CSC321 Lecture 12
Recent advances in conv nets

Roger Grosse and Nitish Srivastava

February 12, 2015

Overview

At this point, you know enough to understand the state-of-the-art methods for many vision tasks!

- E.g. Krizhevsky et al., 2012, ImageNet classification with deep convolutional neural networks

In this lecture, we'll look at what's happened since Geoff made the Coursera videos. (Quite a lot!)

Overview

Biggest “advances” in machine learning:

Overview

Biggest "advances" in machine learning:

1. Datasets have gotten bigger
   - Lots of images and videos from digital cameras
   - Text from web pages
   - Ability to collect lots of labels with Mechanical Turk

Overview

Biggest "advances" in machine learning:

1. Datasets have gotten bigger
   - Lots of images and videos from digital cameras
   - Text from web pages
   - Ability to collect lots of labels with Mechanical Turk
2. Computers have gotten faster
   - Moore's Law
   - Graphics processing units (GPUs)

Data and computing power

Graphics processing units (GPUs) are a kind of highly parallel processor. They're good at performing computations with lots of independent operations and little control overhead. This is a perfect fit to neural nets!

Data and computing power

Graphics processing units (GPUs) are a kind of highly parallel processor. They're good at performing computations with lots of independent operations and little control overhead. This is a perfect fit to neural nets!

Some important operations they speed up

- matrix multiplication
- convolution

Data and computing power

Graphics processing units (GPUs) are a kind of highly parallel processor. They're good at performing computations with lots of independent operations and little control overhead. This is a perfect fit to neural nets!

Some important operations they speed up

- matrix multiplication
- convolution

Software, from highest- to lowest-level

- Theano — you describe your model, and it computes derivatives for you
- GNumpy, which provides a NumPy-like interface
    - great for feed-forward nets, which mostly require matrix multiplication
- CUDAMat, a more low-level interface for linear algebra
- CUDA, NVIDIA's extension of C for GPU programming

Data and computing power

	LeNet (1989)	**LeNet (1998)**	**AlexNet (2012)**
classification task	digits	digits	objects

Data and computing power

	LeNet (1989)	LeNet (1998)	AlexNet (2012)
classification task	digits	digits	objects
categories	10	10	1,000

Data and computing power

	LeNet (1989)	**LeNet (1998)**	**AlexNet (2012)**
classification task	digits	digits	objects
categories	10	10	1,000
image size	16×16	28×28	$256 \times 256 \times 3$

Data and computing power

	LeNet (1989)	LeNet (1998)	AlexNet (2012)
classification task	digits	digits	objects
categories	10	10	1,000
image size	16×16	28×28	$256 \times 256 \times 3$
training examples	7,291	60,000	1.2 million

Data and computing power

	LeNet (1989)	**LeNet (1998)**	**AlexNet (2012)**
classification task	digits	digits	objects
categories	10	10	1,000
image size	16×16	28×28	$256 \times 256 \times 3$
training examples	7,291	60,000	1.2 million
units	1,256	8,084	658,000

Data and computing power

	LeNet (1989)	LeNet (1998)	AlexNet (2012)
classification task	digits	digits	objects
categories	10	10	1,000
image size	16×16	28×28	$256 \times 256 \times 3$
training examples	7,291	60,000	1.2 million
units	1,256	8,084	658,000
parameters	9,760	60,000	60 million

Data and computing power

	LeNet (1989)	**LeNet (1998)**	**AlexNet (2012)**
classification task	digits	digits	objects
categories	10	10	1,000
image size	16×16	28×28	$256 \times 256 \times 3$
training examples	7,291	60,000	1.2 million
units	1,256	8,084	658,000
parameters	9,760	60,000	60 million
connections	65,000	344,000	652 million

