DEEP LEARNING FOR COMPUTER VISION

Summer Seminar UPC TelecomBCN, 4 - 8 July 2016

Instructors

Xavier Giró-i-Nieto

Elisa Sayrol

Amaia Salvador

Jordi Torres

Eva Mohedano

Kevin McGuinness

Organizers

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

telecom BCN

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

DCU Dublin City University
Ollscoil Chathair Bhaile Átha Cliath

Insight Centre for Data Analytics

NVIDIA GPU CENTER OF EXCELLENCE

Co-funded by the Erasmus+ Programme of the European Union

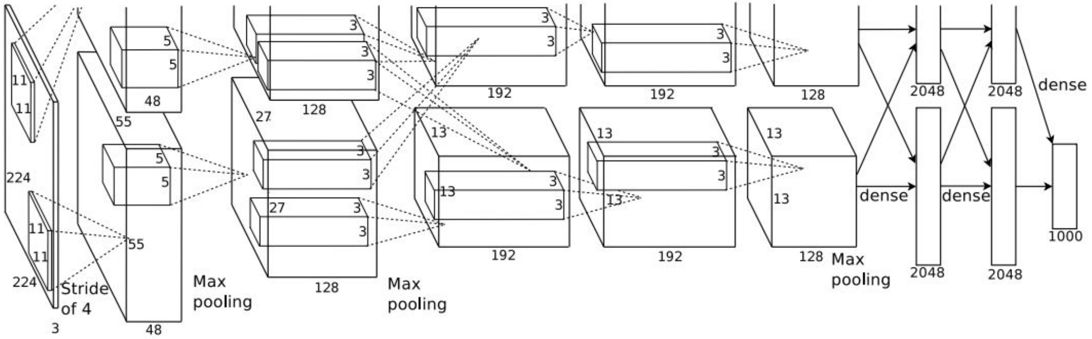
+ info: **TelecomBCN.DeepLearning.Barcelona**

Day 2 Lecture 2

Augmentation

Introduction

[ImageNet Classification with Deep Convolutional Neural Networks](), Krizhevsky A., 2012



ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) **1.2 million training images**, 50,000 validation images, and 150,000 testing images

Architecture of 5 convolutional + 3 fully connected = **60 million parameters ~ 650.000 neurons**.

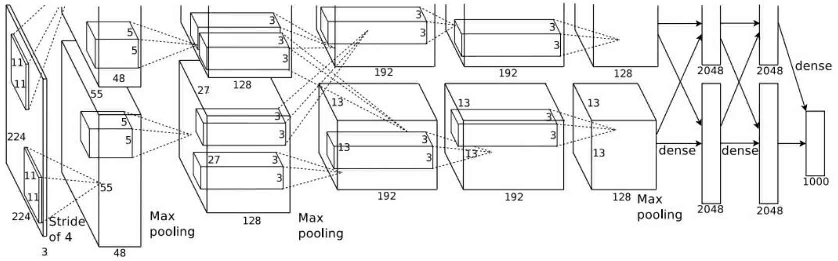
Overfitting!!

