Report: Project 3

Machine Learning for Health Care

Remo Kellenberger, Michael Mazourik, Julian Neff

29.03.2022

1 Introduction

Brain tumors account for the majority of central nervous system tumors [Bhu+20]. Moreover, the 5 year survival rate post diagnosis is low being 27% for men 33% for women in Switzerland [22].

Developing low cost and reliable method to diagnosis early stage developments of these tumors could further increase the survival rates. Furthermore, although MRIs have not been declared for absolute safely from both a medical and psychological perspective, it evident that it is is relatively safe and holds great potential to for analysing the state of the brain [Mar+07].

With recent developments in machine learning and computer vision, the possibility of high quality brain picture analysis is back on the table. In this research project, we aim to analyse the dataset from three different perspective. Firstly, using a random forest classifier which looks at texture, shape and first-order features. Then, an convolutional neural network based on features originating from cooperative game theory. And finally two models with high interpret-ability which are going to be a bag of words (image vocabulary) classifier and the scale invariant feature transform (SIFT) algorithm.

2 Dataset

The dataset is composed of MRI pictures in the png format. The testing and validation sets contain 28 images each and the training set is composed 222 images. Furthermore, the images are transformed into tensors of shape $3 \times 128 \times 128$. The tensor representation indicates a square image with 128 pixels of height and width with 3 channels for the red, green and blue colors.

In the training dataset, 128 out of 222 datapoints with associated images represent positive cases. There are 16 out of 28 positive and 18 out of 28 positive with brain tumor identification in the test dataset, and validation set respectively. This results in the positive percentage to be 57%, 64% and 58% for the test, valid and training set.

3 Task 1: Baseline Random Forest

As a baseline model we made use of the random forest classifier from sklearn. This classifier constructs multiple decision trees during training that then operates as an ensemble. The trees are individually different because each tree creates different training data by bagging (Bootstrap Aggregation). Bagging means that each decision tree is created by a subset from the training data (with likely multiple occurrences). This makes the trees learn different features and reduces overfitting. For the prediction of the forest classifier each individual (decision) tree outputs a prediction and the random forest classifier then takes the class which appears in the majority of predictions.

Using this baseline model on the pyradiomics dataset we achieved an accuracy of **78.6%**.

4 Task 2: CNN Baseline

With the deployment of machine learning systems to healthcare settings, new requirements are being placed on this systems. A ML model does not only need to achieve high accuracy scores, it is also measured by its interpretability. Interpretability means the ability of a model to explain the reasoning behind its results. If a model is able to explain its reasoning, this reasoning could then be verified by human experts ultimately resulting in bigger trust in these models. In this task, we evaluate the interpretability of a simple ML model in a typical application in healthcare. The objective of the model is to detect brain tumors in MRI images. The interpretability of the model is then evaluated by examining its shapley values.

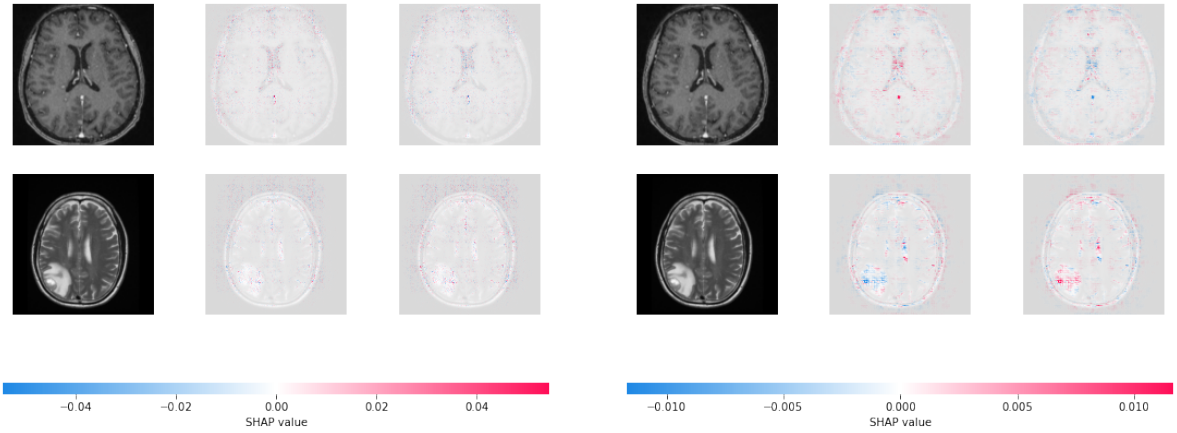
4.1 Results

All experiments have been run for 200 epochs with the dataset split into train (80%), validation (10%) and test (10%) sets. In order to counter the small amount of available data, some basic image augmentation (rotation and flipping) was performed. The results in table 1 show, that all models perform quite similar. A deeper look at the evolution of the models during training shows, that there are still some differences between these models. In the next section, we take a look at the interpretation of the models.

