\documentclass [11pt,letterpaper ,twoside ,openright ]{report}

\title{Localizing Little Landmark with Transfer Learning}

\author{Sharad Kumar}

\usepackage{setspace}

\usepackage{geometry}

\geometry{margin=1in}

\usepackage{sectsty}

\chapternumberfont{\Large}

\chaptertitlefont{\huge}

\usepackage{graphicx}

\usepackage{subcaption}

\usepackage{float}

%\floatstyle{boxed}

%\restylefloat{figure}

\usepackage{chngcntr}

\counterwithout{figure}{chapter}

\usepackage{wrapfig}

\begin{document}

\maketitle

\tableofcontents

\begin{abstract}

In this work we explore the possibility of using context information to localize small objects in an image. To localize on the small object - like location of hand-holding-leash - in a dog walking image, we create a regression model using Convolutional Neural Network (CNN) that is supervised by the coordinates of the small object in an image. Since small objects do not have strong visual characteristics in an image, it's difficult for neurons to discern their pattern because the feature map exhibit low resolution for small objects, which means much weaker signal for the neurons to recognize. Use of context for object detection and localization has been studied for a long time. This idea is explored in [1] for small object localization by using a multistep regression process where spatial context is used effectively to localize on door handle of a car in the image of a car. We extend the idea in this paper and demonstrate that the technique of localizing in steps using contextual information when used with transfer learning can reduce the training time.

\end{abstract}

\chapter{Introduction}

\doublespacing

Because of complexity and variability in static images, semantic understanding of an image has remained one of the most difficult problem to solve in computer vision. While human eye is trained to rapidly cull relevant information from an image to build a language level understanding of the scene, such feat has remained elusive artificial intelligence systems. Understanding relationships between objects in an image can help understand the semantic meaning of an image as it provides evidence for a particular situation being present in the image. For example, the ability of an artificial intelligence system to localize on a "hand-holding-leash" relationship in an image provides a fairly good clue that the image most likely has a dog and a person walking the dog and this information can be used to generate semantic understanding of the image.

In this work we focus to localize on hand-holding-leash relationship in dog walking images. We only consider prototypical images for localization, where the image has only one dog and one dog-walker in the picture. [1] proposed the idea of using successively more relevant contextual information in a sequence of steps using a recurrent architecture to localize on the small object. In this scheme, only the final step of the sequence is supervised by the location of the small object and predicts the target of small object location, while other steps predict where to look next. The learning therefore discovers globally distinctive pattern to start the sequence and conditionally distinctive ones to get closer to the target in discrete steps.

Using Tensorflow® we implement a model of the CNN model described in [1]. We use the car door handle dataset, provided by authors of [1], to train and test the model. We also train this model on our dog walking image dataset. We compare the results and report our findings in section xxx. In section xxx we provide more details about the datasets we used for our experiments.

Our contribution in this work is to incorporate transfer learning to reduce training time for small object localization. The motivation of using transfer learning is the fact that generalizations learned by a model that has been trained on a large dataset can be effectively used as input activations to some other model. We propose a slightly different model by repurposing a pre-trained VGG-16 [2] model where we use activations from pool4 layer of the VGG-16 model. VGG-16 is a deep CNN model proposed by K. Simonyan and A. Zisserman from the University of Oxford that achieved 92.7% accuracy in top-5 test category on Imagenet dataset. The Imagenet dataset [3] is a dataset of over 14 million images belonging to 1000 categories. We demonstrate significant improvement in training time with this approach and report our results in section xxx.

\chapter{Related work}

\doublespacing

Localization of small object in an image using CNN is little understudied. Chen et al. in [6] extend the R-CNN algorithm to detect small objects - like computer mouse on a desk, or a faucet in a kitchen. In their approach, they use a modified Region Proposal Network (RPN) [7] by choosing object proposals many times smaller than used in the original RPN and to use context information, they also crop a region centered at region proposal, but bigger than the region proposal. The region proposal and the context proposal are fed to two parallel CNNs and their concatenated activation are used as input to a third CNN to make predictions. In this work authors use Intersection over Union (IoU) as performance metric whereas in our work we use Euclidian distance between original and predicted coordinated normalized by the bounding box of the object as explained in section xxx.

