Deep Learning Techniques for Classification of Orthodontic Photos into the Correct Treatment

Leoni Paolo 1894985
Morosini Francesco 1875742
Protani Andrea 1860126

Sottile Alessandro 1873637
Tarantino Ramona 2082006

Abstract

The objective of this study is to explore Machine and Deep Learning techniques for the classification of orthodontic images according to the correct dental alignment treatment. The aligner company Sorridi will provide photos taken by dentists to patients for whom a photo pre-evaluation has been requested. This process is used to determine the necessary treatment and therefore have an estimate of duration and cost and is currently manual, the goal will therefore be to automate it by developing a machine learning model that will be able to correctly assign the treatment. The model will be trained and tested on the images of about 2500 patients collected from 2018 to date, for each patient a variable number of photos of different nature may appear (full frontal mouth, right lateral, left lateral, upper arch only, lower arch only ...). The types of treatment performed, "Soft", "Semi-Medium", "Medium", "Semi-Hard" and "Hard" are determined by various factors such as the positioning, overlap and rotation of the teeth. Different approaches will be attempted based both on a preprocessing of the images that can even include the segmentation of the teeth only to give more emphasis to the dental structure, and on leaving the images "raw" as other factors such as gums health can also play an important role in determining the responsiveness of the teeth to the alignment.

1 Main research aim

In recent years, various machine learning models have been applied to different fields/sectors and from our point of view, they can also work well in the dental sector. We are extremely interested in carrying out this type of project as it represents an application of these technologies to real data and could be of great benefit to the owner of the company that provided us with the data. The main objective of our project is, through the use of ML, to classify with good accuracy the right types of treatment for patients in the various dental clinics.
Since this is only a pre-evaluation, we want the model to output two classes, those it deems most probable and then we want to constrain it to propose only adjacent classes (for example Semi-Medium/Medium).

2 Data source & Collection

To collect the images needed to train the model we collaborated with Sorridi which provided us with access to their website from which we were able to download the various images associated with each patient.

(a) Front image (b) Right side image (c) Left side image

(d) Upper internal image (e) Lower internal image

Figure 1: Example of several images relating to a patient.

In these cases it is necessary to maintain anonymity for the patients, so to each of them we assigned a unique identifier.

Many of the patients were discarded due to too low image quality, the final dataset consists of images from approximately 2500 different patients.

Problem 1: not all patients for whom a pre-evaluation was done then completed treatment. For this type of patient there wasn't therefore information relating to the treatment that had actually been carried out but only the proposal of the double class made by the technicians. With their help we therefore chose only one of these two classes to use as a label.

Problem 2: the photos are quite heterogeneous since they are taken by different dentists without precise guidelines. To reduce the variability, different datasets were used, some containing images with all angles and some containing only frontal images. The downside is that in this way the size of the dataset which was already not very large was further reduced.

Problem 3: there are five treatments available:

- Soft
- Semi-Medium
- Medium
- Semi-Hard
- Hard

The classes are not all equally represented, in particular in the "Soft" class there are many fewer patients. The most present class is the "Medium".

Photos taken from different angles for a patient are shown in Figure 1.

3 Model & Methods

For the entire project, Python was used as the programming language, both PyTorch and TensorFlow were used as frameworks.

After making initial attempts based on simple methods such as Histograms of Gradients (HoG) and Histograms of colors, we decided to move on to deep learning techniques. A number of different models and approaches have been tried, most notably CNNs and Visual Transormers.

The models that gave the worst results were the models created from scratch. Given the small size of the dataset, they were not able to learn which were the right features to extract from the images, thus ending up in many cases assigning the same class to all of the pictures. The models that gave the best results were those based on fine-tuning of large models. To perform the fine-tuning, models available on the Hugging Face platform were used.

Techniques as augmentation, segmentation and ensemble of the various angles were also tried but these did not lead to improvements in performance.

Finally, a graphical interface was created using the Gradio Python package with which makes possible to simulate the photographic pre-evaluation process (limited to frontal images only).

