references

Hack Chyson

August 7, 2019

Contents

1 image retrieval using bim and features from pre-trained vgg network for indoor localization

1.1 network

VGG16 and VGG19

ImageNet pretrained network

1.2 two experimental:

1. in corridor

2. in hall

1.3 study content

view overlap
which layer is best

1.4 result

1. version-based method is more efficient

2. the fourth layer feature map is best

3. ImageNet network can extract generic features

1.5 problem

1. large object effect like a picture frame or a poster

2. structure similarity (multiple pictures)

2 very deep convolutional networks for large-scale image recognition

VGG: Visual Geometry Group

2.1 main work

investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting.

VGG16-19 winned that ImageNet Challenge 2014 first in localisation and classification tracks.

VGG representations generalise well to other datasets.

2.2 introduction

ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)

attempts to improve accuracy:

1. utilise smaller receptive window size and samller stride of the first cnvolutional layer (Zeiler & Fergus, 2013; Sermanet et al., 2014)

2. train and test the networks densely over the whole image and over multiple scales (Sermanet et al., 2014; Howard, 2014)

3. increase the depth of the network by adding more convolutional layers (this paper)

2.3 convnet configuration

2.3.1 architecture

input	224 x 224 RGB image
preprocessing	substract the mean RGB value, from each pixel (why?)
conv. filter	kernel=3 x 3 or 1 x 1; stride=1
spatial pooling	max-pooling, 2 x 2, stride=2
activation	rectification(ReLU)

filters with a very small recptive field: 3 x 3 (the smallest size to capture the notion of left/right, up/down, center)

1 x 1 convolution filter: can be seen as a linear transformation of the input channels (followed by non-linearity)

Spatial padding is such that the spatial resolution is preserved after convolution. (i.e. the padding is 1 pixel for 3 x 3 conv. layers) (why?)

LRN: local response normalization (why? it des not improve the performance on the ILSVRC dataset)

2.3.2 configurations

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 **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 **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 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

In spite of a large depth, the number of weights in our nets is not greater than the number of weights
in a more shallow net with larger conv. layer widths and receptive fields. (That's one reason of using small kernel)

benefits of using three 3 x 3 conv. layer instead of a single 7 x 7 layer:

1. incorporate three non-linear rectification layers instead of a single one ==> make the decision function more discriminative

2. less parameters ==> imposing a regularisation on the 7 x 7 conv. filters, forcing them to have decomposition through the 3 x 3 filters

The incorporation of 1×1 conv. layers is a way to increase the non-linearity
of the decision function without affecting the receptive fields of the conv. layers.

2.4 classification framework

2.4.1 training

The training is carried out by optimising the multinomial logistic regression objective
using mini-batch gradient descent (based on back-propagation (LeCun et al., 1989)) with momentum.

batch size	256
momentum	0.9

FC layers
first layer regularisation: weight decay (the L_2 penalty multiplier set to $5 \cdot 10^{-4}$)
second layer regularisation: dropout (dropout ratio set to 0.5)

The learning rate was initially set to 10^{-2} , and then decreased by a factor of 10 when the validation set accuracy stopped improving.

The net converged after 74 epoches because of: (less epoches)

1. implicit regularisation imposed b greater depth and smaller convolution filter sizes

2. pre-initialisation of certain layers

The initialisation of the network weights is important, since bad initialisation can stall learning
due to the instability of gradient in deep nets. ("Understanding the difficulty of training deep feedforward neural networks")

The author of this paper used pre-training emthod tu circumvent the initialisation of the network.
Although, the pre-training method is unnecessary for the initialisation, how is it done?

training set:

1. cropped from rescaled training images (one crop per image per SGD iteration)

2. random horizontal flipping

3. random RGB colour shift

Training image size:
equal to or greater than 224 x 224.
If the image is greater than 224 x 224, a crop will be done.

isotropically-rescaled *ai sou 'tro pi kerli*

scale jittering (one method of training set augmentation):
Each training image is individually rescaled by randomly sampling S from a certain range $[S_{min}, S_{max}]$.
Crop S size input from sampled images.

2.4.2 testing

1. rescale to a pre-defined smallest image side

2. network applied densely over the rescaled image

3. to obtain a fixed-size vector of class scores for the image, the class score is spatially averaged (sum-pooled)

multi-crop evaluation vs dense evalution:

2.5 classification experiments

Dataset:

training	1.3M images
validation	50K images
testing	100K images

classification performance:

top-1 error	the portion of incorrectly classified images
top-5 error	the portion of images such that the ground-truth category is outside the top-5 predicted

2.5.1 single scale evaluation

A deep net with small filters outperforms a shallow net with larger filters. Training set augmentation by scale jittering is indeed helpful for capturing multi-scale image statistics

What is convolution boundary condition?

Emsembling improves the performance duo to complementarity of the models.

2.6 Localisation

bounding box prediction:
SCR: single-class regression?
PCR: per-class regression?

logistic regression –> Euclidean loss

To come up with the final prediction:
greedy merging procedure

1. merge spatially close predictions (by averaging their coordinates)

2. rates them on the class scores

localisation error in ILSVRC criterion:
$IoU = \frac{P \cap G}{P \cup G} < 0.5$

Conclution: The preformance in localisation can be improved with very deep convolution nets.

