APPLICATION TO ORGANIZE PHOTOS USING AUTOMATIC IMAGE TAGGING

CS4089 Project

Midterm Report

Arjun M S, Joe Toms Panikulam, Nitin Sukumar, Noorul Ameen K M
Guided By: Dr. Saidalavi Kalady

October 4, 2017

1 Introduction

Over the past decade,the number of images being captured and shared has grown enormously. There are several factors behind this remarkable trend. In the modern age,it is now commonplace for private individuals to own at least one digital camera,either attached to a mobile phone, or as a separate device in its own right.Furthermore, the popularity of social networking websites such as Facebook has given users an extra incentive to capture images to share and distribute amongst friends all over the world.

The main objective of this project is to classify images by assigning appropriate tags and place them into appropriate folders according to the objects present in the images and enable users to search the images using the tag names or object names that are present in the images and retrieve the images. These objectives are refined down into three sub-objectives:
1. Extract a discriminative set of image features from the datasets.
2.Implement an efficient version of the original image tagging algorithm.
3.Evaluate image tagging accuracy and image retrieval performance on the standard datasets.

Artificial neural networks (ANNs) are computational models inspired by the human brain. At their core, they are comprised of a large number of connected nodes, each of which performs a simple mathematical operation. Each node's output is determined by this operation, as well as a set of parameters that are specific to that node. By connecting these nodes together and carefully setting their parameters, very complex functions can be learned and calculated.

In machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations.When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features (also named a feature vector).

2 Problem Statement

To develop and design a system that would identify certain images on its own and organizes an efficient picture gallery based on the tags they represent.The need for Automated Image Tagging was a source of concern in the field of AI. Machines, unlike human beings was believed to be deprived of the ability to think on its own.To tackle this anomaly we came up with the concept of creating neural networks to train some ex-

isting data such that further instances of such data could be identified and labeled accordingly.To avail this technology high end servers and network clouds and multi-core GPUs are needed. To create such a network within the limitation of a medium end machine was a challenge in itself. Without any high memory storage servers and without high end GPUs , Real-Time Image Tagging could be achieved by a simple web-application using MATLAB[8].

3 Literature Survey

Vailaya[4] proposed a hierarchical classification scheme to first classify images into indoor or outdoor categories, then, outdoor images are further classified as city or landscape; finally, landscape images are classified into sunset, forest, and mountain classes. In other words, three Bayes classifiers are used for the three-stage classification.

Spirit Tagger is a geo-aware tag suggestion tool which uses Flickr that depicts geographically relevant tags for images with GPS coordinates [6]. It is done by combining the geographical context with content-based image analysis. Geographic mining is done by collecting a set of images that are within a certain radius of the candidate image to be tagged. This set of images is narrowed down by using visual similarity techniques. The tags of the images in the set are then compared to their global frequency. Local frequency refers to the frequency of a tag in the result set, whereas global frequency refers to the frequency of a tag in all images on Flickr. Tags with higher local frequency than global frequency are assumed to be relevant for the query image. Experiments have found that Spirit-Tagger works well as compared to baseline methods that only use geographical context [7]

Alex Krizhevsky[1] analyzed the problem of classifying thousands of objects from millions of images and necessity of a model with large learning capacity.The shear complexity of problem meant that the task of object recognition from images could not be specified by a large dataset like ImageNet[1] alone.The requirement of a model that has a lot of prior knowledge led him to use Convolutional neural networks(CNNs)[1,2].The learning capacity of the model could be controlled by varying their depth and breadth and hence they could be trained to make mostly correct assumptions about the nature of images.The performance of CNNs in contrast to a standard feedforward neural network is theoretically comparable and at the same time easier to train as they have much fewer connections and parameters.

