Automatic Portrait Segmentation of Cats and Dogs

Kevin Mueller, Gordon Erlebacher, Jeremy Quinto*, Iman Dhillon*

Department of Scientific Computing, Florida State University
*NewSciLabs, Tallahassee FL

Abstract

With the ever-growing amount of image data available on the internet, there is an increasing need to develop algorithms that can automatically detect regions of interest (ROI) in images. Currently, almost all approaches for solving this task rely on difficult-to-create supervised datasets that require labels for individual pixels. Here we present an end-to-end framework for automatically cropping images of cats, dogs, and humans by utilizing post-manipulated images prepared by graphic designers. We accomplish this task by first applying an image reconciliation technique that creates pixel masks from the post-manipulated images and then implement a machine learning technique that learns the masks from the original images by utilizing Deep Gaze II, a deep learning model for creating saliency maps trained on human eye-tracking data, as pre-training for finding likely ROIs.

Introduction

Image segmentation is a popular problem in machine learning with numerous applications. There are many different approaches for image segmentation, but they can generally be divided into supervised and unsupervised approaches. Supervised approaches use labeled pixels to learn the correct segmented classes, while unsupervised methods make use of local image properties, such as sharp edges, as a way to segment an image. In general, supervised approaches are the most successful. For example, semantic segmentation treats each individual pixel as a label for a set number of classes and learns a mapping (typically with a deep neural network) from the input pixels to the class labels.

Another approach for image segmentation relies on mimicking how humans process images. In particular, various datasets for eye-tracking have been studied as starting points for detecting ROIs. More recently, this research direction has been combined with deep learning to create a power approach for learning saliency maps. Deep Gaze II is one example of this type of network,which works by randomly sampling feature maps from VGG-16 and passing them through a read-out convolutional network.

For this work we consider a retraining of the Deep Gaze II network on a custom dataset for portrait segmentation. Portrait segmentation is abstractly defined as finding the best portrait-like segmentation for a particular image. Typically, it includes background removal and important facial/head features, but may also include parts of the neck and various accessories.

Dataset

- The original dataset consists of approximately 40,000 pairs of photos of cats, dogs, humans, and other animals.
- Each pair consists of an original photo and a post-processed image prepared by professional graphic designers to mimic/represent what subjectively makes a good portrait.
- In order to simplify the problem, we only consider images of cats and dogs.
- The dataset was split into training/test datasets with a 0.2% train/test split.

Image Reconciliation Model

Since the goal of this project is to automatically extract the crop from the original image, it is necessary to first reconcile the post-processed image with the original image. However, since the post-processed image typically undergoes significant modification (e.g. as color, rotation, scale, lighting), this can be a formidable task. In order to create lighting and color invariance, we first apply the Laplacian as a way to extract the edges of both the post-processed and original image. We then loop through a grid of rotations/scales and calculate auto-correlations by sliding the post-processed image as a kernel across the original image.

Due to the brute force method of looping over a large number of rotation/scale combinations, it is useful to utilize GPUs to quickly perform the convolutions. To accomplish this task, we use pytorch to create a convolutional layer (with appropriate padding) with a total number of filters equal to the total number of looped rotations and scales.

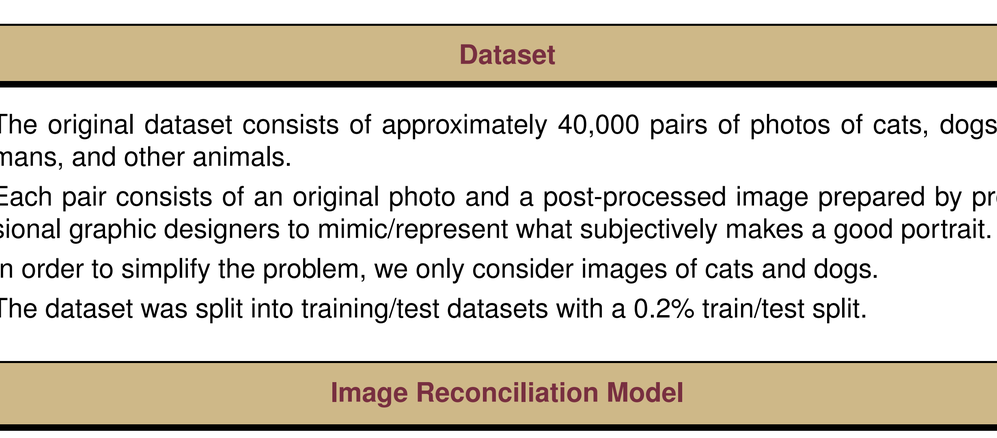

Figure 1: *The Laplacian is applied to the original image and all scales and rotations of the post-processed image. The resized and rotated images are then convolved with the Laplacian of the original image and the maximum convolution value is calculated. The optimal rotation/scale is found by taking the maximum of the maximum convolutions. Finally, the maximum position of the convolution for the optimal rotation/scale is taken to be the final reconciled mask location.*

Automatic Cropping Model

After the correct crop location in the original image has been found the next step is to apply machine learning to automatically crop the head from the original image. To accomplish this task, we apply transfer learning using the architectures of two other powerful deep-learning architectures:

- **VGG-16**: A standard deep learning model for image classification. It consists of 5 convolutional blocks, each with 3 to 4 convolutional layers, and makes heavy use of spatial pooling to rapidly increase the receptive field.
- **Deep Gaze II**: A network designed to build a probability distribution over pixels from eye-tracking data points. It utilizes VGG-16 as a pretrained network and appends a read-out network that consists of convolutional blocks that take sampled VGG-16 feature maps as input.