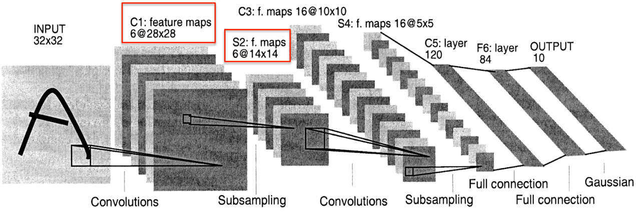
Data and computing power

	LeNet (1989)	**LeNet (1998)**	**AlexNet (2012)**
classification task	digits	digits	objects
categories	10	10	1,000
image size	16×16	28×28	$256 \times 256 \times 3$
training examples	7,291	60,000	1.2 million
units	1,256	8,084	658,000
parameters	9,760	60,000	60 million
connections	65,000	344,000	652 million
total operations	11 billion	412 billion	200 quadrillion (est.)

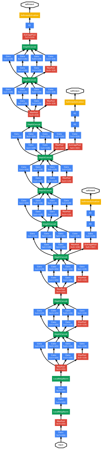
Data and computing power

More computing power allows us to fit deeper networks. E.g.,

- LeNet (1989) had 2 convolutional layers
- Google's *Inception* network (2014) had 22



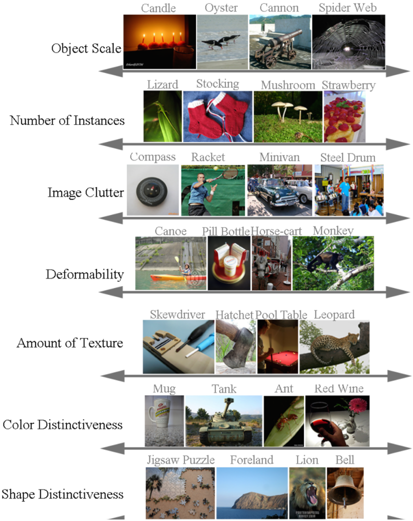
(from Geoff's lecture video)



(Szegedy et al., 2014, "Going deeper with convolutions")

ImageNet

This dataset is responsible for almost all of amazing progress made in applying neural nets for vision. Contains 1.28 million images belonging to 1000 different categories.



Russakovsky et al.

ImageNet

ILSVRC
flamingo
cock
ruffed grouse
quail
partridge
...
Egyptian cat
Persian cat
Siamese cat
tabby
lynx
...
dalmatian
keeshond
miniature schnauzer
standard schnauzer
giant schnauzer
...

Classification

Task : Given an image and a predefined set of categories, find out which category the image belongs to.
There is an annual competition ILSVRC (ImageNet Large Scale Visual Recognition Challenge).

Year	**Model**	**Best Result (Error %)**

Classification

Task : Given an image and a predefined set of categories, find out which category the image belongs to.
There is an annual competition ILSVRC (ImageNet Large Scale Visual Recognition Challenge).

Year	Model	Best Result (Error %)
2010	Hand-designed descriptors + SVM	28.2 %

Classification

Task : Given an image and a predefined set of categories, find out which category the image belongs to.
There is an annual competition ILSVRC (ImageNet Large Scale Visual Recognition Challenge).

Year	Model	Best Result (Error %)
2010	Hand-designed descriptors + SVM	28.2 %
2011	Compressed Fisher Vectors + SVM	25.8 %

Classification

Task : Given an image and a predefined set of categories, find out which category the image belongs to.
There is an annual competition ILSVRC (ImageNet Large Scale Visual Recognition Challenge).

Year	Model	Best Result (Error %)
2010	Hand-designed descriptors + SVM	28.2 %
2011	Compressed Fisher Vectors + SVM	25.8 %
2012	Deep Conv Net	16.4 %

Classification

Task : Given an image and a predefined set of categories, find out which category the image belongs to.

There is an annual competition ILSVRC (ImageNet Large Scale Visual Recognition Challenge).

Year	**Model**	**Best Result (Error %)**
2010	Hand-designed descriptors + SVM	28.2 %
2011	Compressed Fisher Vectors + SVM	25.8 %
2012	Deep Conv Net	16.4 %
2013	Deeper Conv Net	11.7 %

Classification

Task : Given an image and a predefined set of categories, find out which category the image belongs to.
There is an annual competition ILSVRC (ImageNet Large Scale Visual Recognition Challenge).

Year	**Model**	**Best Result (Error %)**
2010	Hand-designed descriptors + SVM	28.2 %
2011	Compressed Fisher Vectors + SVM	25.8 %
2012	Deep Conv Net	16.4 %
2013	Deeper Conv Net	11.7 %
2014	Even Deeper Conv Net	6.6 %
2015	?	?

Classification

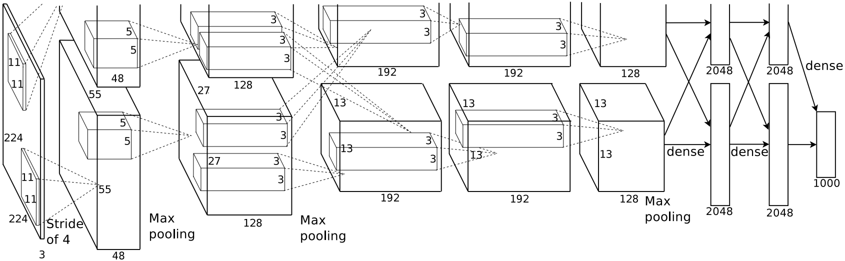
Task : Given an image and a predefined set of categories, find out which category the image belongs to.
There is an annual competition ILSVRC (ImageNet Large Scale Visual Recognition Challenge).

Year	**Model**	**Best Result (Error %)**
2010	Hand-designed descriptors + SVM	28.2 %
2011	Compressed Fisher Vectors + SVM	25.8 %
2012	Deep Conv Net	16.4 %
2013	Deeper Conv Net	11.7 %
2014	Even Deeper Conv Net	6.6 %
2015	?	?

There are already better results now (4.94%).
Human-performance is around 5.1%.

Classification

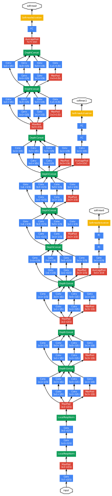
AlexNet, 2012. 8 weight layers. 16.4% Error.



(Krizhevsky et al., 2012)

Classification

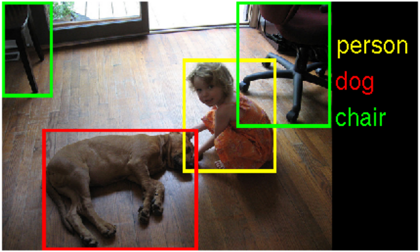
GoogLeNet, 2014. 22 weight layers. 6.6% Error.



(Szegedy et al., 2014)

Detection

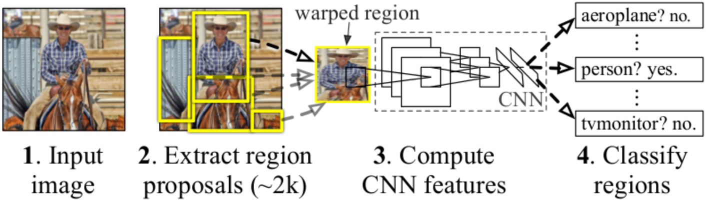
Task : Given an image and a predefined set of objects, find out which objects are present and draw a box around them.
Harder than classification.



For example, if the image has a lot of blue in it, we might classify it as *fish* without knowing anything about what fishes look like.

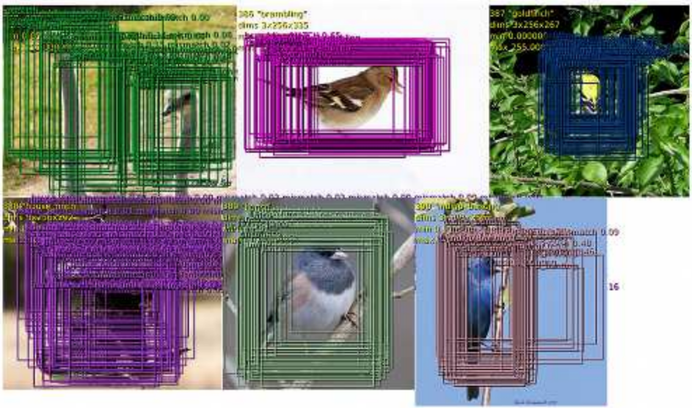
Detection

Region - CNN



Girshick et al. 2014

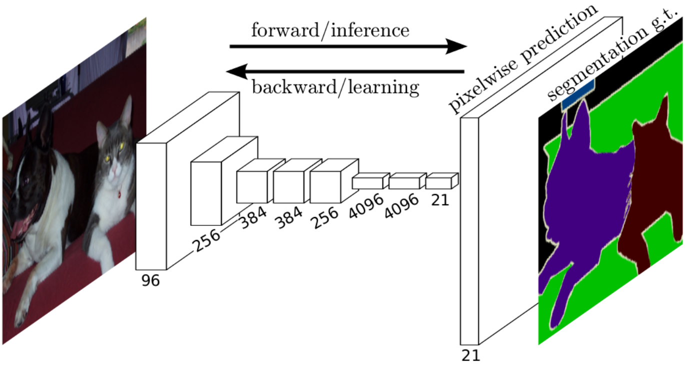
Overfeat - Regression to bounding box coordinates.



Sermanet et al. 2014

Segmentation

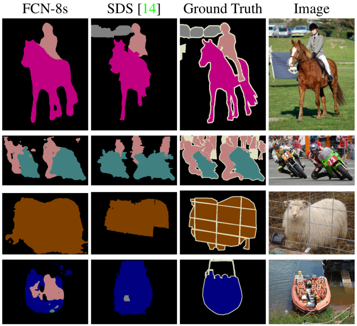
Task : Given an image and a predefined set of objects, find out which pixels belong to which objects.



Long et. al. 2014

Segmentation

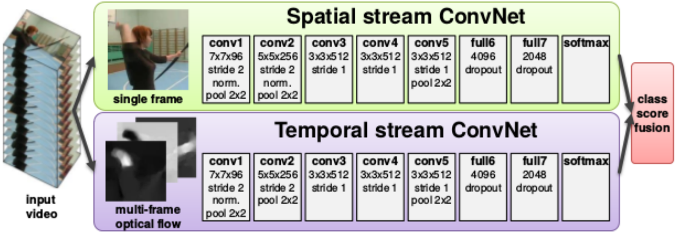
Task : Given an image and a predefined set of objects, find out which pixels belong to which objects.



Long et. al. 2014

Action Recognition

Task: Given a video and predefined set of actions, find out which action is being performed.



Simonyan et. al. 2014

3D convolutions

Instead of convolving in space (2D) convolve in space-time (3D).
Patches of images $\Rightarrow$ Cuboids of space-time.

Transfer

Transfer: Train a model on one dataset, apply to other datasets.

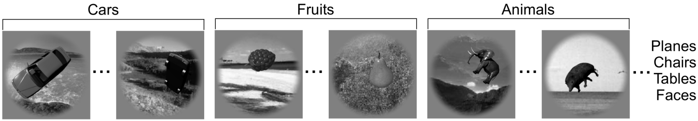
Features extracted from convolutional nets trained on ImageNet have been applied to

- Other Image Recognition / Detection Datasets.
- Many different video datasets.

A general image feature extractor.

Monkey vs Conv Net

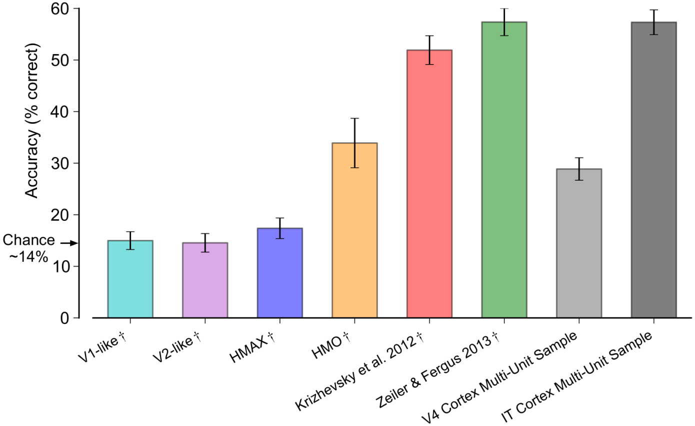
Compare Conv Net features with recordings from monkey brains for this simple task.



