Weekly Report on Road analytics

Faculty: Mehul S. Raval

Date : 24/2/22 – 2/3/22

Member: Yagnik Bhavsar

En. No: AU2149006

Outline of peformed task :

- Literature survey
- Samples distribution over resolution

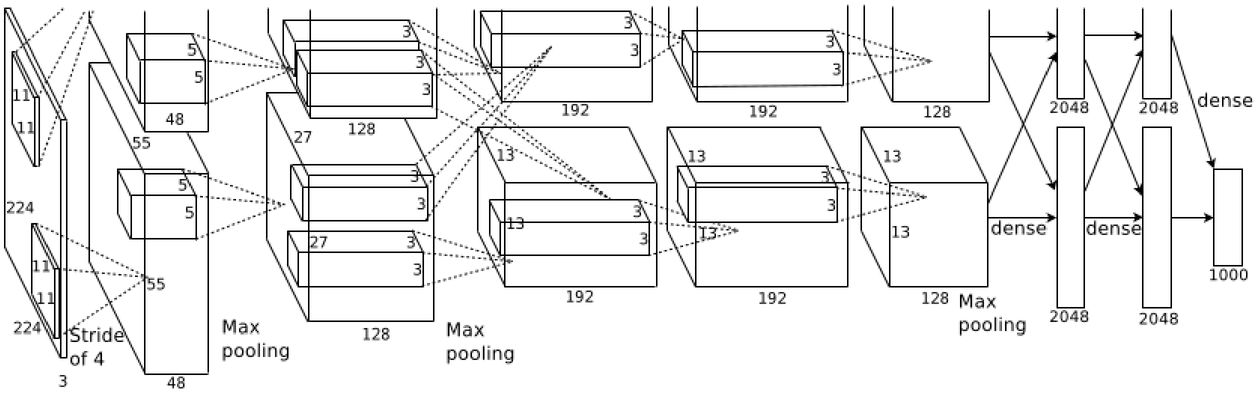
Literature Survey:

[5] ImageNet Classification with Deep Convolutional Neural Networks

Developed a CNN model (Alexnet) with following specifications,

- ImageNet dataset
  - 1.2 million high-resolution training images
  - 50,000 validation images
  - 150,000 testing images
  - 1000 different classes
- Fixed size input images of 224 x 224
- 60 million model parameters
- 650,000 neurons
- five convolutional layers, some of which are followed by max-pooling layers
- three fully-connected layers with a final 1000-way softmax
- “dropout” regularization method
- ReLU activation function
- Training on multiple GPUs
- stochastic gradient descent with a batch size of 128 examples, momentum of 0.9, and weight decay of 0.0005

Architecture :



1st CONV layer :

- 224 × 224 × 3 input image
- 96 kernels of size 11 × 11 × 3 with a stride of 4 pixels

2nd CONV layer :

- Response-normalized and pooled output of the first convolutional layer as input
- 256 kernels of size 5 × 5 × 48

3rd CONV layer :

- Response-normalized and pooled output of the second convolutional layer as input
- 384 kernels of size 3 × 3 ×256

4th CONV layer :

- output of the third convolutional layer as input
- 384 kernels of size 3 × 3 × 192

5th CONV layer :

- output of the fourth convolutional layer as input
- 256 kernels of size 3 × 3 × 192

Last two fully-connected layers have 4096 neurons each and output layer has 100 neurons.

Overlapping Pooling

Pooling layers in CNNs summarize the outputs of neighboring groups of neurons in the same kernel map. Traditionally, the neighborhoods summarized by adjacent pooling units do not overlap,

- neighborhood of size z × z centered at the location of the pooling unit
- grid of pooling units spaced s pixels apart

Here, if s = z leads traditional local pooling but if s < z, then it's overlapping pooling. It is observed that during training that models with overlapping pooling find it slightly more difficult to overfit.

Local Response Normalization

It is found that the following local normalization scheme aids generalization. Denoting by $a^i_{x,y}$ the activity of a neuron computed by applying kernel i at position (x, y) and then applying the ReLU nonlinearity, the response-normalized activity $b^i_{x,y}$ is given by the expression,

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^{\beta}$$

where the sum runs over n "adjacent" kernel maps at the same spatial position, and N is the total number of kernels in the layer. The constants k, n, α, and β are hyper-parameters whose values are determined using a validation set;

- k = 2, n = 5, α = 10 −4 , and β = 0.75.

This normalization is applied after applying the ReLU nonlinearity in certain layers.