Ways to reduce overfitting

- **Reduce network capacity**

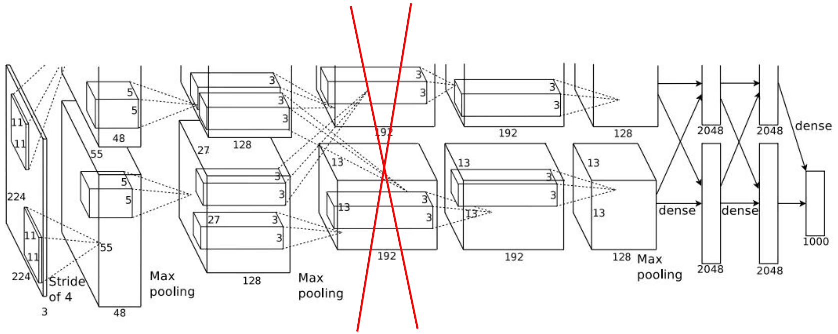
- Dropout

- Data augmentation



Ways to reduce overfitting

- **Reduce network capacity**



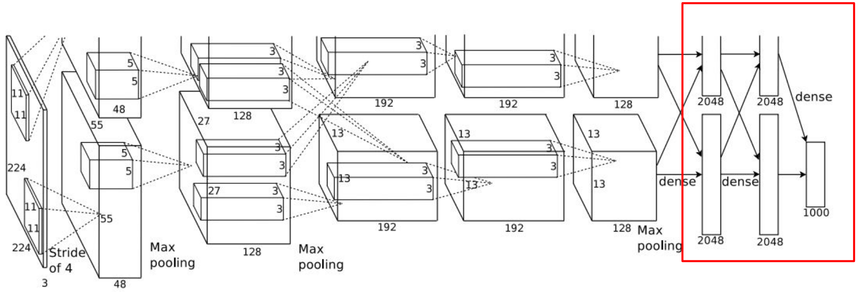
- Dropout

1% of total parameters (884K). Decrease in performance

- Data augmentation

Ways to reduce overfitting

- **Reduce network capacity**
- Dropout
- Data augmentation



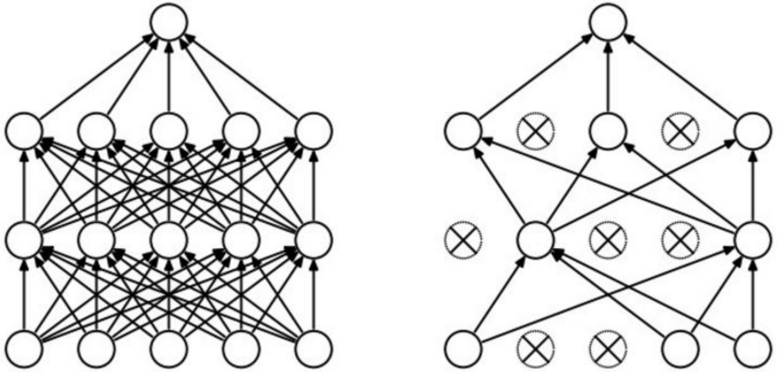
37M, 16M, 4M parametes!! (fc6,fc7,fc8)

Ways to reduce overfitting

- Reduce network capacity

- **Dropout**

- Data augmentation



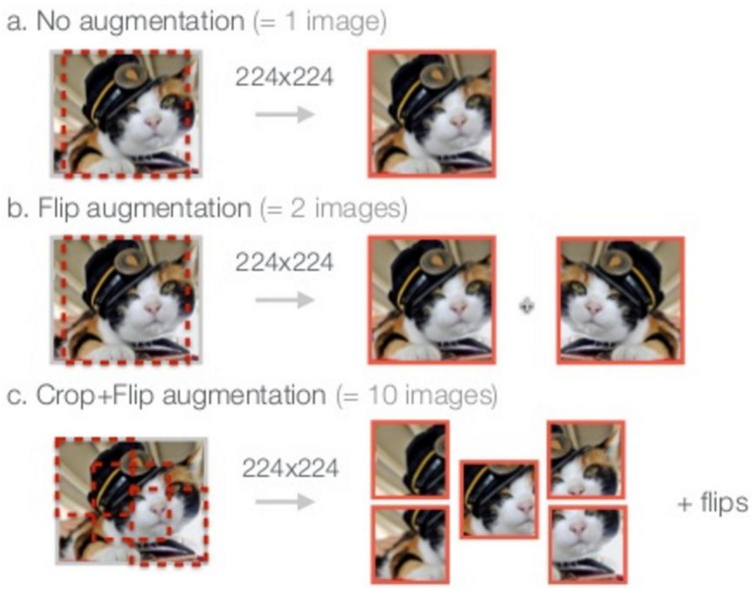
(a) Standard Neural Net (b) After applying dropout.

Every forward pass, network slightly different.
Reduce co-adaptation between neurons
More robust features

More interations for convergence

Ways to reduce overfitting

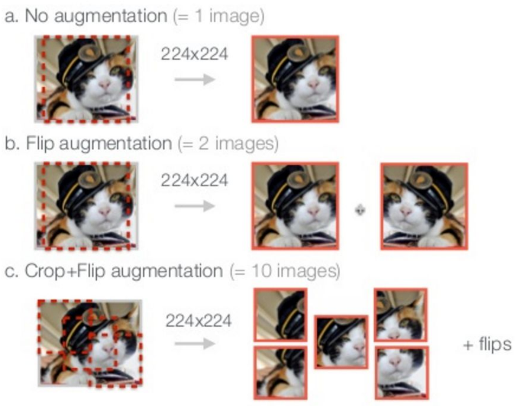
- Reduce network capacity
- Dropout
- **Data augmentation**



Data Augmentation

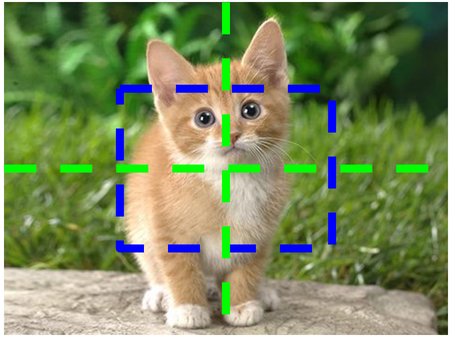
During training, alterate the input image (Krizhevsky A., 2012)

- Random crops on the original image
- Translations
- Horitzontal reflections
- Increases size of training x2048
- On-the-fly augmentation



During testing

- Average prediction of image augmented by the four corner patches and the center patch + flipped image. (10 augmentations of the image)



Data Augmentation

Alternate intensities RGB channels intensities

PCA on the set of RGB pixel throughout the ImageNet training set.
To each training image, add multiples of the found principal components

$$I_{xy} = [I_{xy}^R,\ I_{xy}^G,\ I_{xy}^B]^T$$

$$[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3][\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^T$$

$$\alpha_i \sim N(0,\ 0.1)$$

Object identity should be invariant to changes of illumination

Augmentation for discriminative unsupervised feature learning

Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks, Dosovitskiy, A., 2014

MOTIVATION

- Large datasets of training data
- Local descriptors should be invariant transformations (rotation, translation, scale, etc)