Cadieu et. al. 2014

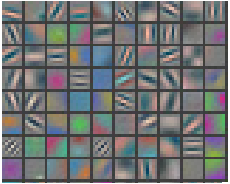
Monkey vs Conv Net

Compare Conv Net features with recordings from monkey brains.



Visualizing the representations

Here are the first-layer filters learned by a state-of-the-art object recognition network from 2013:

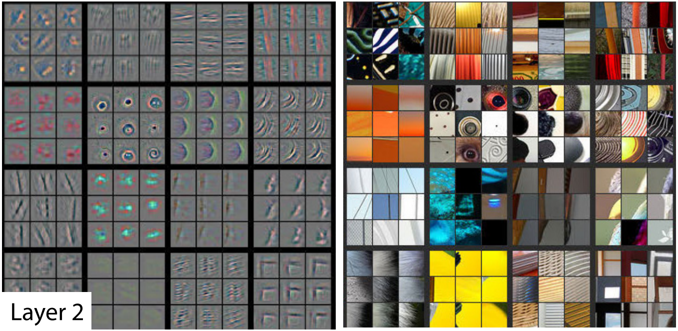


(Zeiler and Fergus, 2013., Visualizing and understanding convolutional networks)

Visualizing the higher-layer filters is much tougher.

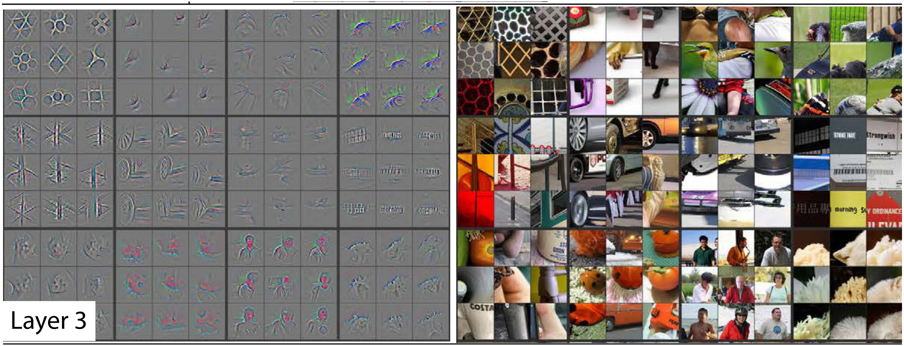
Visualizing the representations

Zeiler and Fergus (2013) came up with a scheme for visualizing the learned representation. For each unit, they picked the 9 largest activations over the whole dataset. They have a scheme for visualizing the responses which we won't talk about.



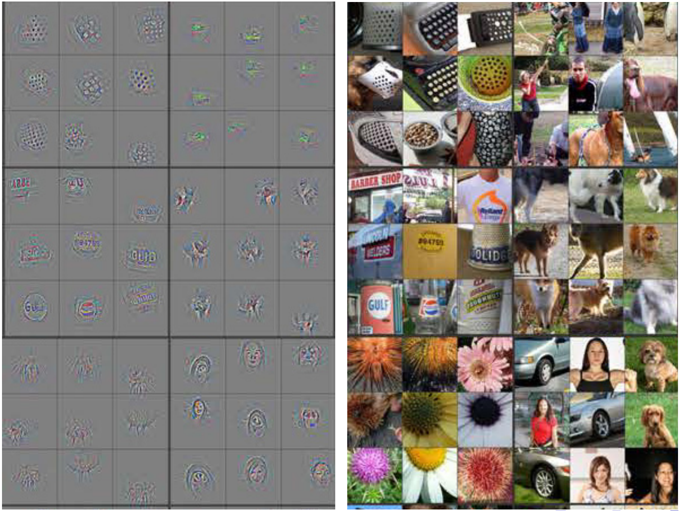
Visualizing the representations

Here's layer 3. The units have larger receptive fields. (Why?)



