Team

Brice TAYART

Hometown

Paris, France

Background

I recently completed a Master degree in AI and work as data scientist at Tribvn Healthcare, a company specialized in histopathological image analysis. I have an interest in both Computer Vision and the health domain, so I'm happy with that!

Prior to that, I worked for 13 years in the geophysical industry, alternatively attending field operations abroad and developing advanced noise filtering algorithms for seismic records, in Matlab and C. The combination sometime gives surprising results, e.g. write a distributed inversion code, then deploy it on a cluster set-up in a backyard shack in Moroccan countryside.

What motivated you to compete in this challenge?

I competed on behalf of Tribvn Healthcare. Personally, I was very interesting in this challenge, as it was a great opportunity to demonstrate skills, gather experience on an exciting dataset and hopefully benefit to patients some day in the future.

Rest of the team

Even though they did not register an account on Driven Data, other members of Tribvn Healthcare IA team assisted in the project, tested some options and contributed though their experience and advice. The team consists in Capucine Bertrand, Saïma Ben Hadj, Solène Chan Lang and especially Tina Rey who trained some of the models which made into the final ensemble. We also had some discussions with pathologists and medical image analysis specialists.

Summary of the approach

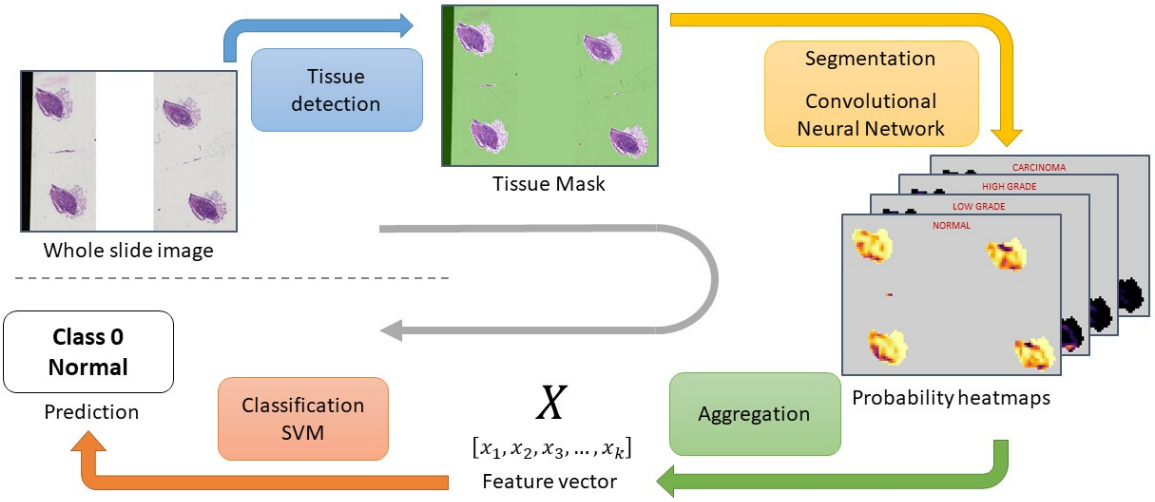

Figure 1: Inference pipeline

The inference process is packed into a parallelized pipeline, so that costly operations that can be deferred to the CPU run in the background while the GPU

is busy with the segmentation – e.g. tissue detection or reading and decoding the tiff files.

Step 1: Tissue detection

With a combination of filters that detect the white background, blurry areas, or some obvious artefacts like black stripes, a mask of the tissue areas is computed. On some slides, these regions of interest (ROI) represents only a very small part of the image.

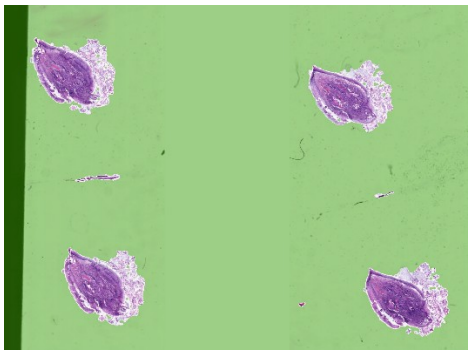

Figure 2: Example of tissue detection

Step 2: Building a dataset and training a DenseNet

The goal is to teach a CNN to classify patches extracted from the whole slide image.

Dataset

A 80/20 train-test split is done on the slides, so that we have 812 slides with 4726 local annotations in the train set and 203 slides with 812 local annotations in the test set. A dataset of labeled images is extracted at a suitable resolution level around the annotations.

This dataset is biased: these images are only representative of what pathologists deems worth of an annotation. It does not contain examples of artifacts, tissue of other nature, etc, all of which should not be classified as cervical lesions. Additional images are extracted from the ROI of slides with a *normal* label – which cannot contain lesions – and added to the dataset (labeled as class 0). This is done both on the training and validation sets (resp. 1069 and 1020 extra images).

Training

A DenseNet 121 is trained on these images, using geometry, color and cutout augmentations.

Calibration

Several sets of parameters give models with an acceptable accuracy, and a cross-plot shows that they do not make errors on the same images. A calibrated ensemble is made, by taking a linear combination of the (pre-softmax) activations of 6 models. The weights and biases are learnt by gradient descent on the training set, all other parameters being frozen.

Step 3: Segmentation and aggregation

The model of step 2 is applied on a sliding window over the ROIs, which gives a segmentation of the tissue. These segmentation maps however contain from less than 10 to several thousand valid data points and are not readily usable.

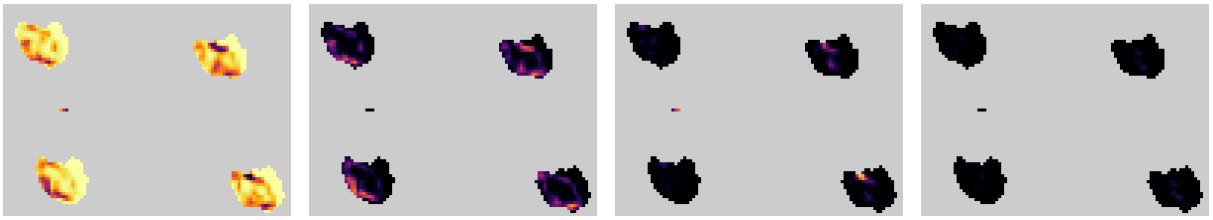

Figure 3:Segmentation heatmaps for each class

For each class, we look at the distribution of probabilities and a small, fixed number of statistics is computed. All the statistics are concatenated into a feature vector that describes the whole slide image.

Step 4: Label prediction

A linear SVM is trained with a K-fold on the validation slides – since the feature vectors on the training set may not be representative of general data due to segmentation model overfitting. The SVM is used to predict the label of the slides as a probability distribution over the 4 classes. Given this distribution and the reward matrix, the average reward expected upon predicting each of the 4 labels is computed. The prediction that gives the highest expected score is picked.