We create our new loss function by maximizing the log-likelihood of the dataset as

$$\mathcal{L} = \frac{1}{N}\sum_i \log p(x_i, y_i, I_i)$$

,
for the pixel locations (x_i, y_i) in the image I_i. This equivalent to minimizing the cross-entropy

$$H(p(x,y), q(x,y)) = -\sum_i \log\left(p(x_i, y_i)\right) q(x_i, y_i)$$

between the pixel locations (x_i, y_i) between the (true) mask q_i and prediction probabilities p_i. Finally, since our output from the network is a log-probability we convert the output to a binary mask by applying a simple thresholding function.

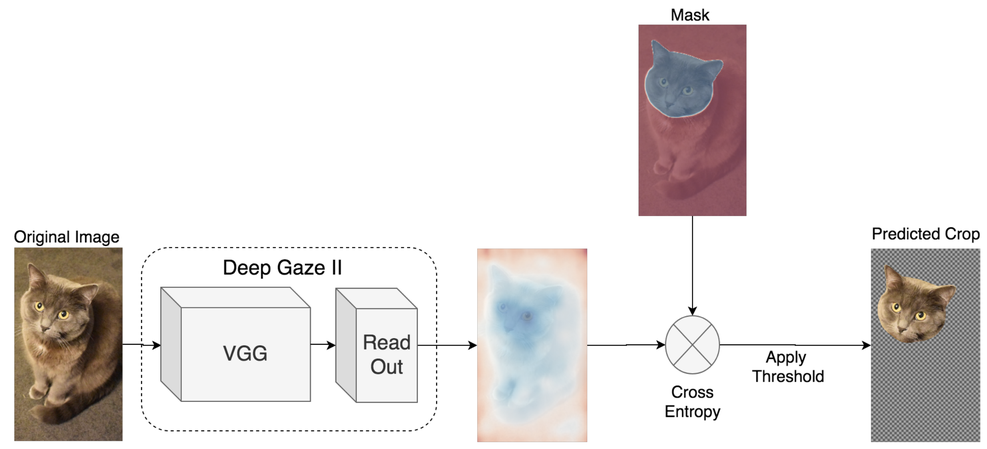

Figure 2: *The original image is first fed into the Deep Gaze II network which returns a log probability of the original image from its prior training. The mask is then used to further guide the network to the area of interest of the correctly cropped head. Finally, a threshold is applied to create a binary mask for the original image*

Results

We use the Deep Gaze II readout network's pre-training but further modify the weights at training. In contrast, the VGG-16 weights are fixed and not further trained. Additionally, we down scale all image/mask pairs to fit into a fixed-size input of 800 x 800 and pad the unused pixels with zeros.

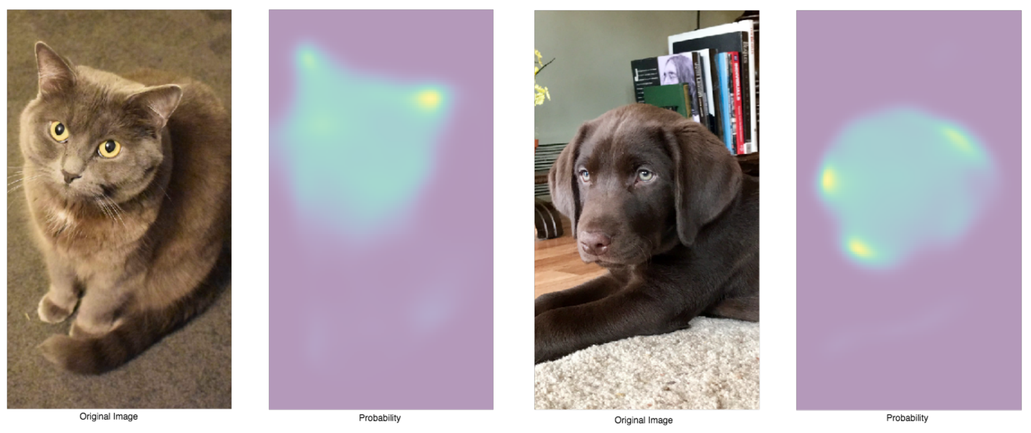

Figure 3: *Examples of original images and their predicted probabilities from the validation set.*

Discussion

- Results are promising for training on sets of dogs and cats jointly
- Using saliency map/eye-tracking data as pre-training could be useful for binary classification image segmentation tasks.
- Transfer learning is a powerful technique for training models on less data.
- The threshold is currently set by hand but could be integrated as part of the learning process.
- The probability output of the image-cropping model provides a substantial amount of information, which could be applied by the image reconciliation task to approximate the optimal scale/rotation to reduce the total number of required convolutions.

References

[1] Matthias Kummerer, Thomas S. A. Wallis and Matthias Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *arxiv*, 1610.01563, 2016.

[2] Karen Simonyan, Michael J. and Andrew Zisserman. Very Deep Convolutional Networks for Large Scale Image Recognition *arxiv*, 1610.01563, 2016.