Model	no augmentation	simple augmentation
BaselineCNN	0.96	1.00
Resnet18	0.96	0.92

Table 1: Accuracy of Baseline CNN for Brain Tumor Detection

4.2 Interpretation



(a) No image augmentation

(b) With image augmentation (rotation, flipping)

Figure 1: Shap values of BaselineCNN for brain tumor detection. The first row contains an MRI scan without brain tumour whereas the scan in the second row is classified as brain tumor.

A shap value describes the impact of a specific feature. In other words, a higher shap value means, that this feature in comparison to its baseline value contributes more to the final prediction. In our case, the feautures are the pixels of the MRI scans. Thus we expect the area of the tumor to have positive shap values for the class 1 (= brain tumor) and negative shap values for class 0 (= no brain tumor). Figure 1 shows, that in the BaselineCNN the area of the tumor is correctly classified as feature with high impact on the outcome of the model. Despite having a high accuracy, the impact of the area with the tumor is less important in the BaselineCNN without image augmentation.

5 Task 3: Additional Models

5.1 Bag of Words (BoW) Classifier

Although bag of word classifiers are traditionally implemented for natural language processing tasks, they can also be implemented to represent images using a visual vocabulary which was inspired and adapted from the computer vision and pattern recognition class at ETHZ. In this approach, the image is chunked using a uniform grid. Then a Sodel filter is applied on the x and y axis to deliver an approximate orientation for the given pixels. The benefit of using a filter for gradient estimation is the relative low computational power. Then, the gradients are classified into an 8-bin histogram called the histogram of oriented gradients. Once this has been done with all the images, a k-means clustering algorithm is applied which generates the cluster means representing the words of the algorithm.

Then the classifier finds the nearest neighbour in the training set based on the visual words which is then used as a proxy for the given label (with or without tumor). The final accuracy is estimated at 89.28 % with 100 % accuracy on the negative cases and 81.25 % on the positive cases which is not ideal in our case. In future case studies, it would be important to focus on the recall as detecting early stage disease developments although we also would like to avoid misdiagnosis of true negatives.

5.2 Scale Invariant Feature Transform (SIFT) Algorithm

Scale Invariant Feature Transform (SIFT) is a feature detection algorithm in Computer Vision [Low99; Low04]. SIFT helps to locate the local features/keypoints in an image. These points are scale and rotation invariant which is the major advantage of SIFT. This makes it in our opinion interesting for tumor detection, because we hope that there are keypoints in brain slices with a tumor which do not appear in a brain slice without tumor. As a direct comparison we will use the model in task 2 for classification. This gives us a good comparison to see if the SIFT algorithm is able to learn relevant keypoints in images with tumors.

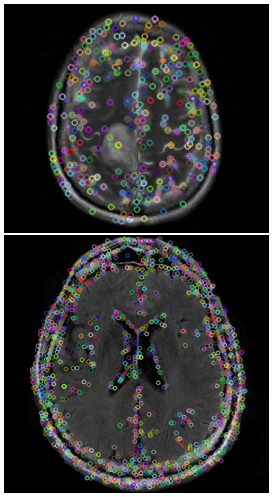


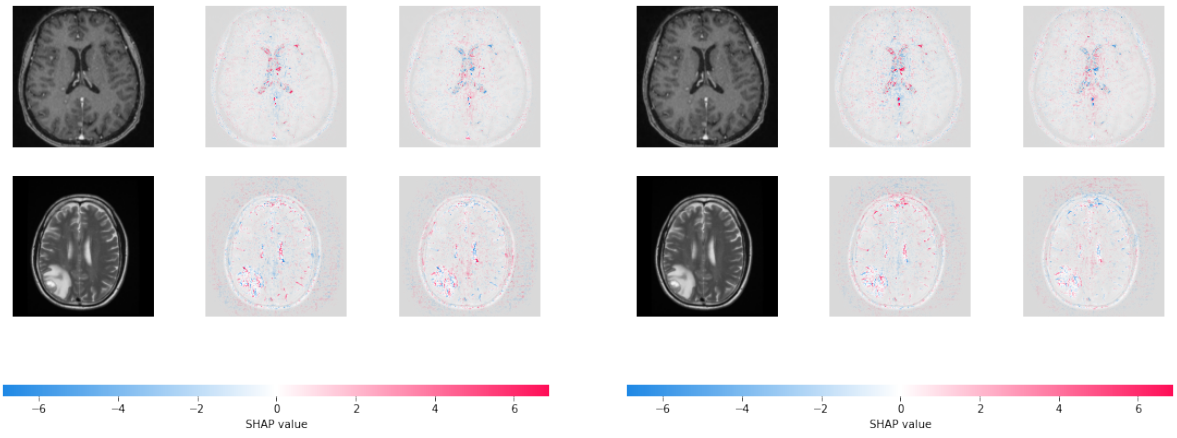
Figure 2: Keypoints of a brain slice with (above) and without (below) a tumor.

