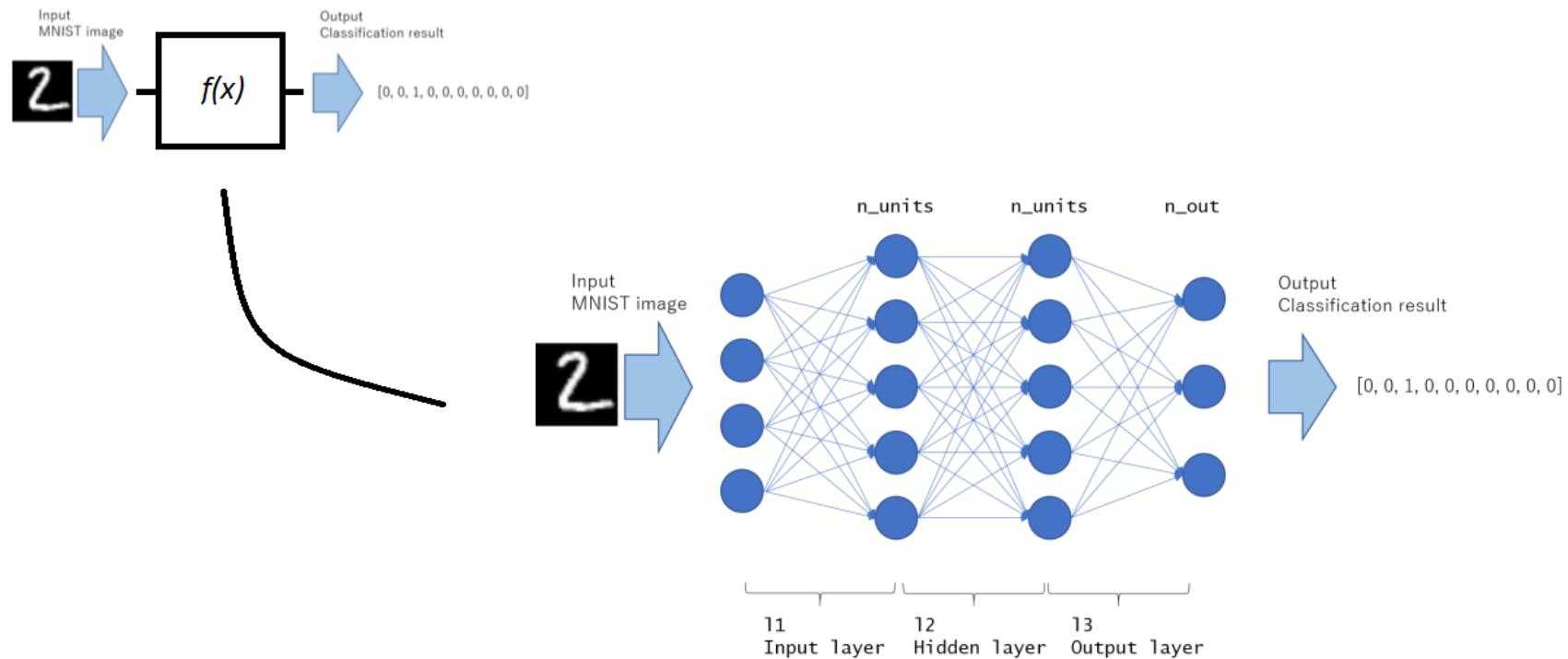


Redes Neurais Convolucionais

Lucas Araújo Pereira
Deep Learning Brasil
Instituto de Informática - UFG

Recapitulando...

- MLP



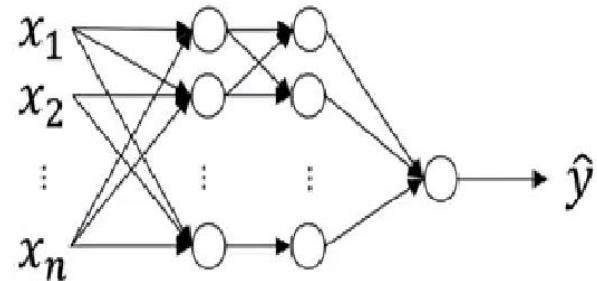
Problemas

- Complexidade



1920 px

1080 px



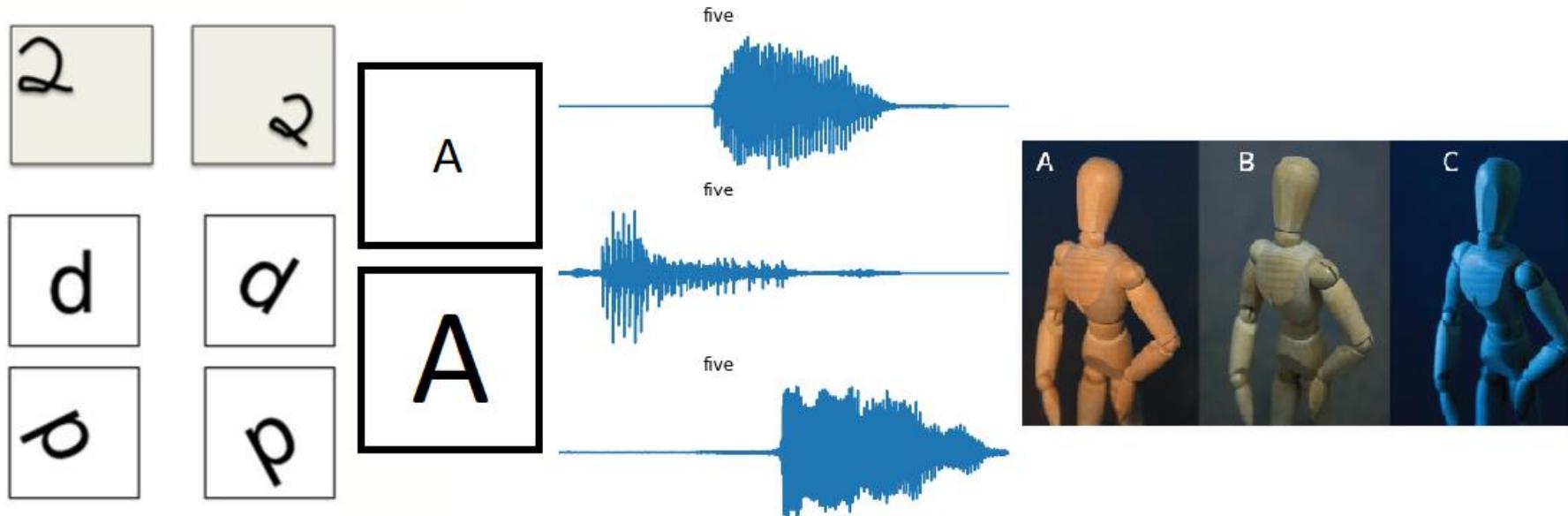
$1080 \times 1920 \times 3 \approx 6M$

1 hidden layer (camada oculta) = 1000 neurônios

≈ 6 bilhões de parâmetros treináveis

Problemas

- Variabilidade dos dados



Problemas

- Diferentes funções

0000000000000000
1111111111111111
2222222222222222
3333333333333333
4444444444444444
5555555555555555
6666666666666666
7777777777777777
8888888888888888
9999999999999999

MNIST 10 dígitos



A 10x10 grid of 100 small images, each depicting a different scene or object from the game. The images include various landscapes, characters, and items, providing a comprehensive visual overview of the game's content.

IMAGENET 1000 objetos/seres



COCO
91 objetos/seres
Localização, relação semântica, pose

Algo em comum?

Algo em comum?

- Relação Estrutural
- Imagens:
 - Relação espacial entre os dados (pixels)
- Audio, texto, vídeo, sequências temporais:
 - Relação temporal entre os dados

E se...

- Yann LeCun, 1989

Generalization and Network Design Strategies

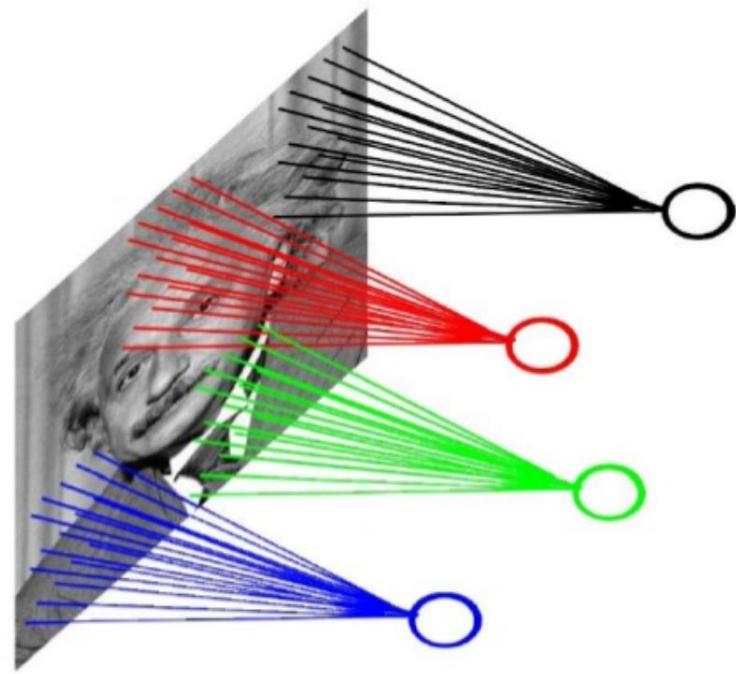
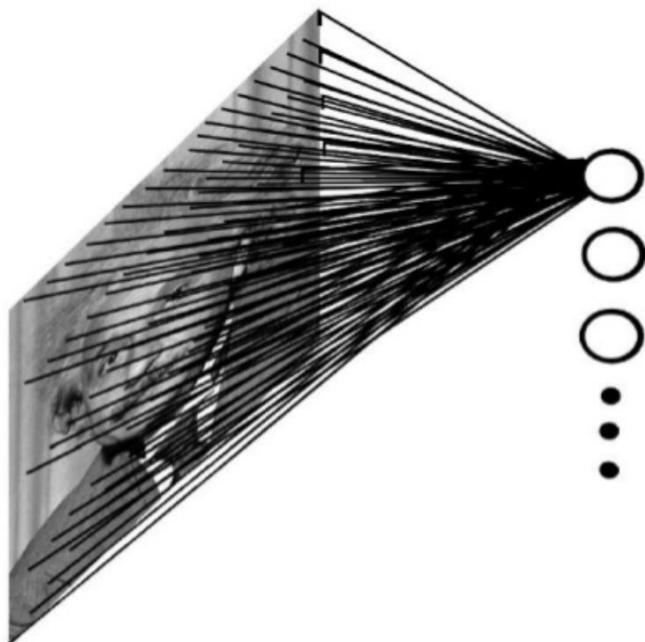
Yann le Cun *

Department of Computer Science, University of Toronto
Toronto, Ontario, M5S 1A4. CANADA.

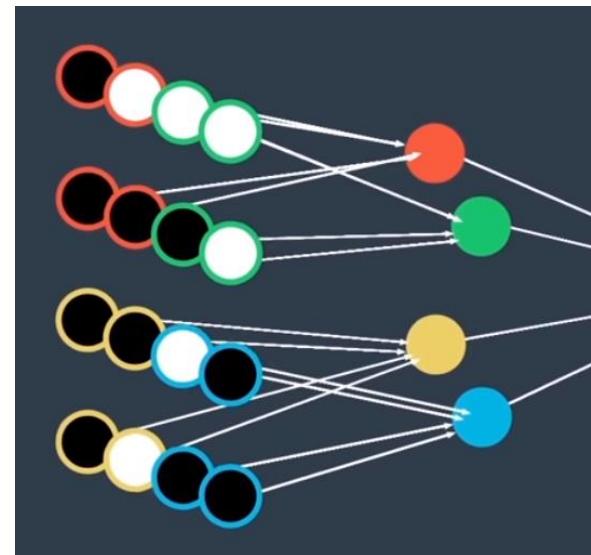
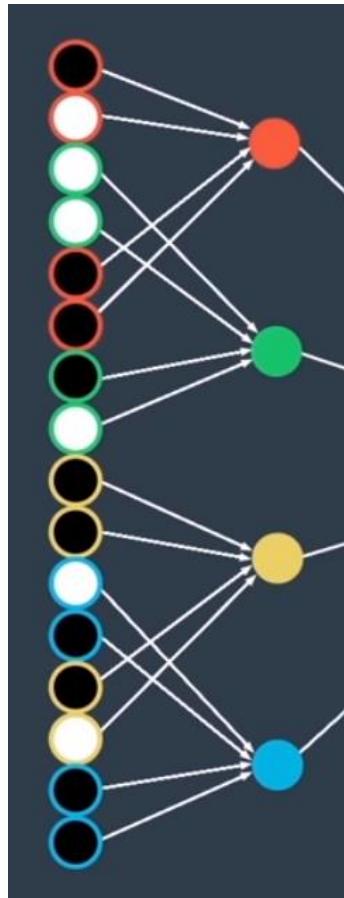
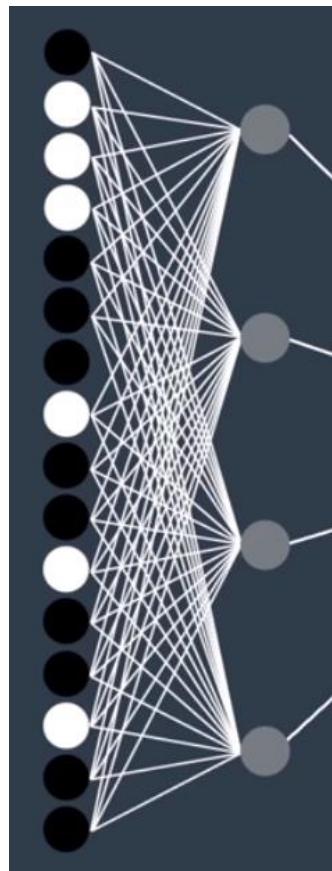
Abstract

An interesting property of connectionist systems is their ability to learn from examples. Although most recent work in the field concentrates on reducing learning times, the most important feature of a learning machine is its generalization performance. It is usually accepted that good generalization performance on real-world problems cannot be achieved unless some *a priori* knowledge about the task is built into the system. Back-propagation networks provide a way of specifying such knowledge by imposing constraints both on the architecture of the network and on its weights. In general, such constraints can be considered as particular transformations of the parameter space

Ligações Locais



Ligações Locais



Créditos imagens: Udacity

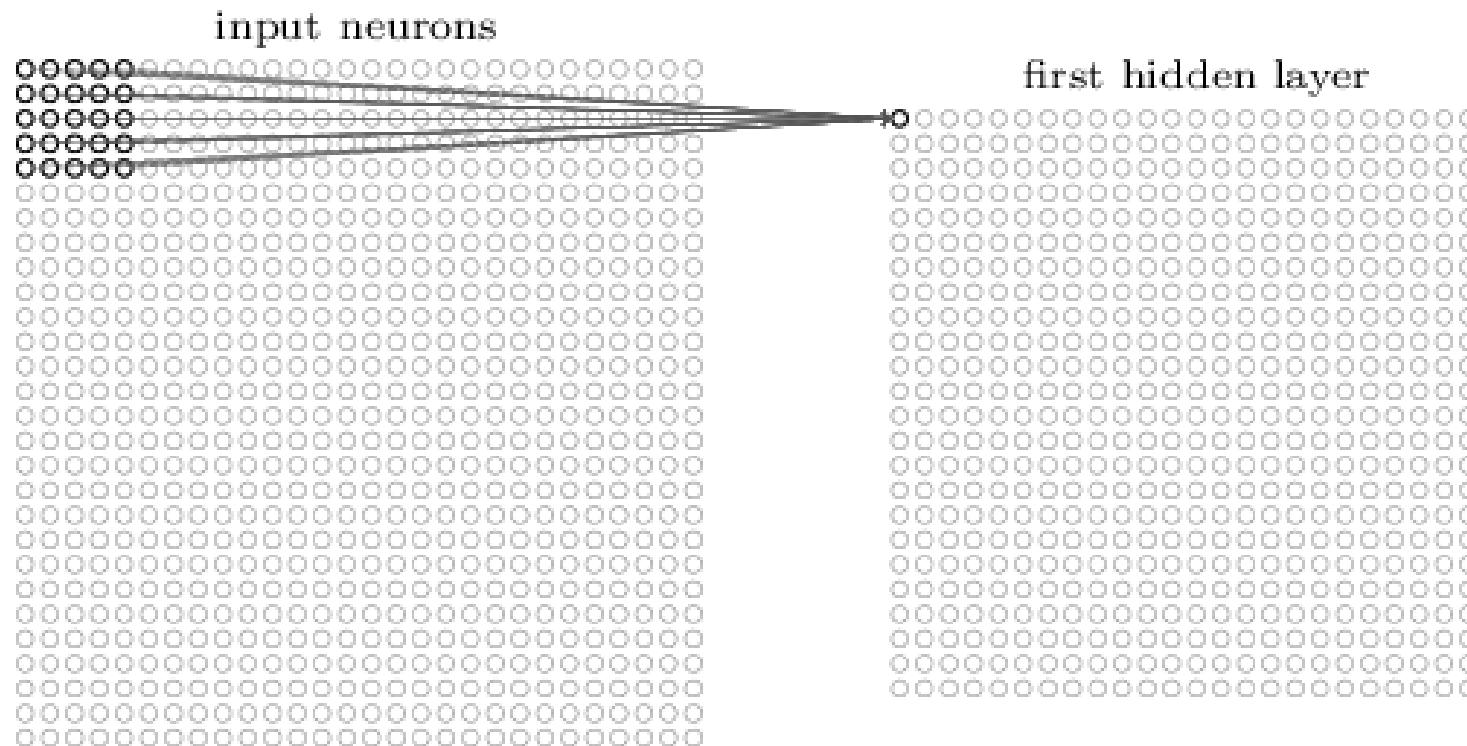
Ligações Locais

- Por que?

