Mapping Features to Labels

Imagine we are teaching a computer how to ‘see’ something. This means that the computer should not only be able to process the image for cosmetic properties like color or shade, but more importantly be able to understand the *contents* of an image.

To simulate this, we could say that the understanding of the contents of the image are effectively blurred...like this:

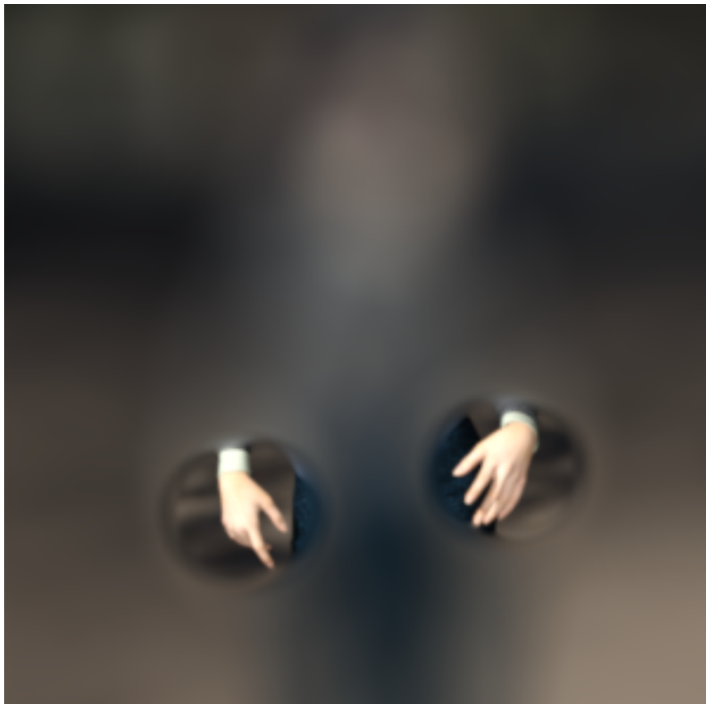
There’s information in there, but we don’t know what it represents, so we simulate that by blurring the image.

Next, imagine that there’s a convolution (filter) that can extract something from the image consistently, and the something that it extracts is *always* present when the image is labeled with

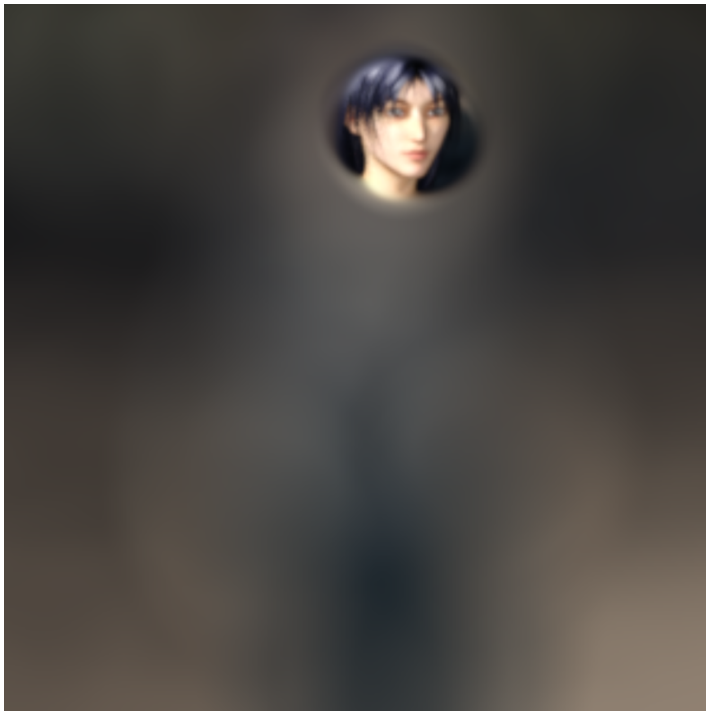
a particular class. For example, consider a filter that always produces something as shown below, when the image is labeled *human*.

We know that these clearly seem to be human legs, but the computer doesn't know that. It just knows that two cylinder-like objects like these tend to show up for a particular filter, and only on images that are labeled human.

