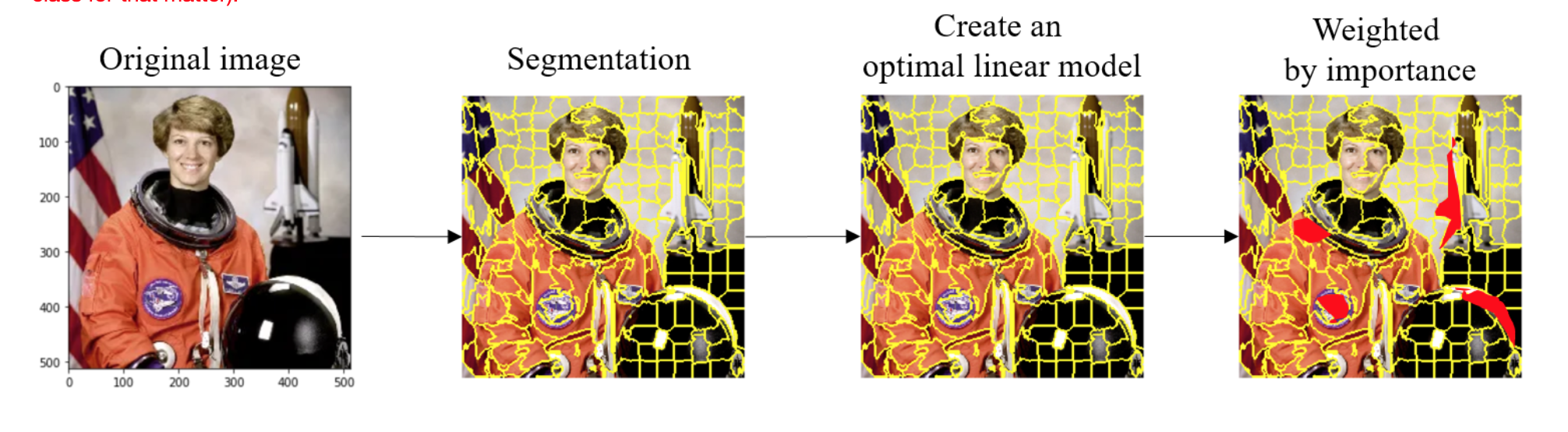
Deep Learning Based Semantic Segmentation to Enhance Local Surrogate Models