They successfully implemented a neural net-Alexnet[1] which had eight layers with weights; the first five were convolutional and the remaining three were fullyconnected. The output of the last fully-connected layer was fed to a 1000-way softmax which produced a distribution over the 1000 class labels.They achieved top-1 and top-5 error rates of 37.5% and 17.0% which was considerably better than the previous state-of-the-art neural network and was the winner of ILSVRC-2012 competition.

Alex Krizhevsky[1] indicated that despite the attractive qualities of CNNs, and despite the relative efficiency of their architecture, they have still been prohibitively expensive to apply in large scale to high-resolution images as performance was seriously affected by lesser clocked CPUS and GPUs.The network took between five and six days to train on two GTX 580 3GB GPUs and this can be seriously improved to minutes with the help of current state of the art GPUs.But the current support to train neural networks on powerful GPUs which grow faster every year with the highly-optimized implementation of 2D convolution meant that the problem of object recognition could be solved much more faster and efficiently.Apart from

this,the fact that we have datasets such as ImageNet containing over 15 million labeled high-resolution images belonging to roughly 22,000 categories[1] made training without overfitting much easier.

As an alternative to using a neural network from scratch,S. J. Pan and Q. Yang[5]did a study on transfer learning and how it can improve accuracy in object classification on neural nets.It aimed to improve the process of learning by using experience gained through the solution of a similar problem. ie; if a solution to particular problem doesnt exist then find a mapping with another problem whose solution exists.

So in transfer learning, a pretrained convolutional neural network for eg;Alexnet[1] is taken and its last few layers(fully connected layers) which is used for the classification of 1000 objects will be trimmed,and converted into a layer such that it will classify only few objects.It will then treat the rest of the convolutional neural networks as fixed feature extractor for the new dataset.

Also transfer learning, allows the domains, tasks, and distributions used in training and testing to be different. In the real world, we observe many examples of transfer learning. [5]Learning to recognize apples might help to recognize pears. Similarly, learning to play the electronic organ may help facilitate learning the piano.

Transfer learning methods can be used to improve the accuracy, but the drawback of this method is that although time taken to train is less than building a neural network from scratch ,it is still higher than compared to the using a pre trained convolutional neural network and the requirement of a good hardware is high.As we understand the study of transfer learning is motivated by the fact that people can intelligently apply knowledge learned previously to solve new problems faster or with better solutions.

As we learned from the work of Matthew D. Zeiler and Rob Fergus[2] who analysed the different layers of Alexnet[1] and visualised the input image through all these layers,every unit/layer in Alexnet[1] extracts a distinctive feature, that is some function of the input and helps the neural net to discriminate between the classes it trains on.(the classes become approximately linearly separable)[2].These predictions from an inner layer are known as activations or features .The network constructs a hierarchical representation of input images. Deeper layers contain higher-level features, constructed using the lower-level features of earlier layers and it allows the output layer to easily separate the classes .

Their research laid out the set of features that each layer in Alexnet[1] identifies.Layer 2 responded to corners and other edge/color conjunctions.Layer 3 had more complex invariances, capturing similar textures (e.g. mesh patterns; text). Layer 4 showed significant variation, and was more class-specific: dog faces; birds legs. Layer 5 showed entire objects with significant pose variation, e.g. keyboards and dogs.[2]

To explore how discriminative the features in each layer was, they tried to classify images based on the features they extracted from each layer and checked for the accuracy of these results.[2]They performed it on both Caltech-101[3] and Caltech-256 datasets using SVM/Softmax classifiers.SVM classifier on Layer-1 gave an accuracy of 44.8% and 24.6% on Caltech-101 and Caltech-256 datasets respectively.It showed a drastic change on Layer-7 when it gave accuracies of 85.5% and 71.7% each on Caltech-101 and Caltech-256 datasets respectively.This supported the premise that as the feature hierarchies become deeper, they learn increasingly powerful features and hence increases accuracy of the neural network.[2]

To make optimum utilization of representational power of pre- trained deep

neural networks we extracted features from Layer-7 of Alexnet[1] as indicated by research of Matthew D. Zeiler and Rob Fergus[2] and used it as as predictor variables to fit a multiclass Support Vector Machine (SVM) using Naive-Bayessian Classifier.Their research has shown features to be far from random, uninterpretable patterns.Rather, they show many intuitively desirable properties such as compositionality, increasing invariance and class discrimination as we ascend the layers.They also showed how these visualization can be used to identify problems with the model and so obtain better results, for example improving on Alexnets[1] impressive ImageNet 2012 result.