Ligações Locais

- Por que?
 - Relação estrutural
 - Diminuição de complexidade

Compartilhamento de Parâmetros



Compartilhamento de Parâmetros

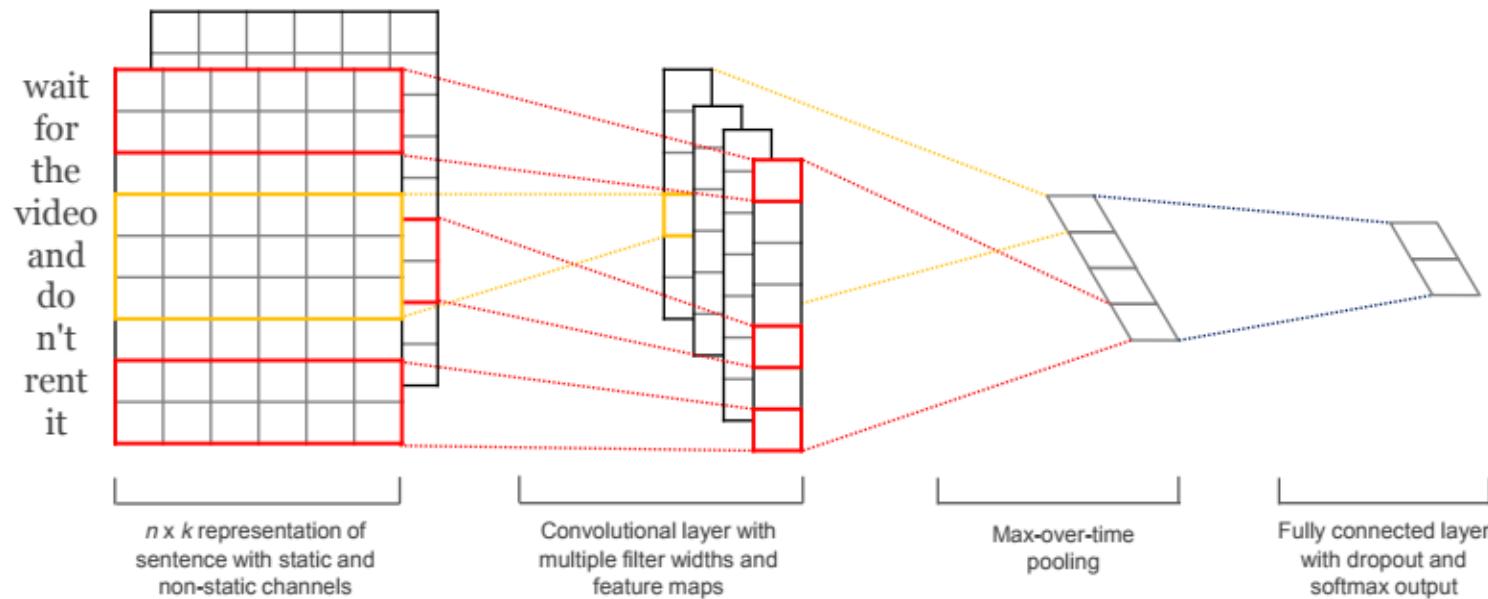
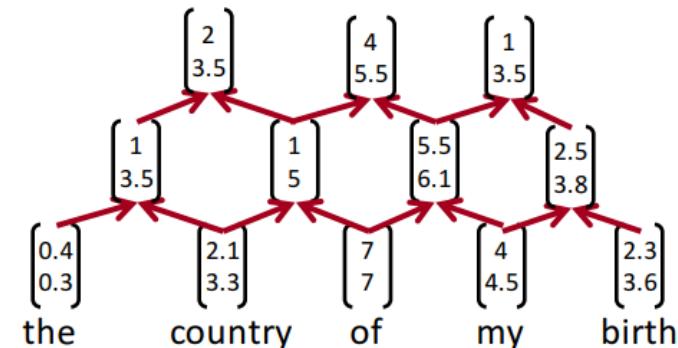
- Por que?

Compartilhamento de Parâmetros

- Por que?
 - Variabilidade dos dados (translação)

Só imagens?

Só imagens? Não!



Testando a hipótese

- Dataset: MNIST
 - Imagens de dígitos (0-9) escritos à mão
 - 60 mil exemplos de treino
 - 10 mil exemplos de teste

Testando a hipótese

- MNIST



Testando a hipótese

- Redes testadas

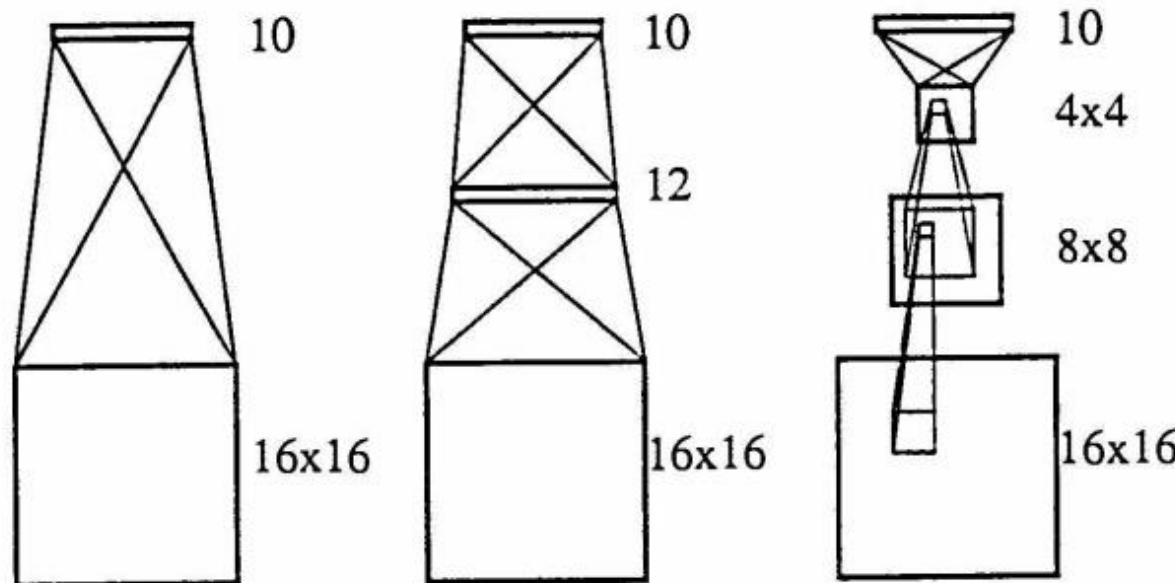


Figure 4: three network architectures Net-1, Net-2 and Net-3

Testando a hipótese

- Redes testadas

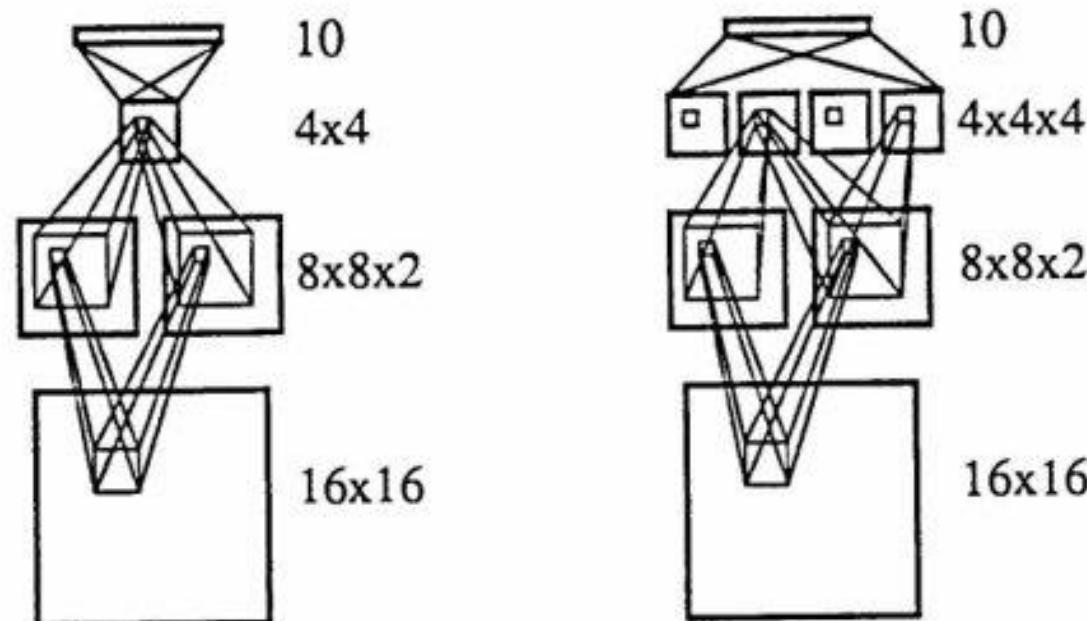
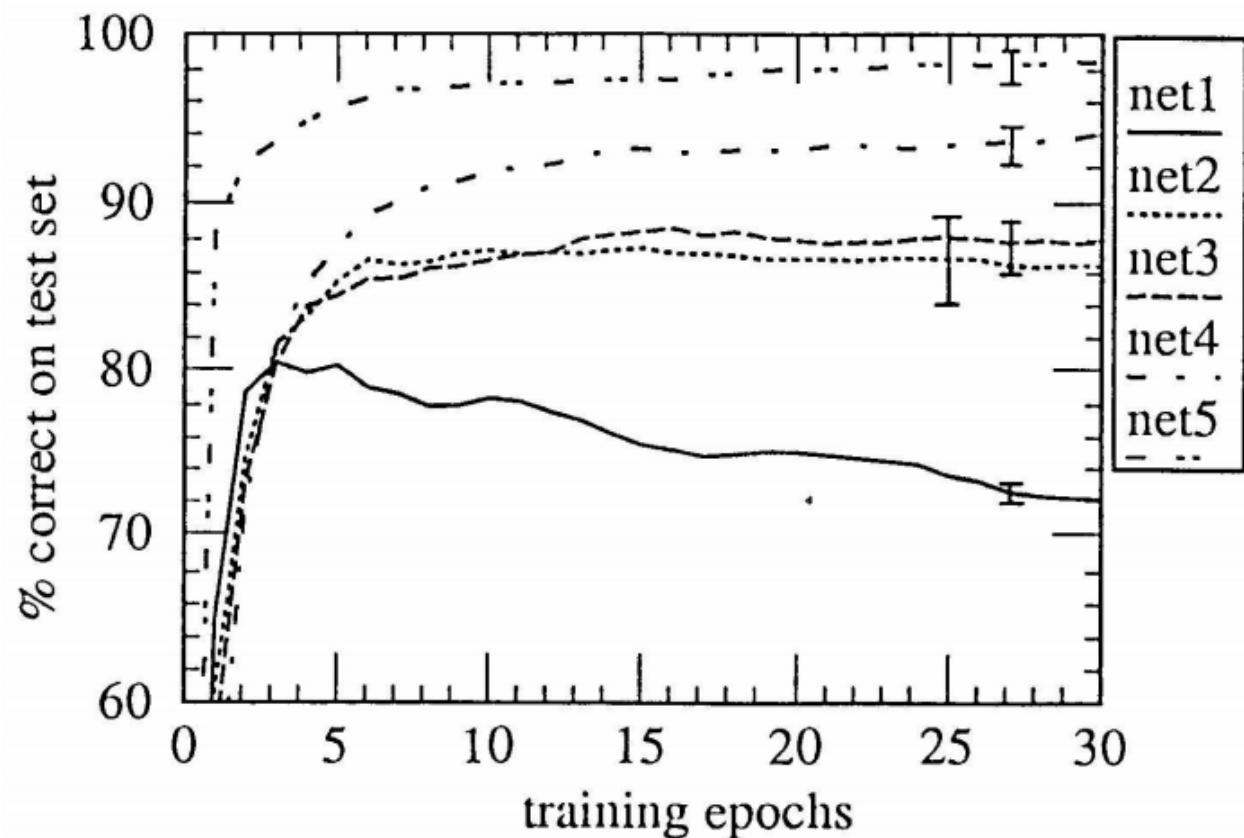