2.7 generalisation of very deep features

ConvNets, pre-trained on ILSVRC, generalise well on other, smaller, datasets,
where training large models from scratch is not feasible due to over-fitting.

How:
remove the lst fully-connected layer and use 4096-D activation of the penultimate layer as image features

aggregation of feature:

1. an image is rescaled

2. the network is densely applied

3. perform global average pooling on the resulting feature map (produces a 4096-D image descriptor)

4. the descriptor is averaged with the descriptor of a horizontally flipped image

5. extract feature over several scales

6. the resulting multi-scale features can be either stacked or pooled across scales

20% of training images were used as a validation set for hyper-parameter selection.
hyper-parameter: set by human being. (learning rate, tree depth)
parameter: learned from a algorithm. (matrix weight of CNN)

If the dataset contains multi-scale image, stacking and pooling of feature are almost same.
Otherwise, stacking allows a classifier to expolit scale-specific representations, and behaves better.

3 visualizing and understanding convolutional networks

3.1 intruduction

Without clear understanding of how and why they work, the development of better models
is reduced to trial-and-error.

To study the CNN:

1. visualizing with multi-layer deconvolutinal network

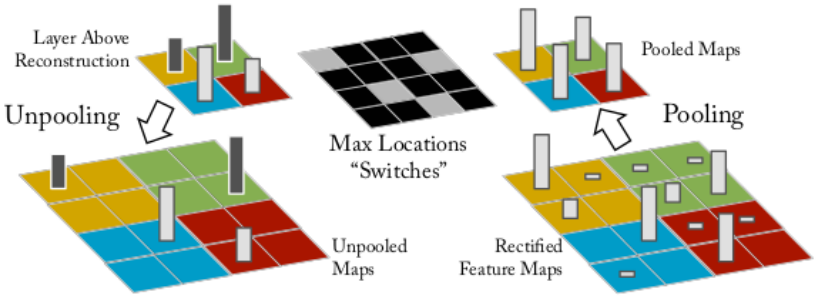
2. sensitivity analysis of the classifier output by occluding portions of the input images.

3.2 Approach

deconvnet: map features to pixels

switches: record the location of the local max in each pooling region.

In convnet, the max pooling operation is non-invertible, however we can obtain an approximate
inverse with **swithes**



As these switch settings are peculiar to a given input image, the reconstruction obtained from
a single activation thus resembles a small piece of the original input image, with structures
weighted according to their contribution toward to the feature activation.

deconvent:

1. unpooling with switches

2. rectification with relu (same with convnet)

3. filtering with transposed filter

3.3 Training Details

preprocess:

1. resize the smallest dimension to 256

2. crop the center 256*256 region

3. substracting the per-pixel mean

4. use 10 different sub-corps of size 224*224

optimization:
SGD with mini-batch (128)

Renormalize each filter whose RMS(root mean square) value exceeds a fixed radius of 10^{-1} to this fixed radius
to avoid a few of filters dominate.

3.4 Convnet Visualization

3.4.1 Feature visualization

1. strong grouping within each feature map

2. greater invariance at higher layers

3. exaggeration of discriminative parts of the image

3.4.2 Feature Evolution during Training

The latter layer need more epoches to converge.

3.4.3 Feature Invariance

For max pooling, the network output is stable to translations and scaling. In general, the output is not invariant to rotation.

3.4.4 Occlusion Sensitivity

method: occlude different parts of the image.

The model is localizing the objects as the probability of the correct class drops significantly
when the object is occluded.

This shows that the visualization genuinely corresponds to the image structure
that stimulates that feature map.

3.4.5 Correspondence Analysis

method: masking out specified parts and random parts of a image.

At layer 5, the model does establish some degree of correspondence
by comparing Mean Feature Sign Change with masking out left eye, right eye, nose and random region.

4 BIM Tracker: A model based visual tracking approach for indoor localisation using a 3D building model

localization with edges search and matching.

5 BIM-PoseNet: Indoor camera localisation using a 3D indoor model and deep

result: indoor localization in real-time with an accuracy of approximately 2 meters.

5.1 Introduction

objective: investigate whether pose estimation can be done by fine-tuning a pre-trained
network using synthetic images derived from a 3D indoor model rather than geotagged images

5.2 Background and related work

The visual localization approaches in the literature can be classified as:

appearance-based image retrieval problem

pose eistimation-base directly estimate the 6-DOF pose of a

matching point features with 3D point clouds requirement of point clouds (usually derived from SfM)

pose regression using RGB-D images fast and precise, but need RGB-D camera

pose regression using images only

RGB-D: RGB + depth (distance between pixel and the sensors)

They fine-tuned a pre-trained network on image samples with ground-truth poses derived from the SfM methods.

They state: deep convolutional neural network trained for the task of classification preserve pose information
till the final layer by leveraging transfer learning,
despite being trained for a different task with a different dataset.

drawback (using ground-truth images): dependent on SfMmethods to estimate the ground truch camera poses,
required during fine-tuning the network.

synthetic images generated from 3D object models is used to eliminate the challenge of creating
manually labelled images (ground truth images)

5.3 Methodology

current CNN network -> based on -> PoseNet(Kendall,2015) -> based on -> GoogLeNet(Szegedy,2015)

Thw weights are updated from a pre-trained network of GooLeNet that is trained on Places dataset(Zhou,2014)

$$p = [x, q] \tag{1}$$

x is a vector representing;
q is a orientation representing;

$$\text{loss}(I) = ||\hat{x} - x||_2 + \beta \left|\left| \hat{q} - \frac{q}{||q||} \right|\right|_2 \tag{2}$$

β is a hyperparameter balancing the error of location and orientation.

The author of (Peng,2015) show that features derived from DCNNs are invariant to color, texture, pose and context.

In other words, if a network is invariant to an object's texture,
then it will have similar activations of neurons for the object with or without texture.
The network hallucinates the right texture when given a texutre-less object's shape.
(? don't understand)

Whether different model renderings and processing the real images to make them similar
to the synthetic images will increase the pose estimation accuracy.

To test this, we transform the synthetic and real images in a common feature space of edge gradient magnitude (gradmag) images.
(Converting images to edge gradmag comes at the cost of loss of information such as
colour and texture, but on the other hand, the main geometrical features of the images are preserved.)

5.4 experiments and result

1. experiment 1: creating a baseline accuracy using real images

2. experiment 2: fine-tuning with synthetic image dataset and test

3. experiment 3: explore accuracy with detail

	Caffe library on Linux
loss optimization	Adagrad gradient descent optimization algorithm
learing rate	10^{-3}.
	NVIDIA GTX980M
batch size	40
resize to resolution	320*240
crop	224*224

5.4.1 dataset

1. synthetic image dataset The BIM contains the main building elements including walls, floors, ceilings, doors, ceiling tube-lights,
   and stairs, but not details such as material, fabrication, assembly and installation information

   The height of the trajectory was kept in the range of $1.5 - 1.8$ meters from the floor.

   we have rendered images along the trajectory at 0.05 meters interval and ± 10 tilt.

2. real image dataset A total number of 1000 images of 640x480 pixels resolution were acquired at a constant 30 frames per second.

5.4.2 baseline performance using real images

The value of β lies in the range of 120 to 750. (beta seleted)

Figure 8: The saliency map predicted by a network fine-tuned with real images, overlaid on the test image. The red colour indicated the pixels in the image that the network considers important for pose regression.

5.4.3 fine-tuning with synthetic images

The author showed that: different parts of a image make a difference in importance.

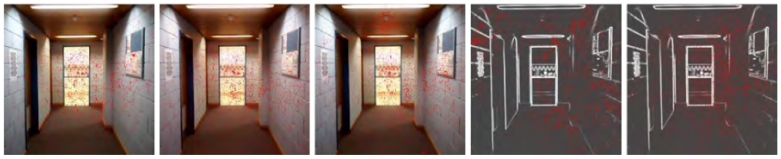


Figure 11: The saliency maps of (a) - (c) a real image and (d) - (e) edge *gradmag* of real image predicted by networks fine-tuned with synthetic image datasets. The red colour shows the pixels in the image the network considers important for pose regression.

elements to case error:

1. photo blur

2. external elements (like poster)

3. structral difference

Lighting of the scene plays a vital role in the appearance of the scene.

The high errors might be a result of the learnt features for each network, which might not be suitable for pose regression with real images.
This fact is reflected in the saliency maps of the real images as predicted by the fine-tuned networks.

5.5 effects of level-of-detail of 3D models

6 Learning to Compare Image Patches via Convolutional Neural Networks

show how to learn directly from image data a general similarity function for comparing image patches.

Requirement: large datasets that contain patch correspondences between images.
This is not suitable from indoor localization, becuase this is no such large dataset of camera pictures.

7 Structure Extraction from Texture via Relative Total Variation

a picture = meaningful structures + textured surfaces (commonly)

inherent variation and relative total variation to distinguish them

In psychology:
the overall structural features are the primary data
of human perception, not the individual details

8 Cross-Domain 3D Model Retrieval via Visual Domain Adaptation

9 Cross-domain Image Retrieval with a Dual Attribute-aware Ranking Network

DARN: Dual Attribute-aware Randing Network

retrieval feature learing.

two sub-networks, whose retrieval feature representations are driven by semantic attribute learning.

attribute-guided learning is a key factor for retrieval accuracy improvement.

9.1 Related Work

1. Fashion Dataset

2. Visual Analysis of Clothing with Fashion Datasetn

3. Visual Attibutes

4. Deep Learning (explicitly use attribute prediction as a regularizer in deep network)

Attributes are usually referred as semantic properties of objects or scenes that are shared across categories.

Richer supervision conveying annotator /'an nou tei ter/ rationales based on visual attributes, can be considered as a form of privileged information. Cross-domain image retrieval can benefit from feature learning that simultaneously optimizes a loss function that takes into account visual similarity and attribute classification.

A poselet describes a particular part of the human pose under a given viewpoint.

9.2 Data Collection