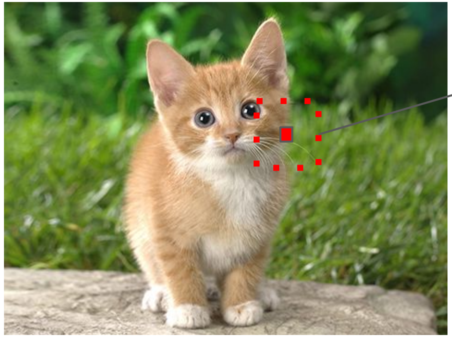
WHAT THEY DO

- Training a CNN to generate local representation by optimising a surrogate classification task
- Task does NOT require labeled data

Augmentation for discriminative unsupervised feature learning

Generate augmented dataset: 16000 classes of 150 examples each

Class k=1, with 150 examples



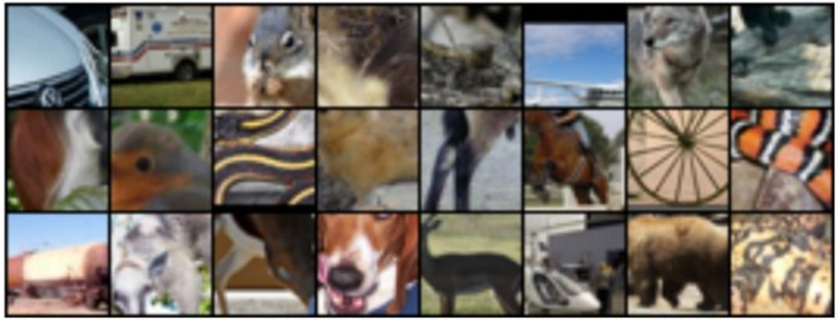
...

Select *random* location k and crop 32x32 window (restrictions: region must contain objects or part of the object: high amount of gradients)

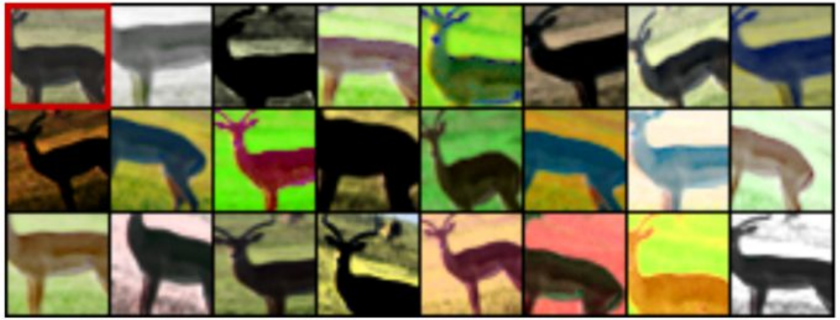
Apply a transformation [translation, rotation, scalig, RGB modification, contrast modification]

Augmentation for discriminative unsupervised feature learning

Generate augmented dataset: 16000 classes of 150 examples each



Example of classes



Example of examples for one class

Augmentation for discriminative unsupervised feature learning

Classification accuracies

Algorithm	STL-10	CIFAR-10(400)	CIFAR-10	Caltech-101	Caltech-256(30)	#features
Convolutional K-means Network [33]	60.1 ± 1	70.7 ± 0.7	82.0	—	—	8000
Multi-way local pooling [34]	—	—	—	77.3 ± 0.6	41.7	1024×64
Slowness on videos [14]	61.0	—	—	74.6	—	556
Hierarchical Matching Pursuit (HMP) [35]	64.5 ± 1	—	—	—	—	1000
Multipath HMP [36]	—	—	—	82.5 ± 0.5	50.7	5000
View-Invariant K-means [16]	63.7	72.6 ± 0.7	81.9	—	—	6400
Ex-CNN Small (64c5-64c5-128f)	67.1 ± 0.2	69.7 ± 0.3	76.5	$79.8 \pm 0.5^*$	42.4 ± 0.3	256
Ex-CNN Medium (64c5-128c5-256c5-512f)	72.8 ± 0.4	75.4 ± 0.2	82.2	$86.1 \pm 0.5^\dagger$	51.2 ± 0.2	960
Ex-CNN Large (92c5-256c5-512c5-1024f)	$\mathbf{74.2 \pm 0.4}$	$\mathbf{76.6 \pm 0.2}$	**84.3**	$\mathbf{87.1 \pm 0.7}^\ddagger$	$\mathbf{53.6 \pm 0.2}$	1884
Supervised state of the art	70.1[37]	—	92.0 [38]	91.44 [39]	70.6 [2]	—

Superior performance to SIFT for image matching.

Summary

Augmentation helps to prevent overfitting

It makes network invariant to certain transformations: translations, flip, etc

Can be done on-the-fly

Can be used to learn image representations when no label datasets are available.