Figure 5 two network architectures with shared weights: Net-4 and Net-5

Testando a hipótese

- Resultados



Testando a hipótese

- Resultados

network architecture	links	weights	performance
single layer network	2570	2570	80 %
two layer network	3240	3240	87 %
locally connected	1226	1226	88.5 %
constrained network	2266	1132	94 %
constrained network 2	5194	1060	98.4 %

Cadê as convoluções?

Cadê as convoluções?

- LeCun, 1990

*Handwritten Digit Recognition with a
Back-Propagation Network*

Y. Le Cun, B. Boser, J. S. Denker, D. Henderson,
R. E. Howard, W. Hubbard, and L. D. Jackel
AT&T Bell Laboratories, Holmdel, N. J. 07733

ABSTRACT

We present an application of back-propagation networks to handwritten digit recognition. Minimal preprocessing of the data was required, but architecture of the network was highly constrained and specifically designed for the task. The input of the network consists of normalized images of isolated digits. The method has 1% error rate and about a 9% reject rate on zipcode digits provided by the U.S. Postal Service.

Cadê as convoluções?

- LeCun, 1990

that has a local receptive field, and store the states of this neuron in corresponding locations in a layer called a *feature map* (see figure 3). This operation is equivalent to a convolution with a small size kernel, followed by a squashing function. The process can be performed in parallel by implementing the feature map as a plane

O que é uma convolução mesmo?

Convolução

- Definição formal (1D)

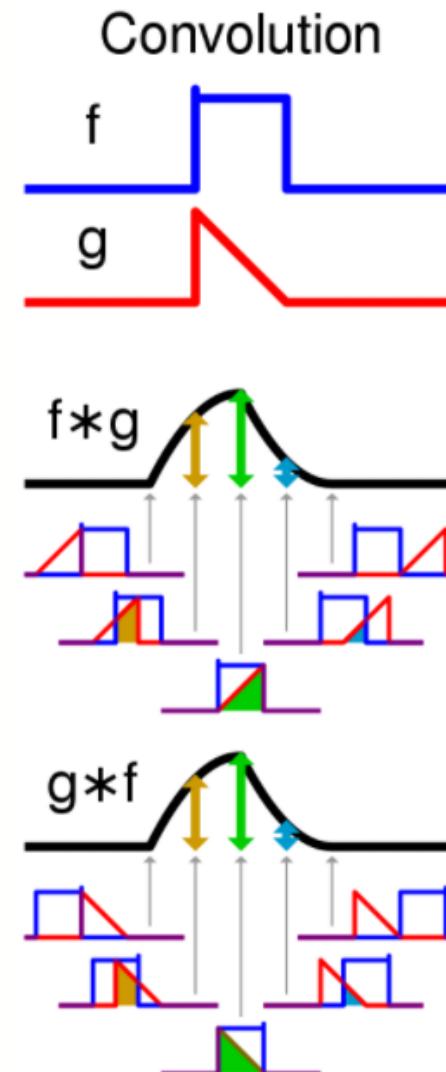
$$s(t) = \int x(a)w(t-a)da.$$

$$s(t) = (x * w)(t).$$

$$s(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a).$$

Convolução

- Visualização (1D)



Convolução

- Definição formal (2D)

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(m, n)K(i - m, j - n).$$

Convolução

- Visualização (2D)

1	1	1	0	0
0	1	1	1	0
0	0	1 _{x1}	1 _{x0}	1 _{x1}
0	0	1 _{x0}	1 _{x1}	0 _{x0}
0	1	1 _{x1}	0 _{x0}	0 _{x1}

Image

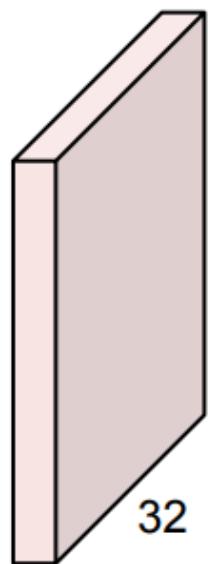
4	3	4
2	4	3
2	3	4

Convolved
Feature

Convolução

- Vocabulário

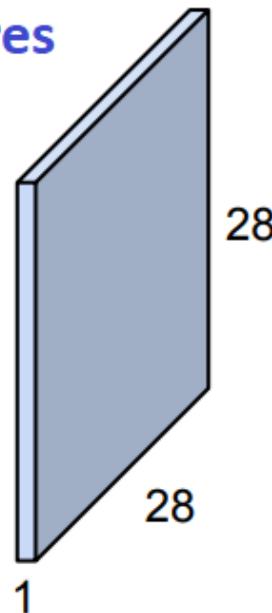
32x32 imagem



5x5

filtro ou kernel

mapa de ativações ou
features



Convolução

- Paper para referência ([link](#))

A guide to convolution arithmetic for deep learning

Vincent Dumoulin¹★ and Francesco Visin²★†

★MILA, Université de Montréal
†AIRLab, Politecnico di Milano

A avó das CNNs

- LeNet-5, 1998

PROC. OF THE IEEE, NOVEMBER 1998

1

Gradient-Based Learning Applied to Document Recognition

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

Abstract—

Multilayer Neural Networks trained with the backpropagation algorithm constitute the best example of a successful Gradient-Based Learning technique. Given an appropriate network architecture, Gradient-Based Learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns such as handwritten characters, with minimal preprocessing. This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit recognition task. Convolutional Neural Networks, that are specifically designed to deal with the variability of 2D shapes, are

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.

A avó das CNNs

- LeNet-5

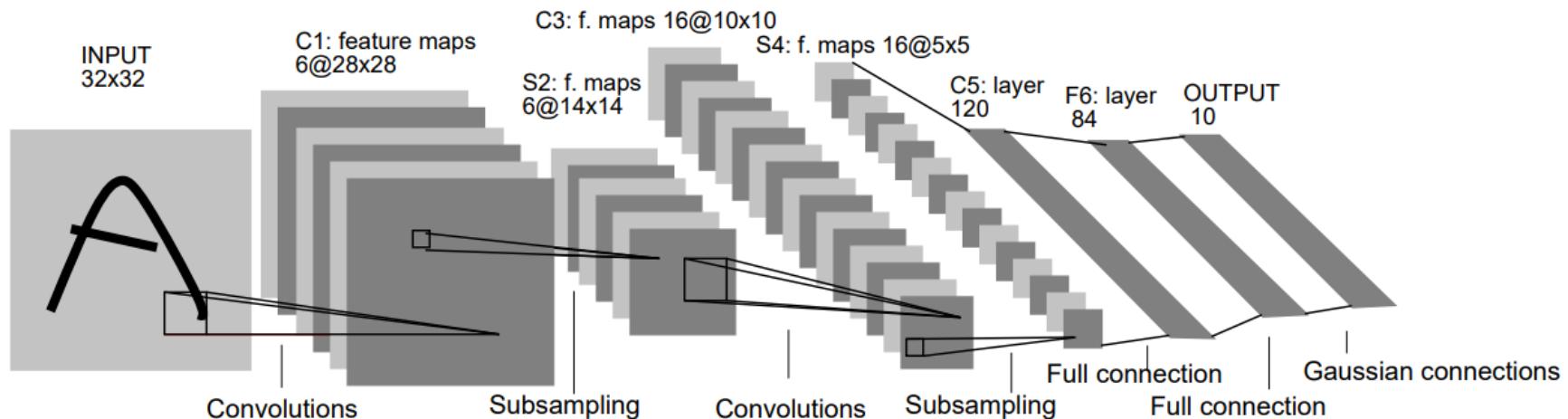
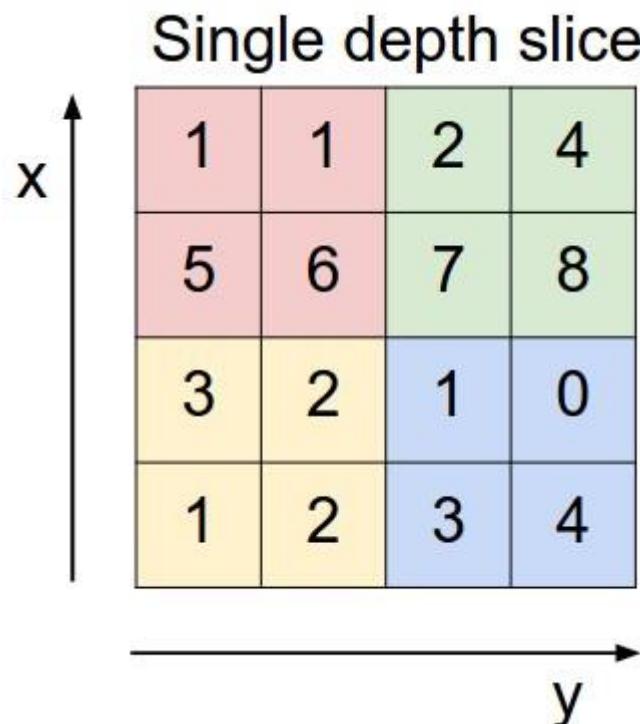