Visualizing the representations

And layer 5. The units respond to high-level semantic properties.



Visualizing the representations

Can we conclude from this that a unit "represents" faces, text, dogs, etc.?

Visualizing the representations

Can we conclude from this that a unit "represents" faces, text, dogs, etc.?

Szegedy et al. (2013) found that the visualization looks just as selective if you pick random linear combinations of units!

(a) Unit sensitive to white flowers.

(b) Unit sensitive to postures.

(c) Unit senstive to round, spiky flowers.

(d) Unit senstive to round green or yellow objects.

(a) Direction sensitive to white, spread flowers.

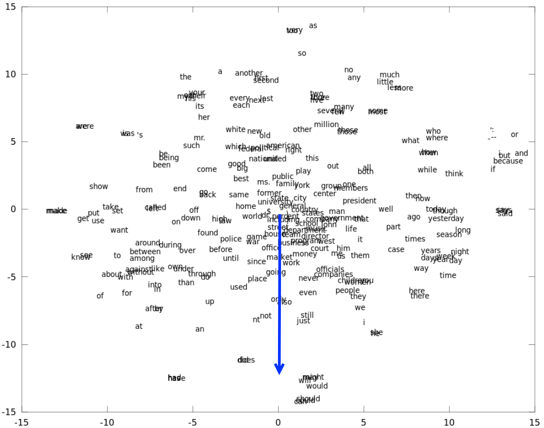
(b) Direction sensitive to white dogs.

(c) Direction sensitive to spread shapes.

(d) Direction sensitive to dogs with brown heads.

Visualizing the representations

By analogy: does the *y*-axis “represent” might, would, should, etc. in the tSNE visualization from Assignment 1?



Adversarial images

In Week 3, we worked through a backprop example. We computed the derivatives with respect to the inputs, even though we never needed them to update the parameters.

Here's something really cool you can do with those derivatives.

Adversarial images

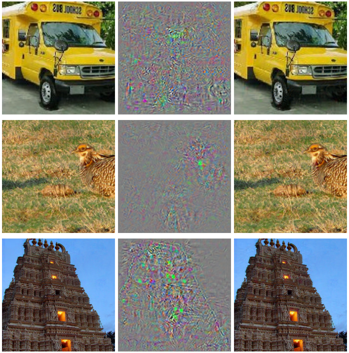
In Week 3, we worked through a backprop example. We computed the derivatives with respect to the inputs, even though we never needed them to update the parameters.

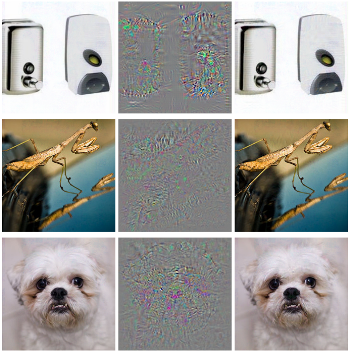
Here's something really cool you can do with those derivatives.

Take a conv net that correctly classifies an image. Do gradient ascent on the image to maximize the probability that it's classified as some unrelated category (e.g. "ostrich"). What do you think will happen?

Adversarial images

Left: original image (which was classified correctly)
Right: adversarial image (which the network things is an ostrich)
Center: difference (adversarial – original), multiplied by 128





Midterm exam

Tuesday, Feb. 24, during class

50 minutes

What you're responsible for:

- Coursera videos up through G (except ones marked optional)
- In-class lectures up through this lecture (especially the problems)
- Assignment 1

The hardest questions will be about things we covered both in the videos *and* in class.

We will not ask for formal proofs, only informal justifications.

There will be less time pressure than in the in-class exercises. We'll focus on conceptual questions, rather than long derivations.

Practice exams and extra office hours TBA.