4 Work Done

We have now envisioned our application as a web interface where users can upload all the pictures and our application groups them into folders based on the objects present in the images.They can also view and download them folder wise.To check the viability of our application we had conducted tests to determine the speed and accuracy of detecting objects via Matlab[8] Application.
We performed training and testing models using CPUS,Parallel CPUS AND GPUS.Methods we used were:
1)Testing with pre-trained neural net like AlexNet.[1]
2)Using transfer learning with Alexnet[1]
3)Using Feature extraction with Alexnet[1] and SVM classifier.

From the results we obtained,we concluded that Feature extraction done with SVM,fitted with Naive-Bayessian Classifier and run on GPU(Nvidia GeForce 920M) outperformed transfer learning.We could successfully classify 4 objects(airplane,frog,deer,bird) trained on 1000 images each with 86.55% accuracy.We used CIFAR-10[9] Dataset and also Caltech101[3] Dataset for the same and tested it on 4000 images.The Neural Net took 2min 15sec to classify the same.A single test image could also be classified with this neural net under 10 secs when trained with 50 images.

5 Future Work

Building a computer system that automatically tags images with the same ease with which we humans do the job is one of the most challenging tasks in Deep Learning. Even though we were able to use a neural network that classified 4 objects with an accuracy of 86.55% in 2min 15 sec, this is far from the end product we envision. In a world where time is of essence, optimizing and increasing the accuracy of these networks is of vital importance. For this, we plan to find ways in which we can optimize both the training and testing time of the neural network further using different techniques in feature extraction.
As the number of images trained increases the accuracy (top 5 or top 1 error rate) of a neural network, incorporating more number of images without losing efficiency is another challenge for which we need to find a solution. We also plan to find ways to maximize the performance of the network by usage of parallel/multi GPU's in our system. As we envision our project to be a user-friendly web application, we will also start implementing the basic interface and design of the application in the upcoming months.

References

[1] Krizhevsky Alex,Sutskever Ilya and Hinton Geoffrey E. *Imagenet classification with deep convolutional neural networks* in Advances in neural information processing systems , 2012 (pp. 1097-1105).

[2] Matthew D. Zeiler and Rob Fergus. *Visualizing and understanding convolutional networks.* in European con-

ference on computer vision 2014 (pp. 818-833).

[3] L. Fei-Fei, R. Fergus and P. Perona.*Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories.* IEEE. CVPR 2004, Workshop on Generative-Model Based Vision.

[4] Dirichl ,A. Vailaya, M. Figueiredo, d H.Zhang,*Image classification for content-based indexing* IEEE Transactions on Image Processing,2001,vol.10,no.1, pp.117130.

[5] S. J. Pan and Q. Yang. *A Survey on Transfer Learning* in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[6] Moxley, E., J. Kleban, and B.S. Manjunath.*SpiritTagger: A geo-aware Tag suggestion tool mined from flickr* in Proc.1st ACM international conference on Multimedia information retrieval. 2008, ACM: Vancouver, British Columbia, Canada. p. 24-30.

[7] Martin Haetta Evertsen.*Automatic Image Tagging based on Context Information* Master's Thesis in Computer Science, Faculty of Science and Technology, University of Troms, June 2010.

[8] MATLAB and Statistics Toolbox Release 2017a, *The MathWorks, Inc.*, Natick, Massachusetts, United States.

[9] Alex Krizhevsky *Learning Multiple Layers of Features from Tiny Images* 2009).