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Pooling ou Subsampling



Pooling ou Subsampling

- Tipos
 - MaxPool - seleciona a ativação máxima do kernel
 - MinPool - seleciona a ativação mínima do kernel
 - AvgPool - calcula a média das ativações do kernel

Pooling ou Subsampling

- Pra que?

Pooling ou Subsampling

- Pra que? (Há controvérsias...)

In all cases, pooling helps to make the representation approximately **invariant** to small translations of the input. Invariance to translation means that if we

do not change. See figure 9.8 for an example of how this works. *Invariance to local translation can be a useful property if we care more about whether some feature is present than exactly where it is.* For example, when determining whether an

[Deep Learning Book, Goodfellow]

A mãe das CNNs

- AlexNet, 2012

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.

A mãe das CNNs

- AlexNet

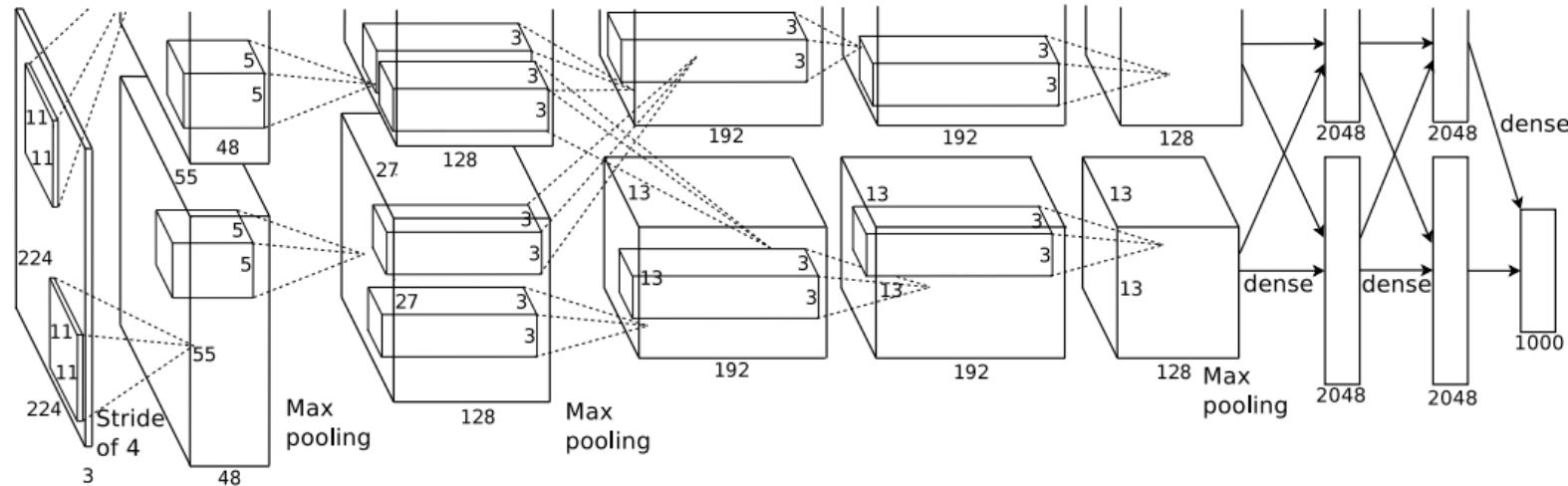


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

ImageNet

- Dataset: [link](#)
 - 1.2M de imagens em "alta" resolução
 - 1000 categorias de objetos e seres a serem classificados
 - 2 tipos de scores: top-1 e top-5

ImageNet

- Performance da AlexNet

Rank	Top-5 Error	Algorithm	Author
1	15.3%	Deep CNN	Krizhevsky et al.
2	26.2%	Non deep learning	Gunji et al.
3	27.0%	Non deep learning	Simonyan et al.
4	27.1%	Non deep learning	Perronin et al.
5	29.6%	Non deep learning	van de Sande et al.

A prole... (os mais famosos)

- VGG, 2014

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.

A prole... (os mais famosos)

- VGG

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

ImageNet

- Performance da VGG

Table 4: ConvNet performance at multiple test scales.

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
B	256	224,256,288	28.2	9.6
C	256	224,256,288	27.7	9.2
	384	352,384,416	27.8	9.2
	[256; 512]	256,384,512	26.3	8.2
D	256	224,256,288	26.6	8.6
	384	352,384,416	26.5	8.6
	[256; 512]	256,384,512	24.8	7.5
E	256	224,256,288	26.9	8.7
	384	352,384,416	26.7	8.6
	[256; 512]	256,384,512	24.8	7.5

A prole... (os mais famosos)

- Inception, 2014

Going deeper with convolutions

Christian Szegedy

Google Inc.

Wei Liu

University of North Carolina, Chapel Hill

Yangqing Jia

Google Inc.

Pierre Sermanet

Google Inc.

Scott Reed

University of Michigan

Dragomir Anguelov

Google Inc.

Dumitru Erhan

Google Inc.

Vincent Vanhoucke

Google Inc.

Andrew Rabinovich

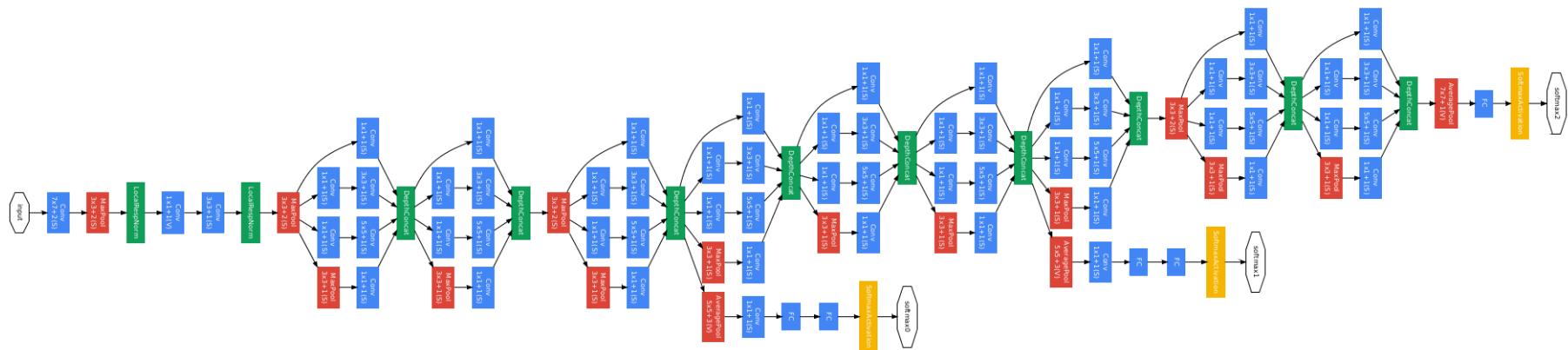
Google Inc.

Abstract

We propose a deep convolutional neural network architecture codenamed Inception, which was responsible for setting 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. This was achieved by a carefully crafted design that allows for increasing 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.