4 Results

The simple "Accuracy" was used to evaluate all created models. To have comparable results between the different models, a train set and a test set with proportions 85-15 were created a priori, the models were also trained with the same number of epochs: 30.

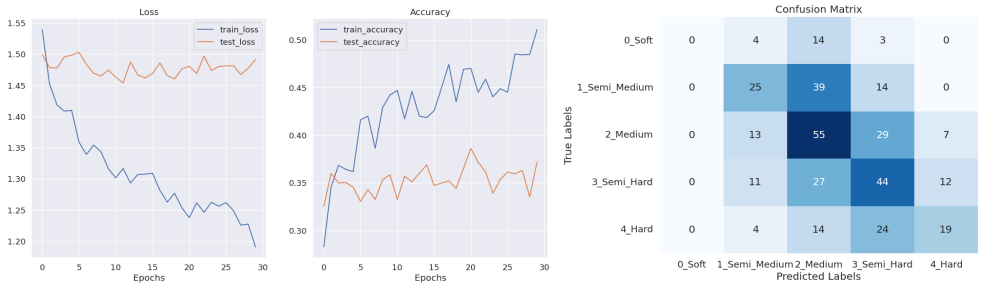
Model	Single	Double	Adjacent
CNN	32.40	58.00	60.00
CNN + augmentation	32.12	57.50	57.50
CNN + early exit	31	48.3	55
CNN + early exit + augmentation	28.5	54	53.8
MobilenetV2	36.60	63.00	59.00
VGG16	36.03	49.72	49.72
VGG19	35.20	51.67	50.27
DenseNet121	35.47	50.27	50.55
Inception	36.60	50.11	50.81
NASnet5	37.06	50.35	50.38
ViT	28.85	55.22	-
PyTorch - VIT_B_16	37.36	62.36	-
Google - vit-base-patch16-224	39.39	62.85	52.23
Swin Transformer	28.85	55.22	-
Microsoft - swin-tiny-patch4-window7-224	39.94	68.16	53.35
Facebook - convnextv2-tiny-1k-224	39.94	68.72	53.63
ViT + augmentation + all angles ensembled	29.00	53.00	-
Human technicians	-	-	≈ 65

As can be seen from the table the results are not optimal. The accuracy levels for single predictions are very low and only reach "decent" levels with double predictions. However, the last row of the table shows the estimated accuracy value of the Sorridi company's technicians in carrying out the pre-evaluations, being around 65% (for the adjacent double prediction) makes it clear how complex this task is. Some of our models manage to exceed the previous value but only for the double prediction without constraints (e.g. Soft/Hard) which in real contexts makes no sense, when we restrict the choice to adjacent pairs the results do not reach the target value.

Correct prediction is very difficult due to several factors:

- limited number of images
- poor image quality
- heterogeneity of images within the same class
- labeling of many photos inaccurate
- particular requests from doctors (for example, even if the aligners should be worn in both arches, they explicitly require only one to be worn)

Now wanting to investigate the results of our models, we can refer to Figure 2, it reports the learning curves of a model and the confusion matrix of the results (they were generally similar for all models). The curves show how the models are unable to learn as the epochs pass, in fact the loss and accuracy



(a) Learning curves

(b) Confusion matrix

Figure 2: Analysis of learning and results.

only improve on the train set while they remain fairly constant on the test. The confusion matrix instead shows how the models have particular difficulty in recognizing images of the "Soft" class as this is the least represented in the dataset. Furthermore, for some of the models the confusion matrices showed that the models (those from scratch) assigned all the images to only one class, the "Medium".

5 Conclusions

At the end of the work the results obtained were not entirely satisfactory, the task is very complicated but the performances obtained are not the best. To address this problem appropriately, it would be necessary to carry out a more controlled and large-scale data collection. The advice given to the company is to give more precise guidelines to dentists for taking and uploading images so that in the future we can retrain the models with more data and of higher quality. A reflection to make is that even human technicians have difficulty carrying out the pre-evaluation based on simple photos, they correctly predict only around 65% of cases. For future work, it would be more important to implement different image augmentation and preprocessing techniques as the different models tested showed similar characteristics so it makes less sense to continue focusing on that.