Another important work for small object identification is done by Eggert et al. in [8] where they modify Faster R-CNN that leverages higher-resolution feature maps for brand logo detection. Their work qualifies as small object detection since they are trying to locate brand logos in pictures that were not intended to capture it - like image of a soft drink brand in a picture taken at an outdoor concert venue, or image of a sport brand on a person's shirt who is walking a dog - and therefore tend to occupy small image area. In their work they attempt to generate better region proposals and assume a perfect classifier. They compare performance of region proposal generator using activations from conv3, conv4 and conv5 layers of pre-trained VGG16. They found that conv3 and conv4 layers' activation performed better than conv5 layer activations.

The paper that we extend from on this work [1] proposes an architecture that is recurrent in the sense that the feature map generated by one step of the model is encode as contextual information and fed as input to the next step in the sequence along with feature map generated by the convolutional layer. Another important work that explores this idea is by Zuo et al in [9]. In their work they argue that convolutional and pooling layers in a CNN are performed locally without considering other regions of the image and therefore fail to capture contextual dependencies for better representation. They propose a model that encodes this correlation for better performance.

We use transfer learning to demonstrate that instead of training a network from scratch, using a pre-trained network may result in significant improvement in training time. Pan et al. in [10] do an in depth study of feasibility of transfer learning and show that knowledge learned by a model in one domain can be transferred to another machine learning model in a different domain even when the feature space and/or the data distribution of source and target systems is not the same. Shin et al. also reiterate this idea in [11] where they employ transfer learning to fine-tune a CNN model pre-trained on natural image dataset (RGB) to medical image (monochrome) task. This idea has been successfully used by numerous researchers and practitioners in image classification and localization tasks by using patterns learned by deep CNN models trained on enormous amount of images.

\chapter{Deep Learning/Convolutional Neural Network(CNN)}

\doublespacing

Since its resurgence in 2012 when a neural network based architecture, called AlexNet [Krizhevsky et al.], was proposed for image classification of ImageNet [3] dataset, neural networks variant deep Convolutional Neural Networks (CNN) have achieved significant improvement in state-of-the-art for classification as well as regression tasks. A CNN model consists of many layers (Figure 1) and each layer learns features or representations at increasing level of abstraction as demonstrated by [4, 5] using a deconvnet that map these representations back to the input pixel space. CNN were created to overcome the scaling problem of traditional neural networks. In traditional neural networks neurons in one layer of the network are connected to all the neurons in adjacent layers, which makes it difficult to scale for intelligent image understanding and analysis. Neurons (also called filters or kernels) in each layer of the CNN are connected only to a small region of the input and are three dimensional (Figure 2).

\begin{figure}[h]

\centering

\begin{subfigure}[b]{0.75\linewidth}

\includegraphics[width=\linewidth]{Images/Figure1-CNN}

\end{subfigure}

\caption{Convolutional Neural Net}

\label{fig:cnn}

\end{figure}

\begin{wrapfigure}{l}{0.5\textwidth}

\begin{center}

\includegraphics[width=0.48\textwidth]{Images/Figure2-CNNFilter}

\end{center}

\caption{CNN Filter}

\end{wrapfigure}

The height and width of the filter determines the region of spatial connectivity of the filter to a layer before. Depth of the filter indicates how many such filters are there in what's called a filter bank. A filter contains real numbered weights. The dot product of these weights with the input region that it's spatially connected to creates output activation map (also called feature map), as depicted in Figure 2 below. A filter convolves (slides) across it's input with a predefined stride. As illustrated in Figure 1, the input is zero padded if the filter does not align with the image at the edges as it convolves. The weight in the filter remains unchanged or is shared as convolutions are performed. The rationale of sharing weights is twofold: first it reduces the number of parameters in the model, and secondly any useful features identified in one part of the image can be re-used everywhere else without having to be independently learned [Murphy 2012 (Machine Learning, section 16.5.1)].\\

