Real-Time Facial Categorization using Convolutional Neural Networks

Somnath Sarkar 15CO247

Mehul Sharma 15CO130

Vikram Rathour 15CO252

Contents

Abstract

For the the Real-Time Facial Categorization using Convolutional Neural Networks project, we will make a website which uses the webcam interface of a computer to categorize the human face in the stream on the basis of a numerous factors. The faces can be categorized on the basis of gender, age, race, attractiveness and many other factors which are outlined at the end of this abstract. The information related to the categorizations would be shown on the website after the image is processed.

Introduction

Convolutional Neural Networks (CNNs / ConvNets)

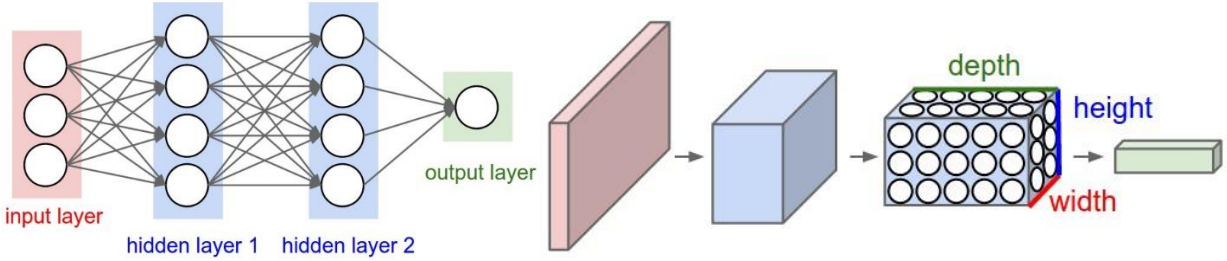
Convolutional Neural Networks are very similar to ordinary Neural Networks: they are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. The whole network still expresses a single differentiable score function: from the raw image pixels on one end to class scores at the other. And they still have a loss function (e.g. SVM/Softmax) on the last (fully-connected) layer and all the tips/tricks we developed for learning regular Neural Networks still apply.

ConvNet architectures make the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture. These then make the forward function more efficient to implement and vastly reduce the number of parameters in the network.

Architecture Overview

Regular Neural Nets don't scale well to full images. In CIFAR-10, images are only of size 32x32x3 (32 wide, 32 high, 3 color channels), so a single fully-connected neuron in a first hidden layer of a regular Neural Network would have 32*32*3 = 3072 weights. This amount still seems manageable, but clearly this fully-connected structure does not scale to larger images.

3D volumes of neurons. Convolutional Neural Networks take advantage of the fact that the input consists of images and they constrain the architecture in a more sensible way. In particular, unlike a regular Neural Network, the layers of a ConvNet have neurons arranged in 3 dimensions: width, height, depth. Note that the word depth here refers to the third dimension of an activation volume, not to the depth of a full Neural Network, which can refer to the total number of layers in a network.