Summary

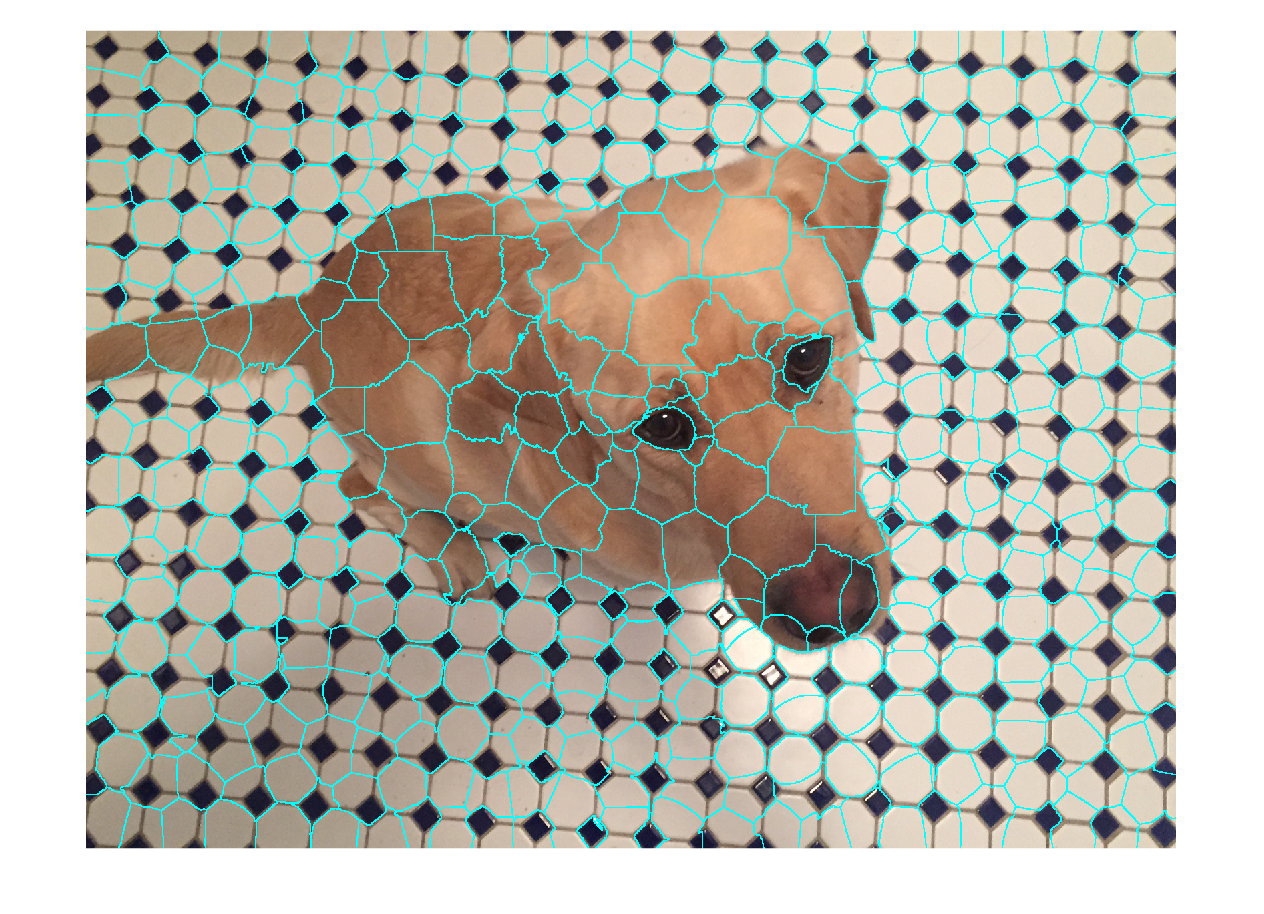
* Image segmentation algorithms used as the first step in surrogate model based model interpretability algorithms don’t create very interpretable segmentation.
* We believe replacing the image segmentation step with deep learning based semantic segmentation can create much more interpretable models.

Using local surrogate models based interpretability algorithms such as LIME on image data work by creating interpretable model-agnostic explanations by first segmenting with image processing algorithms such as slic, watershed, chan\_vese, etc. (see https://scikit-image.org/docs/dev/api/skimage.segmentation.html)

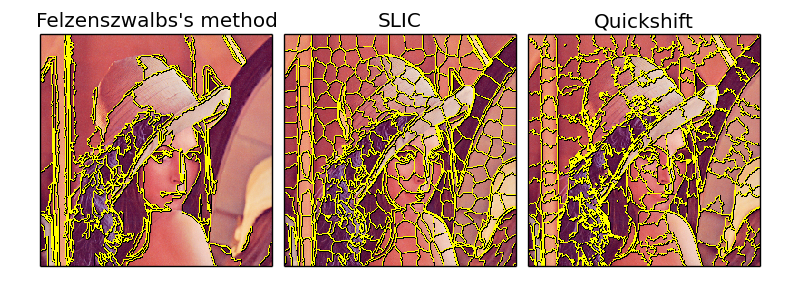
The problem with this kind of image segmentation from a model interpretability standpoint is that even when important predictive features of images are identified they are not mapped to any semantic meaning. In the example below the local surrogate model interpretability LIME, using slic to segment the image, identifies the red regions as important for a making prediction of an astronaut or not. Unfortunately those regions have no semantic meaning. They may be part of the helmet, or the space suit but what one really wants to know is which semantic features are important. Is the space suit or helmet or face or some combination is used to make the prediction?



LIME, using slic to segment the image and identifies the red regions as important for a making prediction.



SLIC based segmentation of a dog picture.



Felzenszwalb, SLIC, and Quickshift based segmentation of a picture.

*Deep Learning Based Semantic Segmentation*

Deep learning based semantic segmentation has been shown to be able segment images into regions much more suited for model interpretability. Rather than segmenting on color and brightness they identify and segment the objects in an image. In the example below, we can see the deep learning based semantic segmentation has identified the regions of the image corresponding to the people, table and TV.



Deep learning based semantic segmentation has identified the regions of the image corresponding actual things.

In this project we propose replacing the image processing algorithms based on color and other statistical properties of images with deep learning based semantic segmentation for the segmentation step in surrogate models based interpretability algorithms such as LIME on images.