Reducing Overfitting

- *Data Augmentation*
  - artificially enlarge the dataset using label-preserving transformations
  - altering the intensities of the RGB channels in training images
- *Dropout*
  - setting to zero the output of each hidden neuron with probability 0.5

[6] Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Existing deep convolutional neural networks (CNNs) require a fixed-size (e.g., 224×224) input image. This requirement is “artificial” and may reduce the recognition accuracy for the images or sub-images of an arbitrary size/scale.“spatial pyramid pooling”, is introduced in existing CNN architecture to eliminate the above requirement.

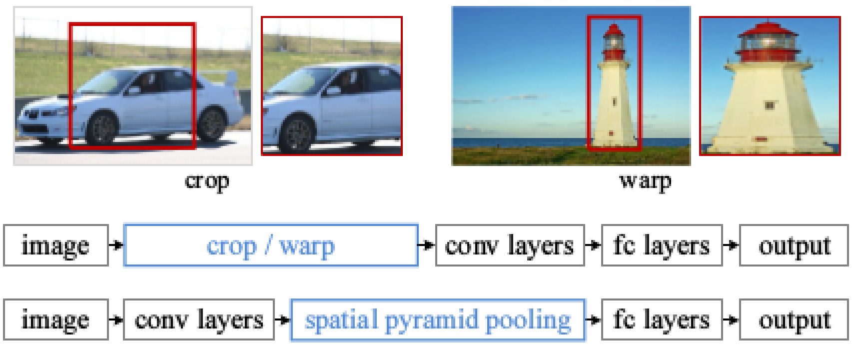


Figure 1: Top: cropping or warping to fit a fixed size. Middle: a conventional CNN. Bottom: our spatial pyramid pooling network structure.

To remove the fixed-size constraint of the network, add a SPP layer on top of the last convolutional layer. SPP is able to generate a fixed-length output regardless of the input size.

SPP is tested on popular seven-layer architectures- 5 CONV layes with 2 FC layers.

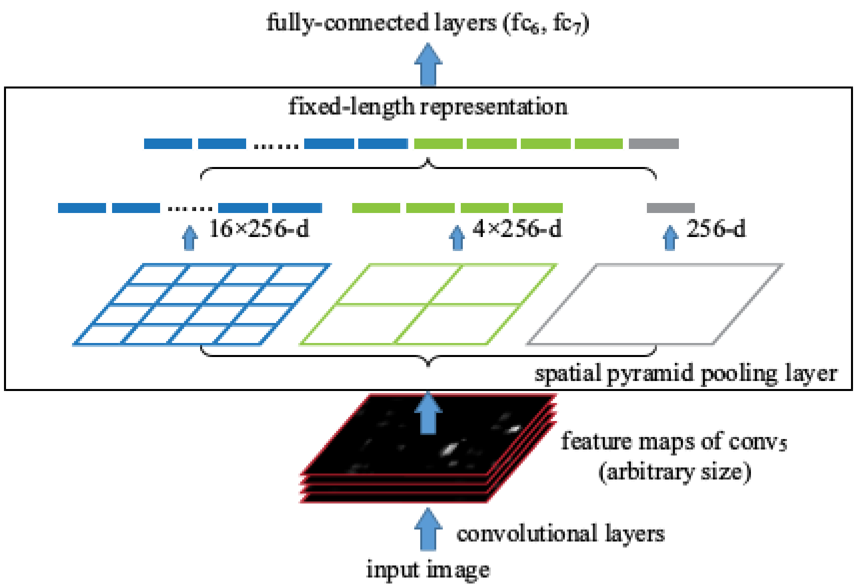


Fig. SPP layer architecture

Replace the last pooling layer (e.g., pool 5 , after the last convolutional layer) with a spatial pyramid pooling layer. In each spatial bin, pool the responses of each filter (max pooling). The outputs of the spatial pyramid pooling are kM - dimensional vectors with the number of bins denoted as M (k is the number of filters in the last convolutional layer). The fixed-dimensional vectors are the input to the fully connected layer.

Consider the feature maps after $conv_5$ that have a size of a×a With a pyramid level of n×n bins, we implement this pooling level as a sliding

- window size win = ceil (a/n)
- stride str = floor(a/n)
- l-level pyramid will have l such layers.

The next fully-connected layer (fc_6) will concatenate the l outputs.