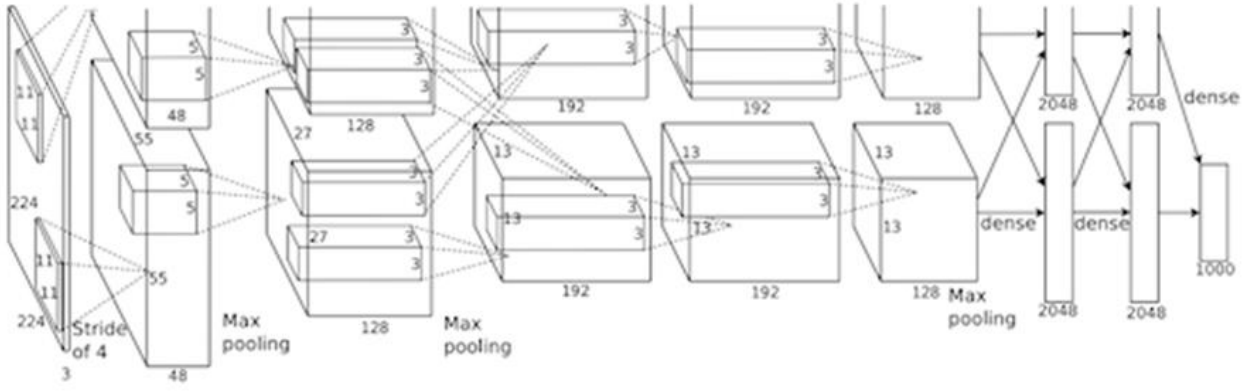
Layers used to build ConvNets

- INPUT [32x32x3] will hold the raw pixel values of the image, in this case an image of width 32, height 32, and with three color channels R,G,B.
- CONV layer will compute the output of neurons that are connected to local regions in the input, each computing a dot product between their weights and a small region they are connected to in the input volume. This may result in volume such as [32x32x12] if we decided to use 12 filters.
- RELU layer will apply an elementwise activation function, such as the max(0,x)
- thresholding at zero. This leaves the size of the volume unchanged ([32x32x12]).
- POOL layer will perform a down sampling operation along the spatial dimensions (width, height), resulting in volume such as [16x16x12].
- FC (i.e. fully-connected) layer will compute the class scores, resulting in volume of size [1x1x10], where each of the 10 numbers correspond to a class score, such as among the 10 categories of CIFAR-10. As with ordinary Neural Networks and as the name implies, each neuron in this layer will be connected to all the numbers in the previous volume.

Literature Survey

Related work

- AlexNet:- In the paper, the group discussed the architecture of the network (which was called AlexNet). They used a relatively simple layout, compared to modern architectures. The network was made up of 5 conv layers, max-pooling layers, dropout layers, and 3 fully connected layers. The network they designed was used for classification with 1000 possible categories.



AlexNet architecture (May look weird because there are two different "streams". This is because the training process was so computationally expensive that they had to split the training onto 2 GPUs)

https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

- VGG-16:-

Simplicity and depth. That's what a model created in 2014 (weren't the winners of ILSVRC 2014) best utilized with its 7.3% error rate. Karen Simonyan and Andrew Zisserman of the

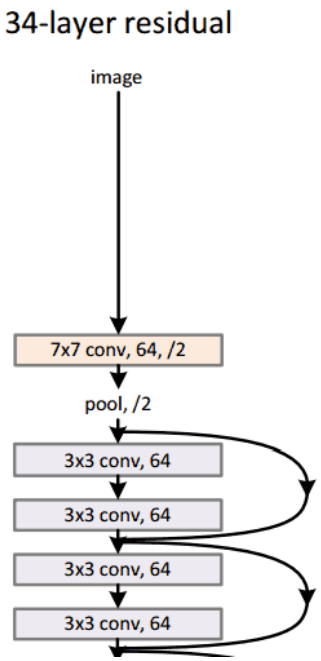
University of Oxford created a 19 layer CNN that strictly used 3x3 filters with stride and pad of 1, along with 2x2 maxpooling layers with stride 2.

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 imag)					
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					

The 6 different architecures of VGG Net. Configuration D produced the best results

https://arxiv.org/pdf/1409.1556v6.pdf

- ResNet:- Imagine a deep CNN architecture. Take that, double the number of layers, add a couple more, and it still probably isn't as deep as the ResNet architecture that Microsoft Research Asia came up with in late 2015. ResNet is a new 152 layer network architecture that set new records in classification, detection, and localization through one incredible architecture. Aside from the new record in terms of number of layers, ResNet won ILSVRC 2015 with an incredible error rate of 3.6% (Depending on their skill and expertise, humans generally hover around a 5-10% error rate.



https://arxiv.org/pdf/1512.03385v1.pdf

- Attribute and simile classification for Face Verification:-

1. Attribute Classifiers: We introduce classifiers for face verification, using 65 describable visual traits such as gender, age, race, hair color, etc.; the classifiers improve on the state-of-the-art, reducing overall error rates by 23.92% on LFW.
2. Simile Classifiers: We introduce classifiers for face verification, using similarities to a set of 60 reference faces; the classifiers improve on the state-of-the-art, reducing overall error rates by 26.34% on LFW. The simile classifiers do not require the manual labeling of training sets.
   http://www.cs.columbia.edu/CAVE/publications/pdfs/Kumar_ICCV09.pdf

