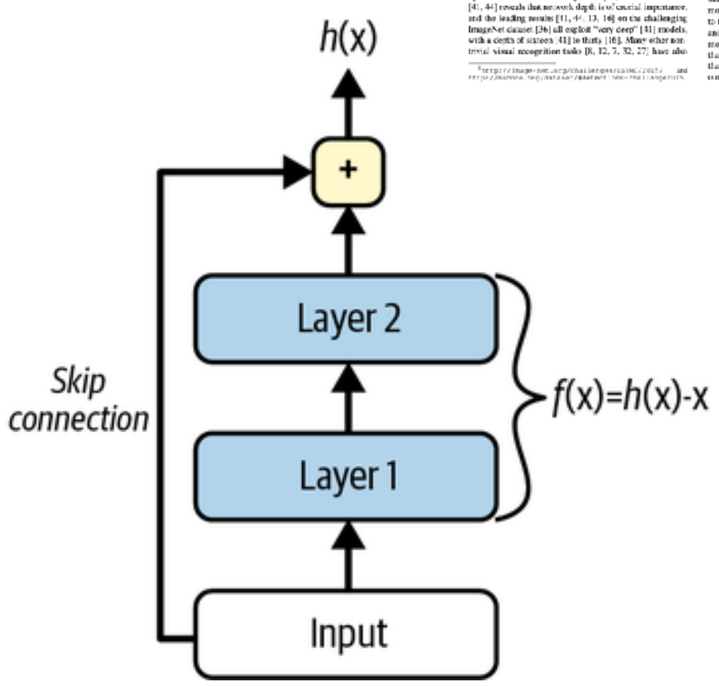

Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research

Abstract

1. Introduction





Deep Computer Vision

CNN Architectures

ILSVRC ImageNet

Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

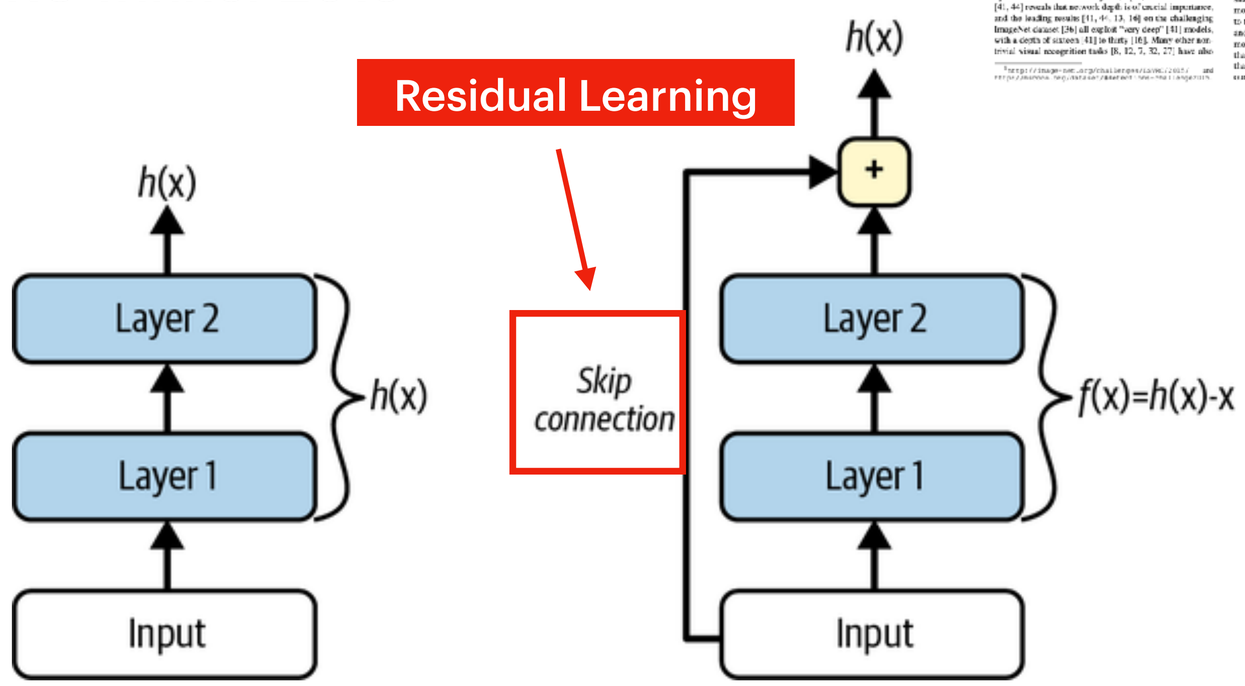
ResNet [Residual Network - Microsoft]
ILSVRC winner 2015


Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research

Abstract

1. Introduction



Deep Computer Vision

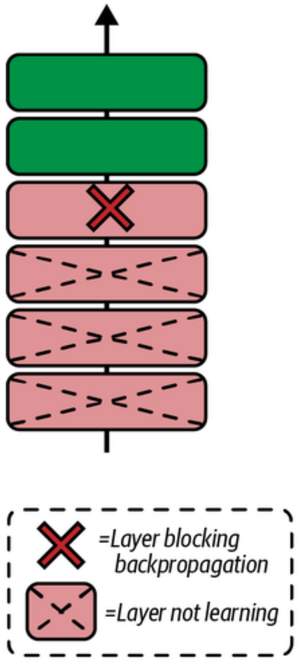
CNN Architectures

ILSVRC ImageNet

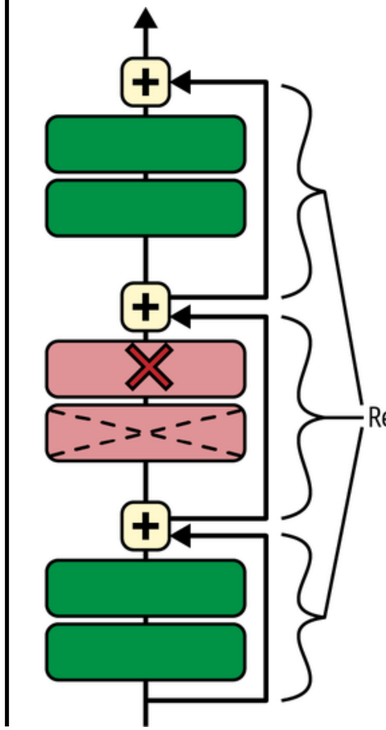
Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

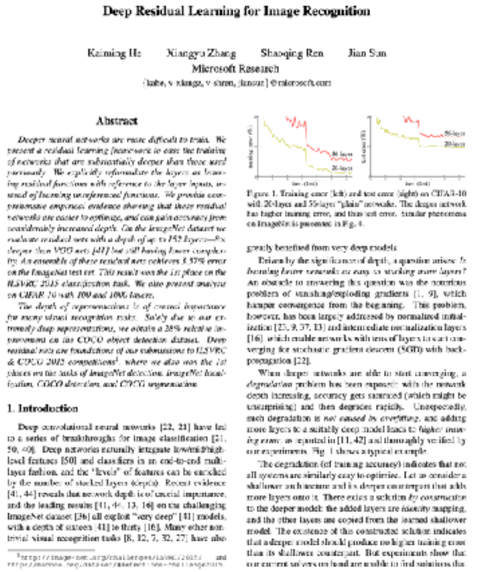
ResNet [Residual Network - Microsoft]
ILSVRC winner 2015



Regular Neural Network



Residual Neural Network



Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research

Abstract

Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers—8× deeper than VGG nets [41] but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. We also present analysis on CIFAR-10 with 100 and 1000 layers.

The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions[1], where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

1. Introduction

Deep convolutional neural networks [22, 21] have led to a series of breakthroughs for image classification [21, 50, 40]. Deep networks naturally integrate low/mid/high-level features [50] and classifiers in an end-to-end multi-layer fashion, and the "levels" of features can be enriched by the number of stacked layers (depth). Recent evidence [41, 44] reveals that network depth is of crucial importance, and the leading results [41, 44, 13, 16] on the challenging ImageNet dataset [36] all exploit "very deep" [41] models, with a depth of sixteen [41] to thirty [16]. Many other non-trivial visual recognition tasks [8, 12, 7, 32, 27] have also greatly benefited from very deep models.

Driven by the significance of depth, a question arises: *Is learning better networks as easy as stacking more layers?* An obstacle to answering this question was the notorious problem of vanishing/exploding gradients [1, 9], which hamper convergence from the beginning. This problem, however, has been largely addressed by normalized initialization [23, 9, 37, 13] and intermediate normalization layers [16], which enable networks with tens of layers to start converging for stochastic gradient descent (SGD) with back-propagation [22].

When deeper networks are able to start converging, a *degradation* problem has been exposed: with the network depth increasing, accuracy gets saturated (which might be unsurprising) and then degrades rapidly. Unexpectedly, such degradation is *not caused by overfitting*, and adding more layers to a suitably deep model leads to *higher training error*, as reported in [11, 42] and thoroughly verified by our experiments. Fig. 1 shows a typical example.

The degradation (of training accuracy) indicates that not all systems are similarly easy to optimize. Let us consider a shallower architecture and its deeper counterpart that adds more layers onto it. There exists a solution *by construction* to the deeper model: the added layers are identity mapping, and the other layers are copied from the learned shallower model. The existence of this constructed solution indicates that a deeper model should produce no higher training error than its shallower counterpart. But experiments show that our current solvers on hand are unable to find solutions that

Deep Computer Vision

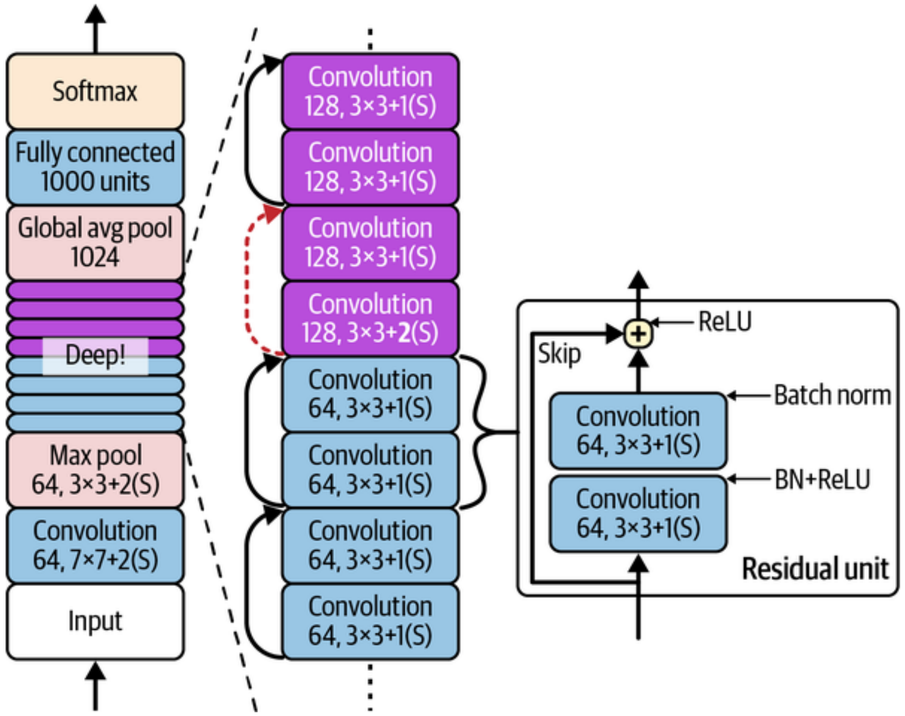
CNN Architectures

ILSVRC ImageNet

Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

ResNet [Residual Network - Microsoft]
ILSVRC winner 2015




Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research

Abstract

1. Introduction

Residual Neural Network

Deep Computer Vision

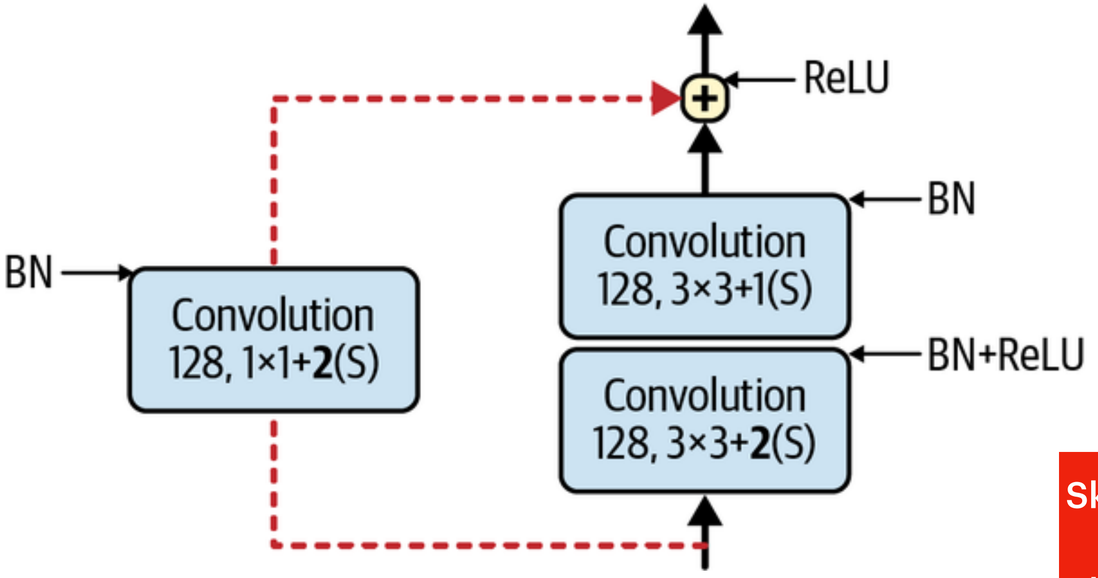
CNN Architectures

ILSVRC ImageNet

Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

ResNet [Residual Network - Microsoft]
ILSVRC winner 2015




Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research

Abstract

1. Introduction

Skip the connection when changing the size and depth of the feature map

Deep Computer Vision

CNN Architectures


Xception: Deep Learning with Depthwise Separable Convolutions

François Chollet
Google, Inc.

Abstract

1. Introduction

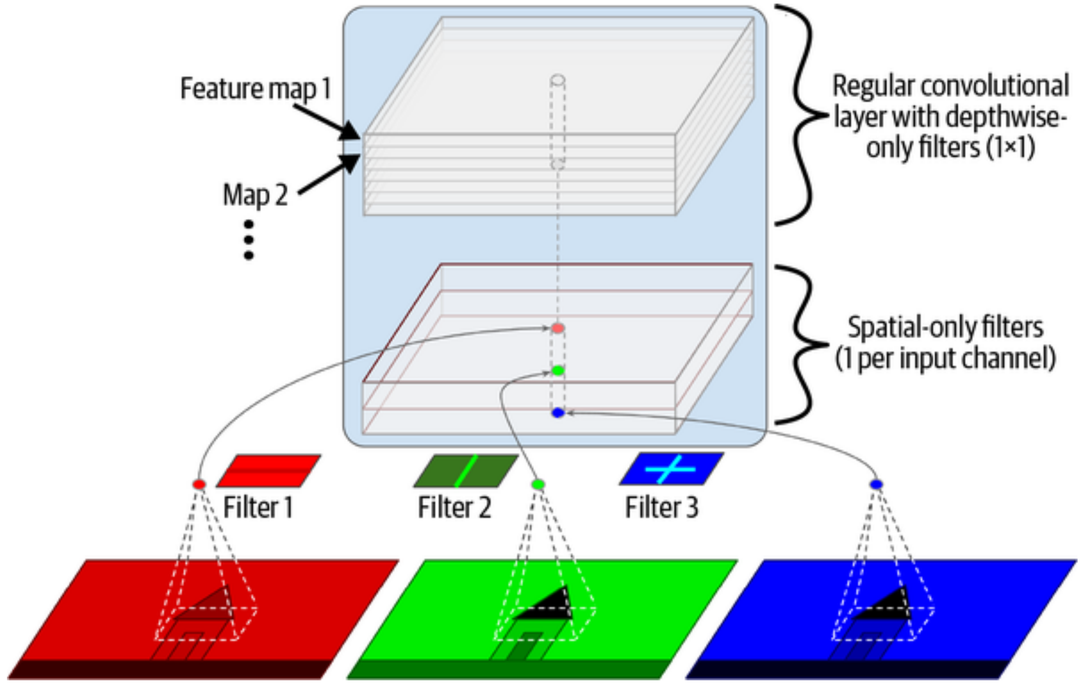
1.1. The Inception hypothesis

ILSVRC ImageNet

Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

Xception [Extreme Inception - 2016]
Keras Author: François Chollet



Separable Convolutional Layers

- *use fewer parameters*
- *less memory*
- *fewer computations*
- *they often perform better*

Deep Computer Vision

CNN Architectures

ILSVRC ImageNet

Error rate from 26% to 2.3% in 6 years!!

- *256 pixels height*
- *1000 class (120 dog race)*
- *Outputs describe CNN evolution*
- *Research progresses in deep learning*

Other Architectures:

ResNeX	MobileNet	EfficientNet
DenseNet	CSPNet	YOLO