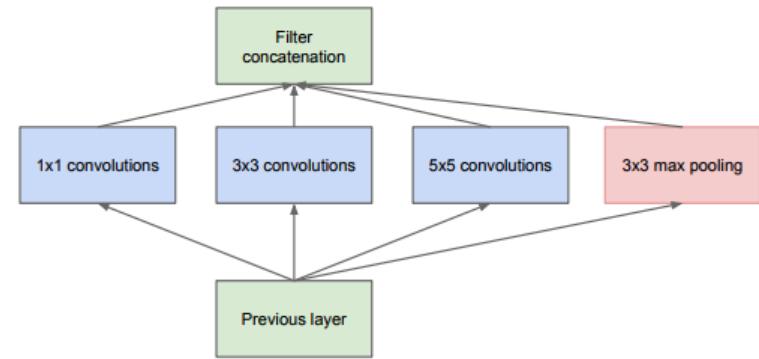
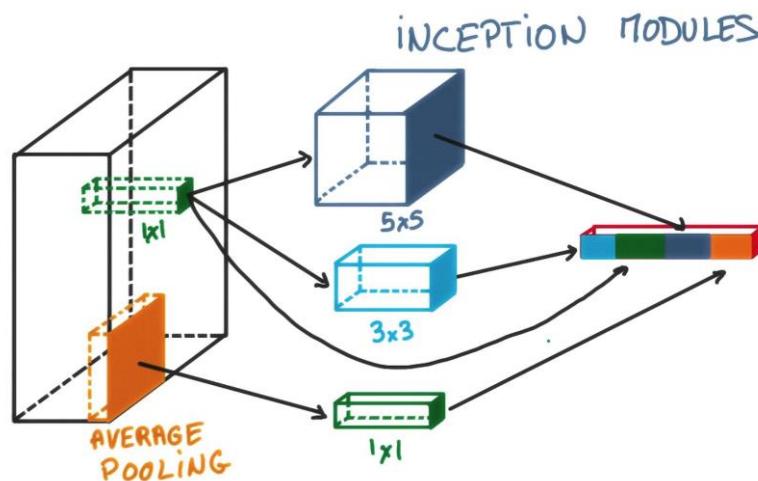
A prole... (os mais famosos)

- Inception

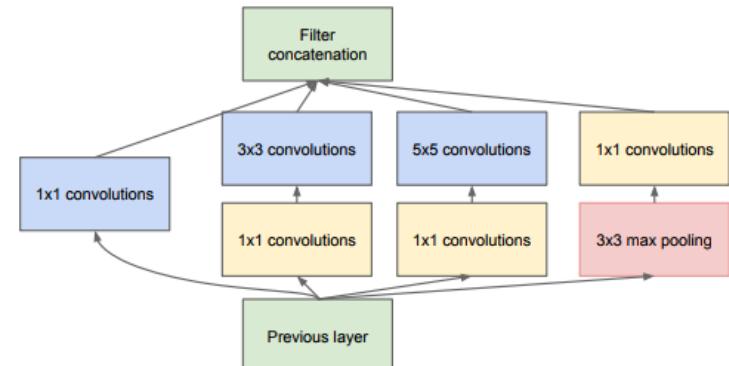


A prole... (os mais famosos)

- Inception



(a) Inception module, naïve version



(b) Inception module with dimension reductions

A prole... (os mais famosos)

- Inception

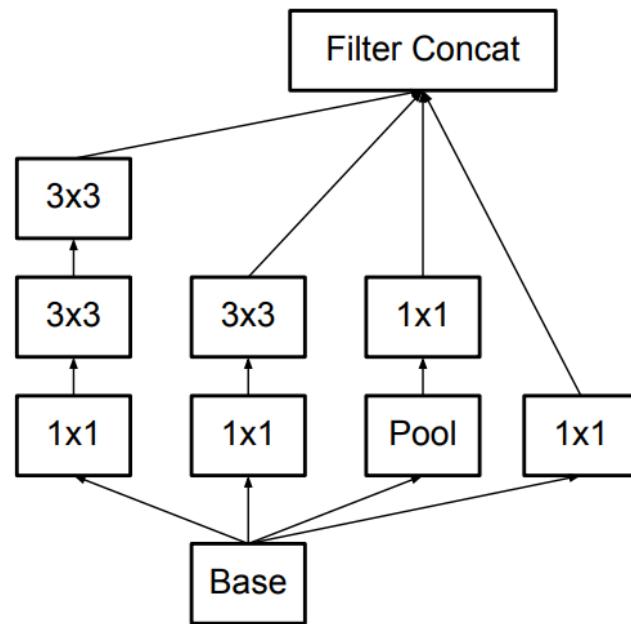


Figure 5. Inception modules where each 5×5 convolution is replaced by two 3×3 convolution, as suggested by principle 3 of Section 2.

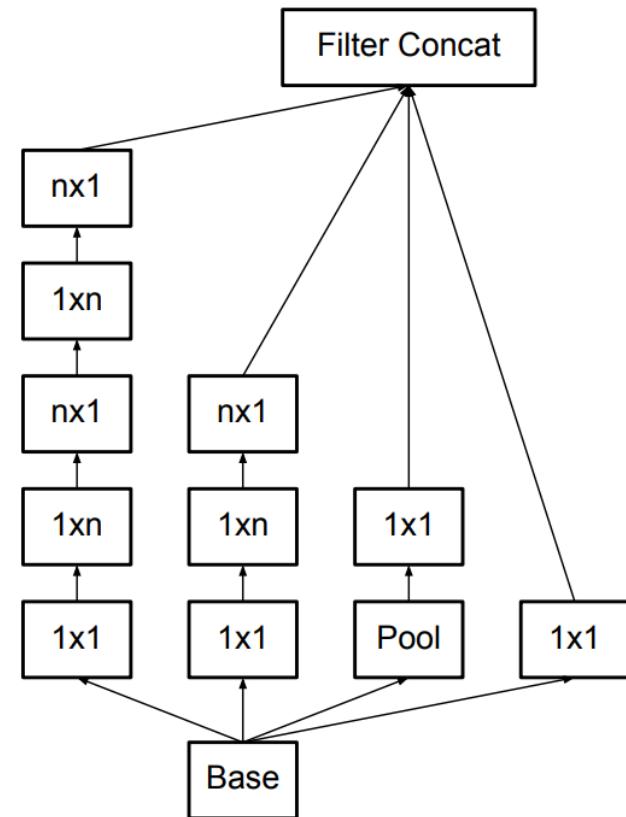


Figure 6. Inception modules after the factorization of the $n \times n$ convolutions. In our proposed architecture, we chose $n = 7$ for the 17×17 grid. (The filter sizes are picked using principle 3)

ImageNet

- Performance da Inception

Team	Year	Place	Error (top-5)	Uses external data
SuperVision	2012	1st	16.4%	no
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	no
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	no
VGG	2014	2nd	7.32%	no
GoogLeNet	2014	1st	6.67%	no

A prole... (os mais famosos)

- ResNet, 2015

Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun

Microsoft Research

{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

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 unreference 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 competitions¹, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

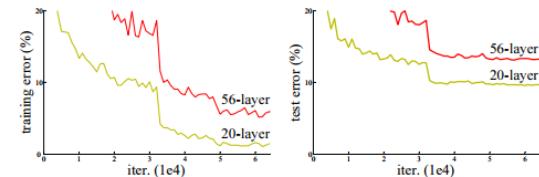


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer “plain” networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

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

A prole... (os mais famosos)

- ResNet

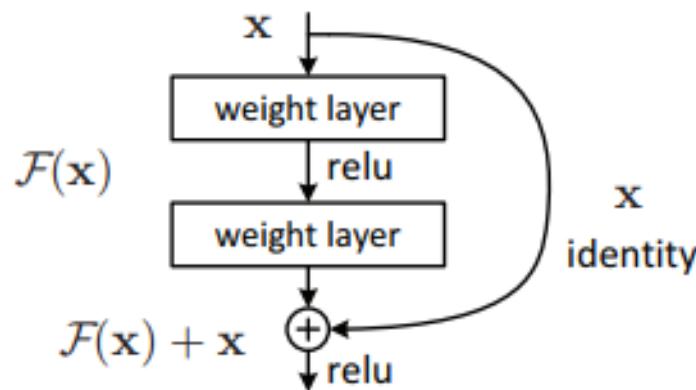
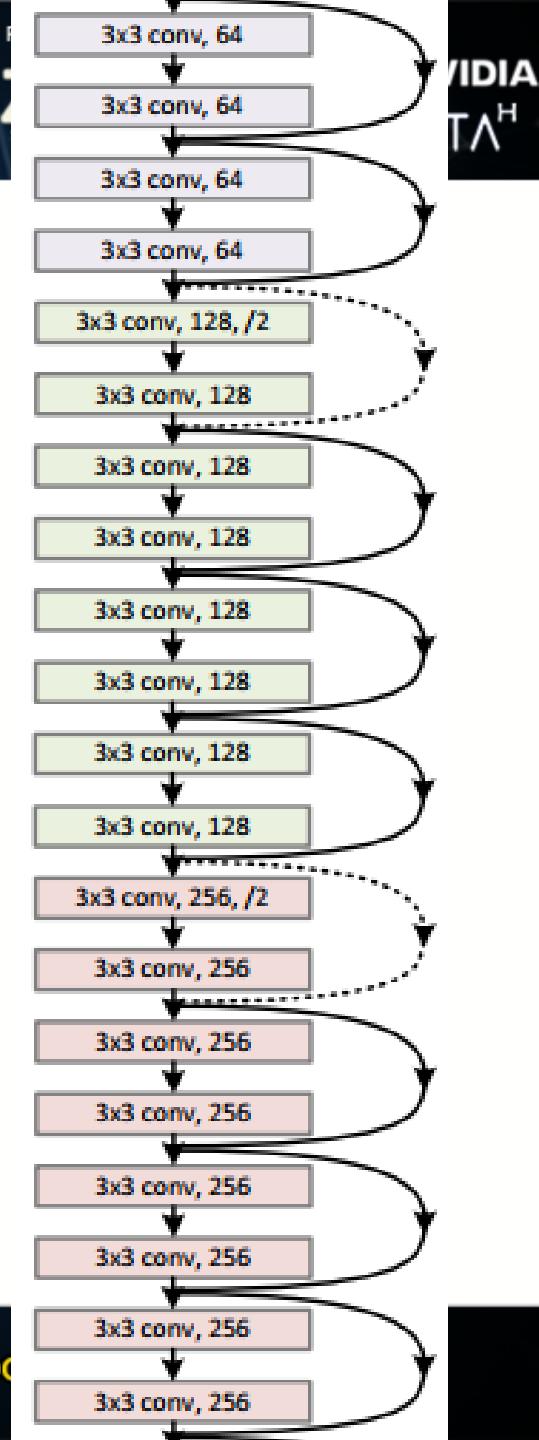


Figure 2. Residual learning: a building block.



