Background: Footwear and Forensic Analysis

In forensic science, shoe prints and outsole characteristics fall into the category of pattern evidence. When a shoe print is found at a crime scene, there are a number of questions that can be asked. For example, an investigator may want to determine the make or model of the show that made the print, or possibly tie the shoe print to a specific individual. Another question, more difficult than the first two, is how common the shoe type, or features of that shoe type, are in a local population. This question is more difficult than the first two because it relies not only on the information contained in the found print, but also on characteristics of many other types other shoes. Thus, any sufficient answer to this question requires a way to automatically and efficiently classify many different types of shoes within a common system.

Classification

Visual classification is a complex task that our brains have been trained to do very well. Our eyes detect a large variety of features of an object, including color, shape, and texture, and send that information to our brain. Our brain then learns which combinations of features to associate with a given label, and should be able to apply those rules to future objects with similar characteristics. For example, an orange caterpillar and a baby carrot may be of similar color, shape, and size, but one is distinctly more fuzzy than the other. Thus, our brains learn that when faced with a small, cylindrical orange object, texture becomes an important feature when assigning a label to that object (which keeps us from accidentally ingesting caterpillars).

Convolutional Neural Networks (CNNs)

While our brains are adept at parsing images and classifying the objects within them, the task has proved much more difficult for computers. Convolutional neural networks (CNNs) are a tool for supervised deep-learning that have become standard in recent years for automatic image classification. CNNS use combinations of convolutional and pooling hidden layers to filter raw information into features, which are then fed into densely connected layers which are trained to associate given sets of features with their desired labels. This translation-invariant automated classification mimics the human eye-to-brain classification process and has become one of the most widely used machine learning techniques for image classification.

Pre-trained CNNs

I don't love this paragraph. Not sure how much to include, awkward flow. Pre-trained CNNs are CNNs that have been trained on a standard data set. The standard data set comes from ImageNet, a database containing over 14 million images in about 22,000 categories (called "synsets", short for "synonym sets").The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) was established in 2010 as a contest for CNN accuracy on a specific subset of ImageNet. Various CNN structures are tested on about 1.2 million images spanning 1,00 categories. These categories range from natural and man-made objects (e.g., daisy, chainsaw) to living creatures (e.g., ring-tailed lemur, sea lion, and dingo). There are also many categories which require subtle distinctions, such as differentiating between a grass snake and a vine snake. *Something about the

best CNNs for this task are the ones that are famous.* Some of the most well-known pre-trained CNNs include AlexNet, GoogLeNet/Inception, VGG, and ResNet. *Can use just structure and train weights yourself, use fully trained model to reproduce ILSVRC results, or just use pre-trained weights for feature detection*

VGG16 Architecture The main difference between different CNNs is their structure, meaning the number of layers they contain and the pattern those layers are in. In our research, we have tested a few pre-trained CNNs, and we are currently using VGG16. Developed by Oxford's Visual Graphics Group, VGG16 has 16 "functional" (i.e., convolutional and densely connected) layers and 5 max-pooling layers, which function more to alter the structure of the information at each step.

Filters and Convolution Convolutional Neural Networks are named to highlight their use of convolution to extract information from an image. To a computer, an image is stored as a 3-dimensional array with a length and width corresponding to its number of pixels and a depth of 3 to represent the typical RBG color channels. A single convolutional filter is a small array (say 5x5x3) of real valued weights that, when applied over a similarly sized portion of the image, returns a single value representing the strength of presence of a particular feature in that image portion. This application, which is mathematically executed by multiplying the filter values by their corresponding image values and summing, can then be done over the entire image. The result of this process for a single filter is a two-dimensional array of values that represent how

A convolutional layer of a CNN takes a large number of these filters and passes them over the image to