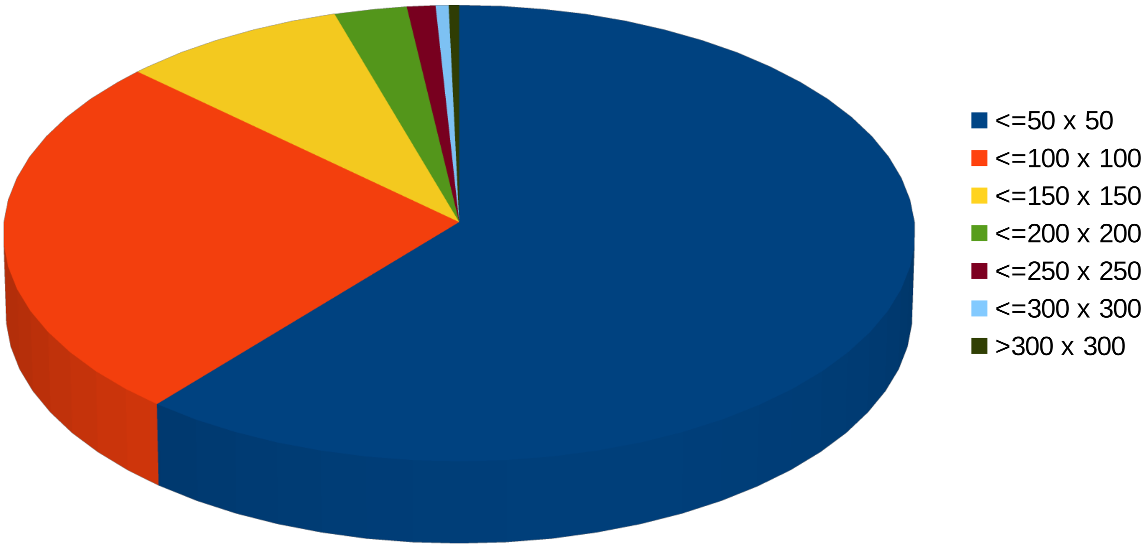
Samples distribution over resolution

We are using VisDrone dataset for model training. Sometimes it is important to have idea about the training input images resolution because some models only allows specific sized training images. Also, smaller resolution images may disturb object detection accuracy of model.

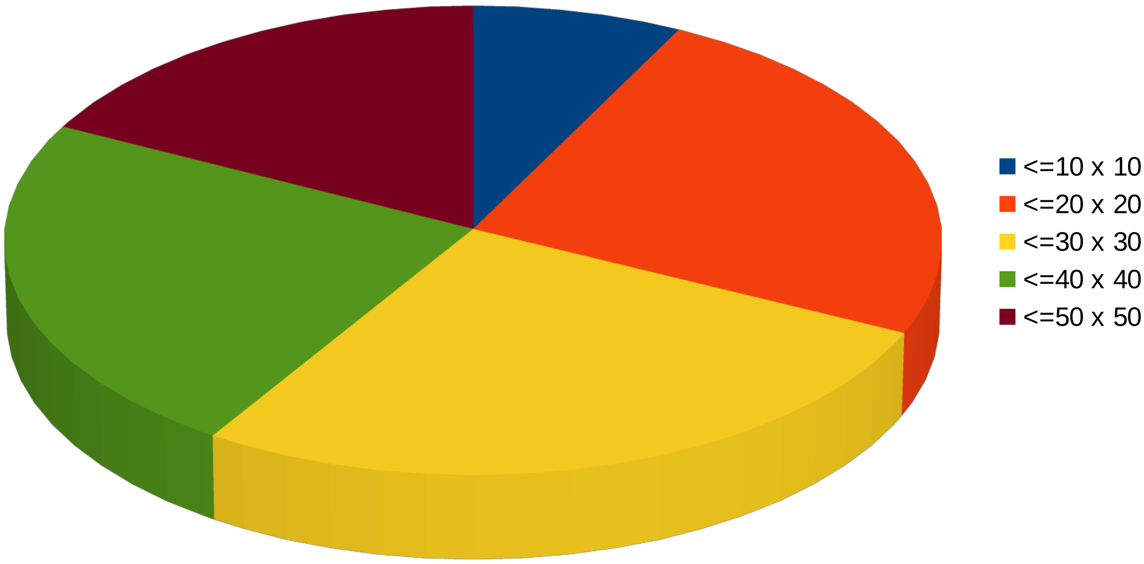
Class	Sample Size	<= 50 x50	<= 100 x 100	<= 150 x 150	<= 200 x 200	<= 250 x 250	<= 300 x 300	> 300 x 300
bicycle	10480	8378	1819	225	44	12	1	1
bus	5926	2277	2104	820	386	156	72	111
car	144867	84578	40225	13301	4079	1582	689	412
motor	29647	25281	3827	431	85	19	4	0
truck	12875	5344	4203	1907	714	312	160	235
van	24956	13730	7176	2535	868	344	178	125
Total(%)	100	61	26	8.4	2.7	1.06	0.5	0.4

So, we can see that 61% of images are of resolution less than 50 x 50.

Image Resolution Vs Sample distribution



Sample Distribution of Images (<= 50 x 50)



Tentive list of tasks for next session :

- Understand Alexnet and SPP