Será?



A prole... (os mais famosos)

- ResNet

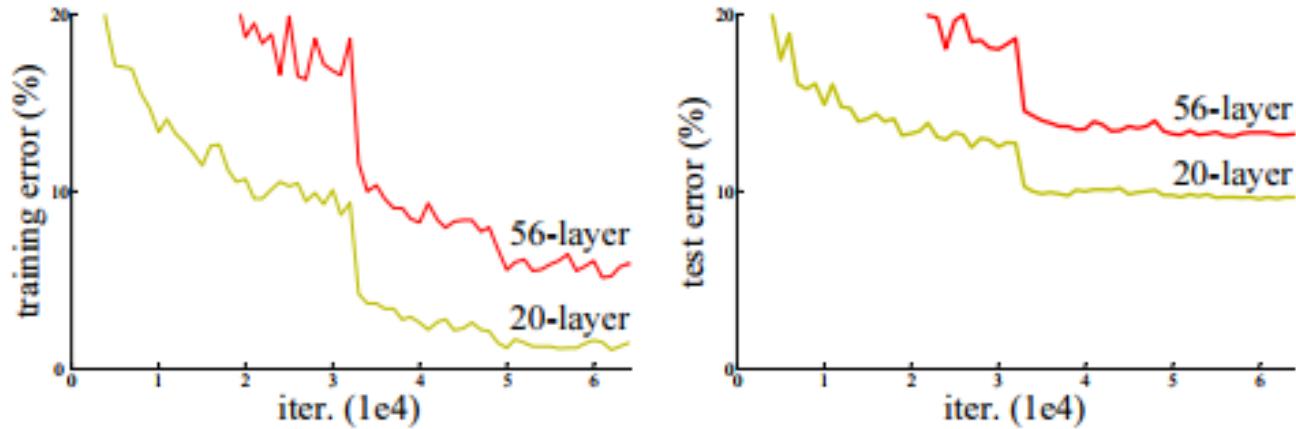


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer “plain” networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

ImageNet

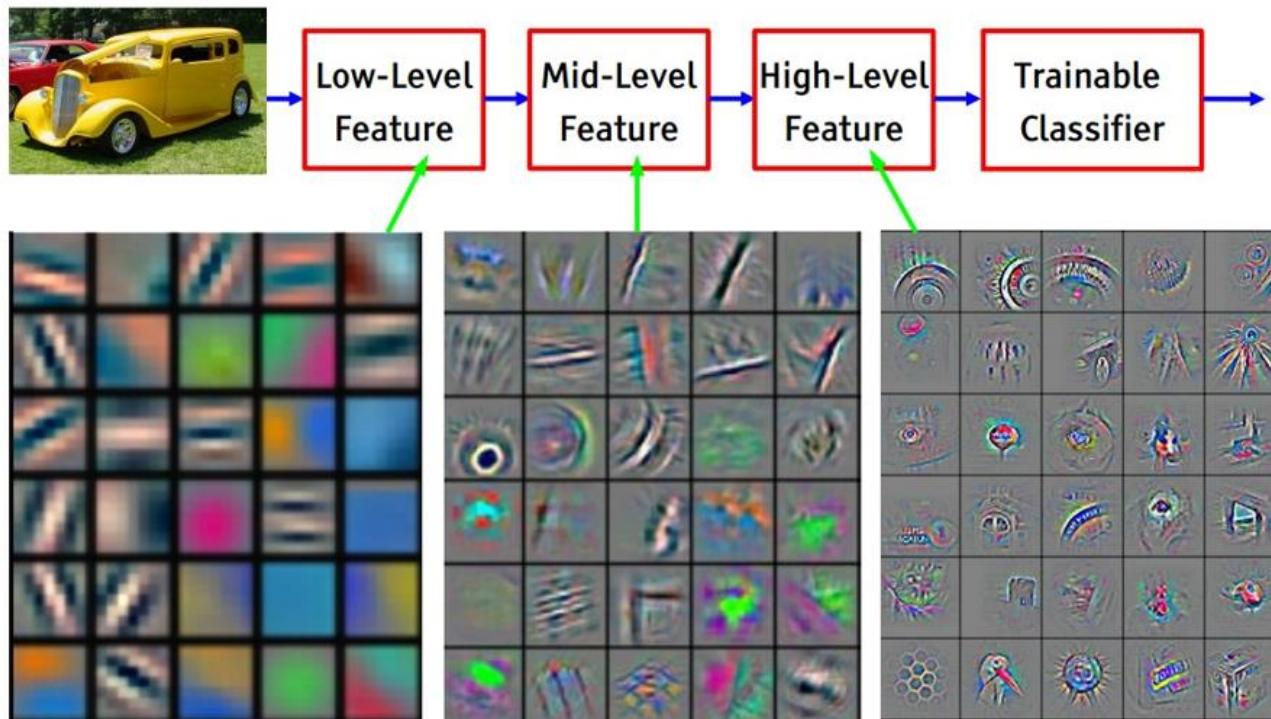
- Performance da ResNet

method	top-1 err.	top-5 err.
VGG [41] (ILSVRC'14)	-	8.43 [†]
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	19.38	4.49

Table 4. Error rates (%) of **single-model** results on the ImageNet validation set (except [†] reported on the test set).

Visualização das camadas

- Acredita(va)-se



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Não é tão intuitivo assim...

- Enganando uma CNN

Intriguing properties of neural networks

Christian Szegedy

Google Inc.

Wojciech Zaremba

New York University

Ilya Sutskever

Google Inc.

Joan Bruna

New York University

Dumitru Erhan

Google Inc.

Ian Goodfellow

University of Montreal

Rob Fergus

New York University
Facebook Inc.

Abstract

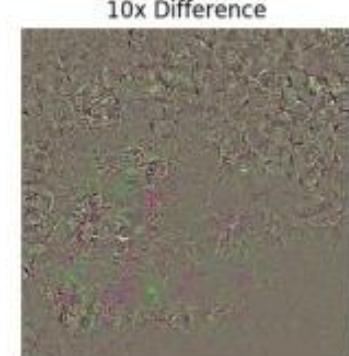
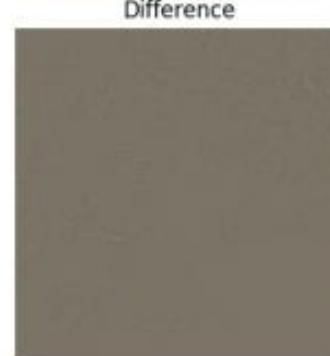
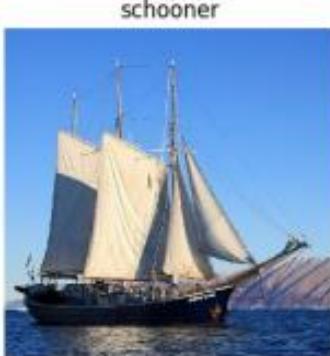
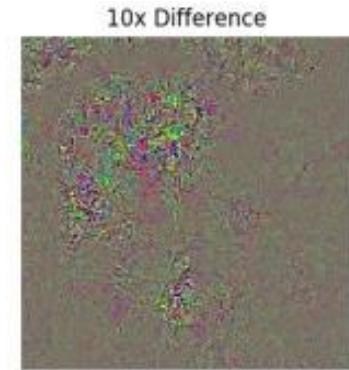
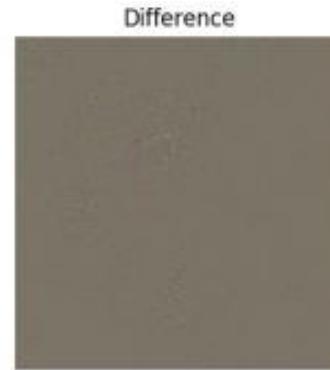
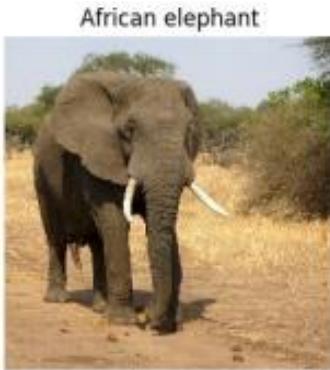
Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the reason they succeed, it also causes them to learn uninterpretable solutions that could have counter-intuitive properties. In this paper we report two such properties.

First, we find that there is no distinction between individual high level units and random linear combinations of high level units, according to various methods of unit analysis. It suggests that it is the space, rather than the individual units, that contains the semantic information in the high layers of neural networks.

Second, we find that deep neural networks learn input-output mappings that are fairly discontinuous to a significant extent. We can cause the network to misclassify an image by applying a certain hardly perceptible perturbation, which is found by maximizing the network's prediction error. In addition, the specific nature of these perturbations is not a random artifact of learning: the same perturbation can cause a different network, that was trained on a different subset of the dataset, to misclassify the same input.

Não é tão intuitivo assim...

- Enganando uma CNN



Não é tão intuitivo assim...