%\paragraph

\noindent

\textbf{\textit{Forward Pass}} \\

There are mainly three types of components in a CNN layer - convolution, Rectified Linear Unit (ReLU) and pooling - and the layers are arranged in a sequence, as depicted in Figure 1. Each CNN layer transforms one volume of activations into another volume during the forward pass of processing. Filters are mainly involved in computation in the convolution layer and process its input as explained above. ReLU layer introduces non-linearity into the model by applying max (0,x) function element-wise to the activations produced by the filters in the convolution layer. Use of max (0,x) non-linear function results in accelerated training as compared to logistic or hyperbolic tangent function [Krizhevsky, ImageNet Classification with Deep Convolutional Neural Networks], where the gradient in the saturating part of the activation function graph becomes very small (for very large or very small weights) resulting in sluggish training because of minuscule weight change during back-propagation. Introducing non-linearity with ReLU ensures that the CNN model does not collapse into a large linear model [Murphy 2012 (Machine Learning, section 16.5.1), Hastie et al., Essentials of Statistical Learning (Section 11.3)]. Neural networks can therefore be thought of as a nonlinear generalization of the linear model and by introducing the nonlinear transformation it greatly enlarges the class of linear model [Hastie et al., Essentials of Statistical Learning (Section 11.3)].

\begin{wrapfigure}{r}{0.5\textwidth}

\begin{center}

\includegraphics[width=0.48\textwidth]{Images/Figure3-Pooling}

\end{center}

\caption{Average Pooling}

\end{wrapfigure}

Activations from the ReLU layer are subsampled by the pooling layer. As shown in Figure3, pooling layer subsamples the activations spatially and is usually done either by averaging or by computing the max over a small window of activations produced by the ReLU layer. Subsampling reduces the number of parameters in the model and hence the amount of computation in the network, it also results in small shift invariance as the image activations propagate through the network. Because of the subsampling, there's a progressive reduction in size of activations, as shown in Figure 1. \\

\noindent

\textbf{\textit{Back Propagation}}\\

Briefly explain about how backprop is used in little landmark model

Weights in the filters are the parameters of a CNN model and model fitting involves learning these weights using back-propagation. Back-propagation is a optimization algorithm where the goal is to find a set of value for model parameters that generalizes the learning in such a way that the model achieves superior classification or regression accuracy on test dataset. This goal is achieved by minimizing the error on training dataset by using gradient descent (or its variant) method for optimization. Validation dataset complements the training dataset to find training hyper-parameters that generalizes learning. Stochastic gradient descent is one of the most effective techniques for back-propagation where gradient of the cost function and chain rule of derivatives are used to update model weights iteratively in minibatches. Use of minibatches in stochastic GD accelerates the training and also helps to obtain unbiased estimate of the gradient by taking the average gradient on a minibatch, assuming examples in minibatches are drawn in identically and independently distributed (IID) fashion [Goodfellow et al. (Deep Learning, 2016, Section 8.3)].\\

\noindent

\textbf{\textit{Overfitting and regularization}}\\

Because of large number of parameters, CNN models have proclivity to overfit when not designed with suitable regularization parameters or when there is not sufficient data to train the model. Large number of parameters in a neural network may result in a model that shows high variance when the model is trained on different training datasets [12]. Overfitting is a manifestation of this problem in neural network where training fails to achieve good model generalizations resulting in good training accuracy, but poor performance on test dataset. Bias in a model is the other source of prediction error which is usually caused by wrong selection of model for the task. In case of overfitting, variance rather than bias dominates the estimation error. Regularization addresses the problem of overfitting by reducing variance significantly while not increasing the bias [13]. L1 and L2 are two most commonly used regularization techniques in machine learning in general, including neural networks. Dropout [14] is another regularization technique used specifically with neural network. In our implementation of localizing on dog walking image, we explored L1 as well as L2 regularization as well as dropout and we report our findings in the experiments section xxx. In [1] authors use L2 regularization as explained in section xxx.

\chapter{Detecting Small Object in an Image}

\doublespacing

An efficient object detection model in an image may use region proposals, as one used by Girshick et al. in R-CNN [7, 15]. This two stage architecture is current state-of-the-art where the first stage extracts region proposals and the second stage uses these region proposals as input to a CNN which is pre-trained on an auxiliary dataset. This setup has shown remarkable performance improvement for prominent object detection in an image, but is not suitable for small landmarks as small objects do not have very distinctive features. A simple solution to address the issue of lack of distinctness for little landmark in an image can be either to magnify the image or take a high resolution image. As explained by Eggert et al. [8] this simple approach may not be very effective since computing convolutions in CNN grows quadratically with image dimensions and will result in unnecessary computation. Moreover, as explained by Chen et al. [6], low resolution inputs for small objects is deeply embedded in the nature of visual perception and a robust vision system should be able to deal with it.