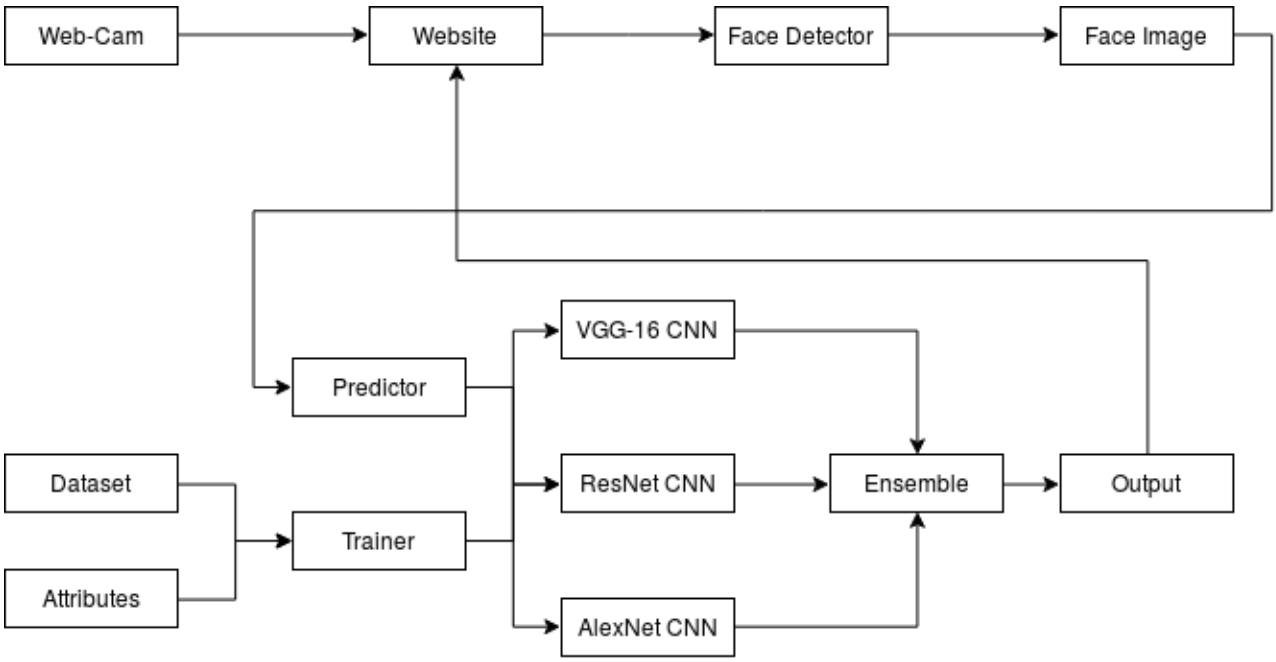
- HAAR:-
  This paper brings together new algorithms and insights to construct a framework for robust and extremely rapid object detection. This framework is demonstrated on, and in part motivated by, the task of face detection.

  https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/viola-cvpr-01.pdf

Problem Statement

- Implement a website which can categorize faces on various factors like gender, age, race etc. after taking an image of the user's face using a webcam.

Design



Implementation

Planned

We are planning on training multiple convoluted neural networks on a large data set of human faces. The data set has been crowd-labelled with numerous descriptors which we intend to predict on the provided user image. For training, we will apply several architectures and combine the results to form a strong predictor. On the front-end we will have a web interface which applies computer vision algorithms on a live webcam feed from the user, to isolate the face and pass it to our backend, which consists of a series of compiled trained models generated in the training step. Here, we pass the face images as inputs to each of the networks and return the output to the user back through the interface. We intend to make use of convolutional neural networks as the main model behind our classification system.

Finished

Pre-processing

The 2 datasets used for the project are as follows:

- Labelled Faces in the Wild : 13,000 Images of 5749 people labelled on a linear scale with a size of 250px x 250px. All the images are in RGB format.. The face are not aligned, and are collected from the internet.
- Celeb-A : 200,000 images of 10,177 people. The images are labelled on a liner scale with a size 218px x 178px. All the images are in RGB format. The faces are aligned and have been collected from the internet.

Training

- VGG-16: The model has been trained and categorization results related to gender, race and age have been achieved.

- Resnet: The coding phase of the architecture is complete.

Work to be done:

Backend

- AlexNet
- ResNet
- Adding Additional attributes

Front-End

- Implementing the Face Detection using OpenCV
- Web Interface