Then, similarly, there's another filter that produces an image as shown below, and only for human images:

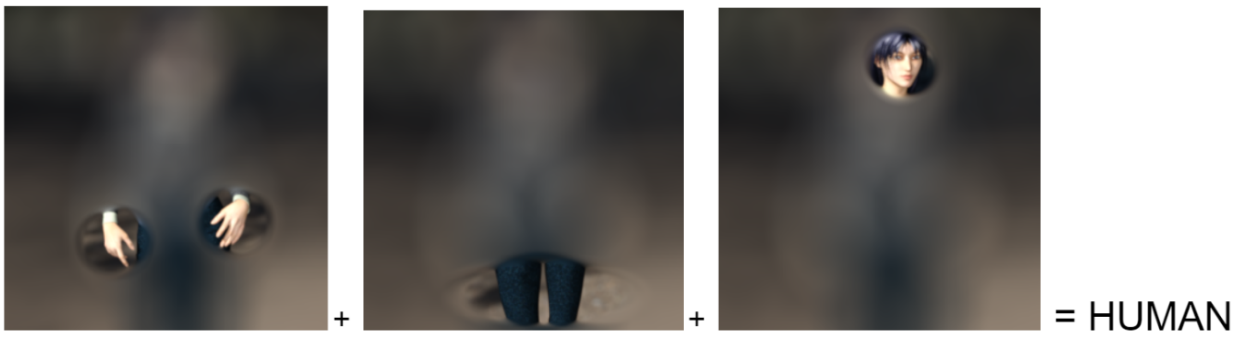


...and then another that produces this, or something similar, again, only for human images:



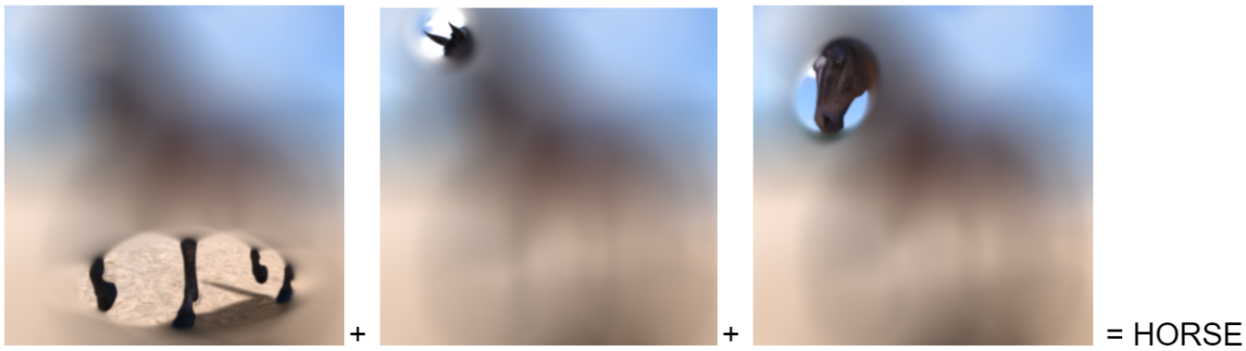
We recognize this as a human face. But the computer doesn't. It only knows that something clear is regularly extracted by that filter, when the image is labeled as human.

Thus, when a set of filters is learned that consistently extracts content as shown in the images above when the image is labeled as human, we could say that the following 'equation' holds:



If there are 64 filters, for example, in the final layer, they may all return 'nothing' and it's just these three that end up having significance for this class.

Similarly, a different set of filters could return values for the label HORSE, and everything else (including the three filters that gave us human hands, legs and face) would return nothing, so we'd get:



Now we have a set of filters that a model has learned. These filters can extract the features that indicate what a horse is and what a human is!

Do note that for this example, we saw features that us humans can easily recognize - hands, legs and feet - and can use them to easily distinguish between humans and horses. The computer, however, is NOT limited to learning such simple features. It might be able to consistently ‘see’ patterns in images that we do not, and that might be a more accurate determinant of the class of the image. The field of convolutional visualization studies this, and it’s fascinating to learn the interpretability of images that are classified using this method!