- Enganando uma CNN

Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

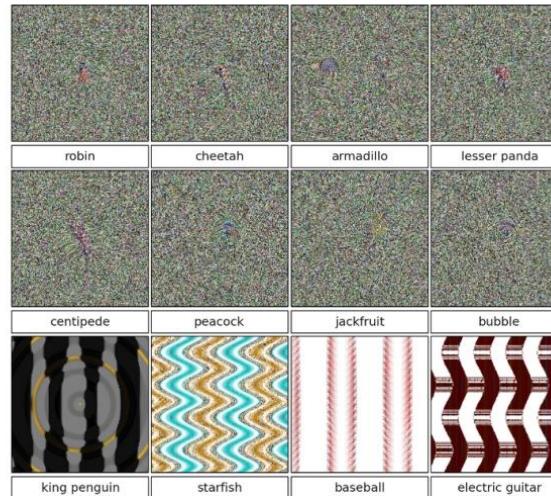
Anh Nguyen
University of Wyoming
anguyen8@uwyo.edu

Jason Yosinski
Cornell University
yosinski@cs.cornell.edu

Jeff Clune
University of Wyoming
jeffclune@uwyo.edu

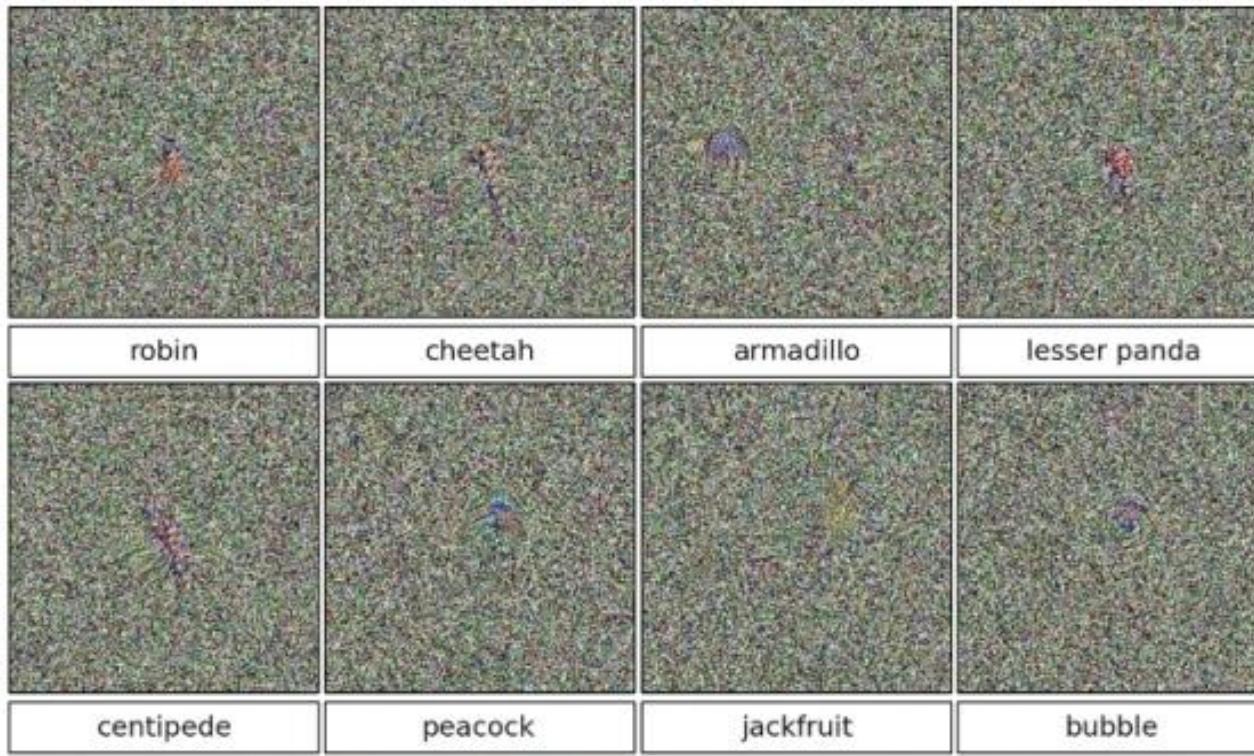
Abstract

Deep neural networks (DNNs) have recently been achieving state-of-the-art performance on a variety of pattern-recognition tasks, most notably visual classification problems. Given that DNNs are now able to classify objects in images with near-human-level performance, questions naturally arise as to what differences remain between computer and human vision. A recent study [30] revealed that changing an image (e.g. of a lion) in a way imperceptible to humans can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library). Here we show a related result: it is easy to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs believe to be recognizable objects with 99.99% confidence (e.g. labeling with certainty that white noise static is a lion). Specifically we take convolutional neu-



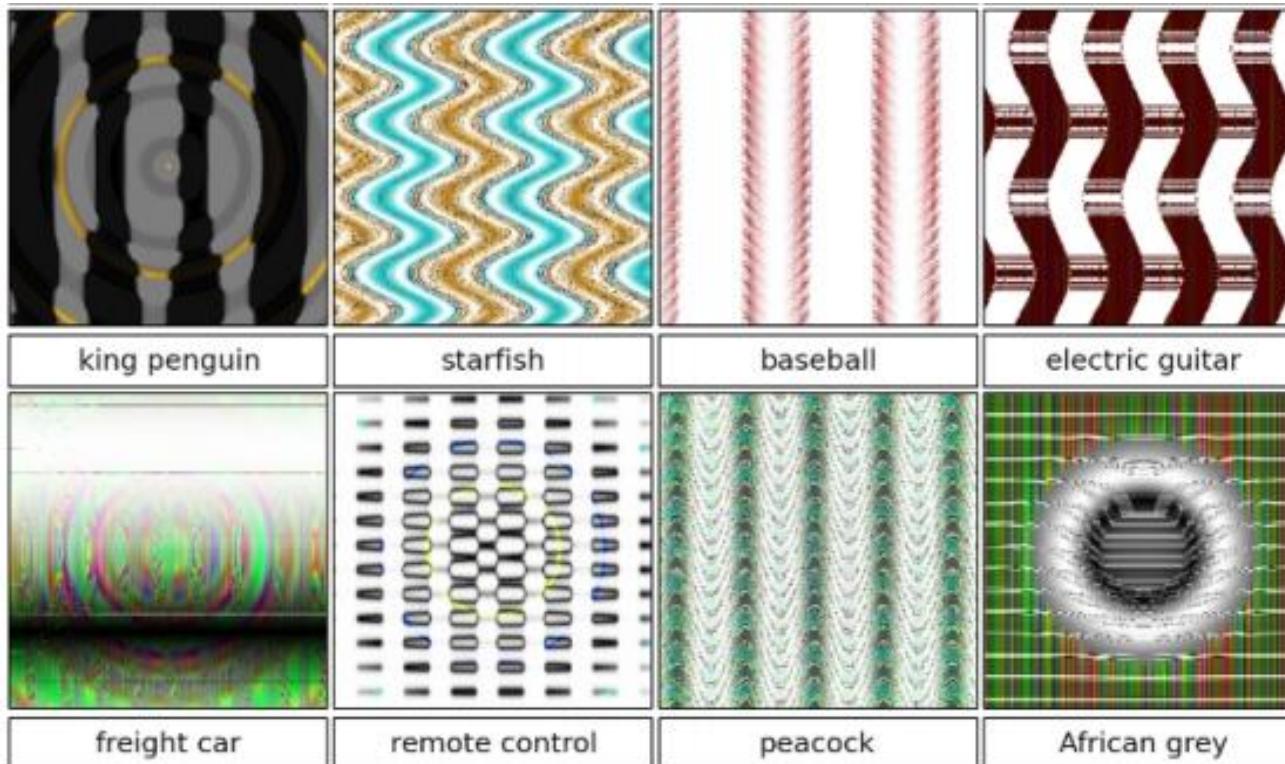
Não é tão intuitivo assim...

- Enganando uma CNN



Não é tão intuitivo assim...

- Enganando uma CNN



Não é tão intuitivo assim...

- Enganando uma CNN

Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with $\geq 99.6\%$ certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (*top*) or indirectly (*bottom*) encoded.

Algoritmos Genéticos contra-atacam!

Não é tão intuitivo assim...

- Enganando uma CNN (Dez 2017)
-

Adversarial Patch

Tom B. Brown, Dandelion Mané*, Aurko Roy, Martín Abadi, Justin Gilmer
{tombrown,dandelion,aurkor,abadi,gilmer}@google.com

Abstract

We present a method to create universal, robust, targeted adversarial image patches in the real world. The patches are universal because they can be used to attack any scene, robust because they work under a wide variety of transformations, and targeted because they can cause a classifier to output any target class. These adversarial patches can be printed, added to any scene, photographed, and presented to image classifiers; even when the patches are small, they cause the classifiers to ignore the other items in the scene and report a chosen target class.

Não é tão intuitivo assim...

- Enganando uma CNN (Dez 2017)



Figure 1: A real-world attack on VGG16, using a physical patch generated by the white-box ensemble method described in Section 3. When a photo of a tabletop with a banana and a notebook (top photograph) is passed through VGG16, the network reports class 'banana' with 97% confidence (top plot). If we physically place a sticker targeted to the class "toaster" on the table (bottom photograph), the photograph is classified as a toaster with 99% confidence (bottom plot). See the following video for a full demonstration: <https://youtu.be/i1sp4X57L4>

Por hoje é só....

De convolução. Tem parametrização ainda...