We had to make two small change in the model from task 2. Namely we now have a one-layer input instead of three and we also excluded the transformations (apart from flipping) from the Dataset because SIFT is already shift- and scale-invariant. Apart from this the whole model structure and the parameters we used were exactly the same.

Using this approach we were able to get an accuracy of **95.6%**. This is a significant upgrade over the results from the baseline and the model succesfully made use of the keypoints.

5.3 Transfer Learning with Shapley Values

Machine learning has widely been used for image tasks such as classification or captioning. Researchers came up with a variety of complex and good performing models. Training machine learning models with images often leads to complications. Sometimes, it is expensive to get enough samples to train a model. This is the part where transfer learning comes into play. The idea is to use a pretrained model to extract reasonable features from the images and then train a custom network on these features. In our case, we decided to use Resnet18 as a feature extractor and a fully connected layer to obtain logits from the feature embeddings. As we can see in table 1, the model performs quite good. Nevertheless, it fails to identity the area of the tumor as an important feature for the final classification result as can be seen in figure 3.



(a) No image augmentation

(b) With image augmentation (rotation, flipping)

Figure 3: Shap values of Resnet18 for brain tumor detection

6 Conclusion

For this project, we compare various different methods to evaluate brain tumor detection using the MRI dataset. Starting with the baseline random forest classifier we achieved an accuracy of 78.6%. This classifier is very intuitive as it works on multiple decision trees, however it lacks in accuracy. We further used a CNN model where we achieved an accuracy of 96%. We can interpret each classification by looking at what each cell focuses on. In the bag of words approach which classifies the nearest neighbor in the training dataset achieves 89.3 %. Finally, using another computer vision algorithm achieved SIFT 95.6% while remaining explainable with visual features. This is convenient as we can look at the datapoints in each layer and see which are relevant for classification.

We have seen that high accuracy scores alone are not enough. Some applications call for another metric to evaluate the performance of a model. In task 2, we investigated the interpretability of a baseline cnn model for the classification of brain tumor. The experiments have shown, that the BaselineCNN model with basic image augmentation applied can be well interpreted with shap values.

7 Appendix

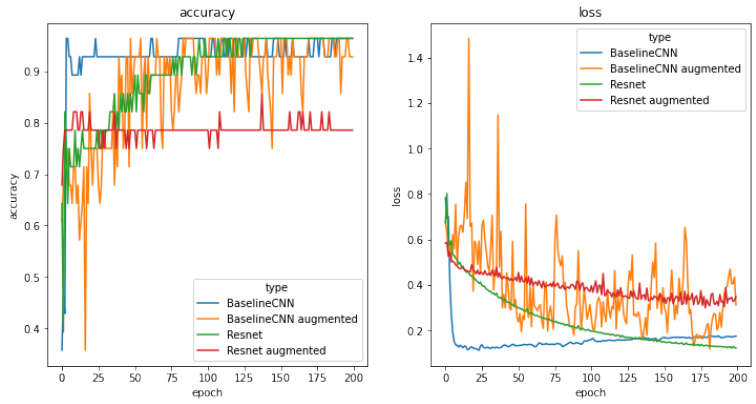


Figure 4: Validation loss and accuracy for brain tumor detection

References

[22] 2022. URL: https://www.liguecancer.ch/a-propos-du-cancer/les-chiffres-du-cancer/-dl-/fileadmin/downloads/sheets/chiffres-le-cancer-en-suisse.pdf.

[Bhu+20] S. Bhuvaji, A. Kadam, P. Bhumkar, S. Dedge, and S. Kanchan. "Brain tumor classification (mri)". *Kaggle, doi* 10 (2020).

[Low04] D. G. Lowe. "Distinctive image features from scale-invariant key points." *International Journal of Computer Vision 60(2)* (2004), pp. 91–110.

[Low99] D. G. Lowe. "Object recognition from local scale-invariant features." *Proc. 7th International Conference on Computer Vision (ICCV'99)* (1999), pp. 1150–1157.

[Mar+07] J. Marshall, T. Martin, J. Downie, and K. Malisza. "A comprehensive analysis of MRI research risks: in support of full disclosure". *Canadian journal of neurological sciences* 34.1 (2007), pp. 11–17.