This limitation of R-CNN beckons a different deep learning architecture to localize small objects in an image. Singh et al. [1] propose a stepwise regression model for this task and they call this model an architecture for localizing little landmark. This is a recurrent model that is trained end to end using the location of the little landmark. Authors demonstrate the robustness of this model by testing and training it on several datasets. One of contribution of their work [1] was to create and annotate a dataset of little landmarks and for this task authors use Stanford's car dataset [15] and annotate the dataset with location of the car door handle for little landmark localization. We use this Car Door Handle (CDH) dataset and our own Dog Walking Images (DWI) dataset for training and testing our models. Details of these datasets are provided in section xxx.

\section{Architecture Overview}

The CNN architecture for small object detection in [1] exploits the fact that a little landmark is defined by it's context. Authors define latent landmark as a location that are more prominent than the little landmark and can be used as context to detect the little landmark conditionally. The localization in this architecture happens in a sequence of steps where a series of latent landmarks lead to the target little landmark by providing contextual information. Figure 4 below illustrates this process with two examples. In the image on the left hand side, the car door handle is detected in a sequence of three steps where a trained network finds the first latent landmark (in red) near the front wheel, the second latent landmark in the bottom half of the front door (in green) which helps find the target location (in blue) in the third step. Similar pattern can be seen in the image on the right hand side in Figure 4 where edge of the electrical switchboard is used as the starting point to localize on the electrical switch on a wall in three steps.

\begin{figure}[h!]

\centering

\begin{subfigure}[b]{0.45\linewidth}

\includegraphics[height=4.5cm, width=6.75cm]{Images/Figure4a-CDH}

\caption{Coffee.}

\end{subfigure}

\begin{subfigure}[b]{0.45\linewidth}

\includegraphics[height=4.5cm, width=6.75cm]{Images/Figure4b-LS}

\caption{More coffee.}

\end{subfigure}

\caption{The same cup of coffee. Two times.}

\label{fig:coffee}

\end{figure}

Figure 5 below shows a schematic diagram of the little landmark localization architecture that localizes on a car door handle in three steps. In this recurrent CNN architecture, the bottom convolutional layers are used to extract image features which is common input to all steps. Three columns in the upper portion of the figure corresponds to the stepwise procedure that detects little landmark in the last step. Every prediction step, s in the model generates location of the latent landmark (red blob) and predicts the location of next latent landmark (blue blob). The prediction for next step is encoded as a feature map, P(s) and along with image features is used as input to the next step. Loss calculation is explained later in section xxx, but it's worth noting here that the difference between prediction (blue blob) in step s-1 and the generated latent landmark location (red blob) in step s contributes to total error of an iteration.

\begin{wrapfigure}{l}[9mm]{0.6\textwidth}

\begin{center}

\includegraphics[width=0.48\textwidth]{Images/Figure5-Architecture}

\end{center}

\caption{Architecture}

\end{wrapfigure}

Weights in row two and three of s steps are shared. It must be noted that this CNN model does not use the pooling layer. Although authors do not explain why, but we surmise that since pooling will result in a low resolution feature map than the original one, it may make identification of little landmark more difficult as it already has non-distinctive characteristics.

\section{Image Pre-processing}

A batch of ten images is created as input to the model. Image height and width are adjusted so that all images in a batch have same size. When necessary, images are zero padded to match the longest width and longest height among all images in a batch. Images are normalized by performing mean subtraction and standard deviation scaling. Random scale jitter is added to images and image are also randomly flipped to introduce noise to ensure that model does not overfit on training data.

A location grid, \textit{\texttt{out\\_locs}}, is created for the batch which corresponds to spatial dimension of the output from the CONV/RELU6 layer of the bottom portion of the model. This location grid is used for generating the latent landmarks (red blobs in figure 5) as explained later in section xxx.

\end{document}