What's the story

Concept Learning - understanding the dynamics of relationship between objects - like person, dog and leash. This is

done by attempting to localize on of hand-holding leash and attached-to-leash part of the image.

What are the challenges in localizing on hand-holding-leash - the small object localization problem.

Learning by use of context. Use other examples.

**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.

**Introduction**

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 achieves 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.

**Related work**

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 assuming 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.

**Deep Learning/CNN**

What is deep learning/CNN

How is it used for regression

Why could this be a good fit for this problem

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).

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)].

|  |  |
| --- | --- |
| nput  24x24  Convolution  Feature maps  4@20x20  Feature maps  4@10x10  Subsampltng  Feature maps  8@8x8  Convolution  Feature maps  8@4x4  Subsampling  Output  20@1x1  Convolution | Macintosh HD:Users:sharadkumar:Library:Group Containers:UBF8T346G9.Office:msoclip1:01:110B11FE-E84D-0E4E-BA75-546CFD0CA5A3.png |
| Figure 1: Explain    *From:* [*https://www.slideshare.net/hanneshapke/introduction-to-convolutional-neural-networks*](https://www.slideshare.net/hanneshapke/introduction-to-convolutional-neural-networks) | Figure 2: Illustrates filter's transformation of the input into output. The figure illustrates two-dimensional input, but usually the input also has a third dimension of depth which is matched by filter's third dimension |

***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)].

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.

|  |
| --- |
| 5 9  12 7  Average Pooling  •aan  21 12  18 10  Max Pooling |
| Figure 3: Explain    *From:* <https://medium.com/@Aj.Cheng/convolutional-neural-network-d9f69e473feb> |

***Back Propagation***

Weights in the filters are the parameters of a CNN model and model fitting involves learning these weights using back-propagation as forward and backward passes are made over the network. 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)].

Challenges of overfitting: regularization, momentum, dropout

Weight initialization

***Classification vs. Regression***

Dataset

CDHD and Dog Walking images

Papers used and explored

Little Landmark - Singh

VGG-16 paper

Alexnet paper

Deconvnet paper (Adaptive deconvolutional networks for mid and high level feature learning) - Fergus

Dropout - Hinton

Visualizing and Understanding Convolutional Networks - Fergus

ROC Paper

Learning Invariant Feature Hierarchies - Yann LeCun (<http://yann.lecun.com/exdb/publis/pdf/lecun-eccv-12.pdf>)

ROC curves for regression (José Hernández-Orallo)

Main Paper

Explain architecture/model

Why was max pooling not used

Explain the recurrent part of the architecture

Explain with help of pictures how the coordinates are localized in three steps

Explain preprocessing

Explain how out\_locs values are created - using calculation from Stanford's notes

Performance of version with distortion and without distortion. Right now the training results I have are all with distortion. Answer the question why are distortions done.

Explain Performance Metric

Performance Metrics for Evaluating Object and Human Detection and Tracking Systems

My implementation of the model

Tensorflow/Python

My Models and Experiments

Discuss what transfer learning is

Different layers of the network learn different types of abstractions (explain more). Murphy pp 995 - explain transfer learning using this

Discuss what VGG-16 model is and how was transfer learning incorporated in the model

Discuss Results

It took 27 hours to achieve same result with transfer learning

Training the entire model vs. fine tuning

Explain two models

First for localizing hand-holding-leash

Second for extending the model for identifying two small objects in one training

Explain the vanishing gradient problem that happened in getNormalizedLocationWeightsFast() function and how it was solved by using normalization. Overlapping Pooling may have solved this problem (ImageNet Classification with Deep Convolutional Neural Networks, section 3.4)

Pooling layer is not used since downsampling by maxpooling results in low resolution feature map which may make it difficult to identify small image. (logo paper)

Explain the problem of over-fitting

that it may be because the context around hand holding leash is not sufficient for localizing

It may be because of not enough data and transfer learning did not help in this case

Use pictures to show where latent coordinates and the final coordinates fall during training, which is pretty close

Discussed ideas used to solve the over-fitting - different hyper-parameters (momentum, weight decay and dropouts)

dropout also has the amazing property that it prevents co-adaptation of feature detectors which improves a network’s ability to generalize (<https://dmm613.wordpress.com/2014/10/15/intriguing-properties-of-neural-networks/>)

Discuss about use of different types of optimizers - Adam and SGD

In this case fine tuning is done to the columns. The assumption is that VGG 16 model is generalized enough and does not have overfitting problem. Because of this the weights in the VGG-16 model are not changed.

Miscellaneous

Include discussion about bias and variance and explore if that can be used in explaining something

Future work:

Explore possibility of removing pooling layers from pre-trained VGG model. This may help since downsampling in the pooling layer renders the low resolution feature map making it difficult to identify a small object (logo paper, page 1, Introdiution section)

**Conclusion:**

One of the reasons why DWI did not localize:

[1] The target landmark may have a local appearance that is similar to many other locations in the image. However, it may occur in a consistent spatial configuration with some pattern, such as an object or part, that is easier to find and would resolve the ambiguity.

There are much more possibility of the location of small objects (RCNN for small objects)

Metric used for measuring performance of small object detection is not consistent - some researchers use IoU whereas others use some sort of normalization. A future work for this is developing good metric for measuring performance of small object detection.

Bibliography

[1] Singh, Hoiem and Forsyth. Learning to Localize Little Landmarks. In CVPR 2016.

[2] VGG-16 model

[3] Imagenet dataset

[4] Visualizing and Understanding Convolutional Networks - Fergus

[5] Adaptive deconvolutional networks for mid and high level feature learning

[6] R-CNN for Small Object Detection, Chen, C.; Liu, M.-Y.; Tuzel, C.O.; Xiao, J.

[7] Faster R-CNN

[8] Company Logo Detection

[9] Convolutional Recurrent Neural Networks: Learning Spatial Dependencies for Image Representation

[10] A Survey on Transfer Learning

[11 Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning