Project proposal

Image Understanding

Authors

- Jorge López Rosende
- Rafael Martín Lesmes
- José Vences Rodríguez
- Manuel Rincón Martínez
- Kévin Alberto López Porcheron

What is the problem that you will be investigating?

We are going to analyse the images in the dataset to develop a classifier capable of determinating wether a given image is cancer positive.

What data will you use?

We are going to explore the colorectal_histology dataset from Tensorflow's database of Image Classification datasets.

Describe the data you will use

Tensorflow's colorectal_histology dataset contains 5000 150x150x3 RGB images that can be from one of eight different classes. The total size of the dataset is 246.14 MiB.

Preliminary exploratory analysis of the data

We can obtein some information about the dataset observing the histogram of the images. We discover some classes have similar histograms, for example, **dipose** and **empty** classes. In the same group we observe diferent cell sizes and tissue staining. This properties make the clasification harder.

Metadata

Samples's shapes and types of (raw) data

```
{'image': (150, 150, 3), 'label': (), 'filename': ()}
{'image': tf.uint8, 'label': tf.int64, 'filename': tf.string}
(150, 150, 3)
<dtype: 'uint8'>
```

Number and names of classes

```
8
['tumor', 'stroma', 'complex', 'lympho', 'debris', 'mucosa', 'adipose', 'empty']
```

Dataset visualization