Medical Imaging - Kaggle Challenge

Daniel Amar & Avigail Zerrad (MASH team)

March 2021

Introduction

We present in this report our work for the Kaggle challenge of Deep learning for medical imaging course.

In computer vision, a very common task is to use labelled data to classify images. There exists a lot of architectures and models that perform very well for this purpose. In some cases however, this approach may be inappropriate because of the nature of the problem we want to solve. In our problem, we have bags of images for each patient. The labelling is at the patient level. Then, we have no access to the labels of the instances (images of patients). This setting is known as Multi-Instance Learning.

Let's now explain in detail the context of the data. Lymphocytosis is an increase in the number or proportion of lymphocytes in the blood. It can occur after an illness and be without risks, but it might represent something more serious, such as a blood cancer. The diagnosis is made on the basis of visual microscopic examination of the blood cells, together with the integration of clinical attributes (age, lymphocyte count...). Some clinical tests are required to validate the diagnosis, like flow cytometry.

The goal of this challenge is to provide a MIL framework which will predict with high accuracy if a new patient has Lymphocytosis. It would be very useful for two reasons :

- First, it automates the analysis of blood cells of the patients.
- Second, it is a good way to determine which patient should be referred for additional analysis.

Using the bags of images of 163 patients together with some clinical attributes like age and lymphocyte count, we tried to propose a coherent approach to solve this challenge. In a first part, we will explain the MIL framework we used, its architecture and some interesting experiments. Then, we will focus on model tuning and on the global approach we propose for the challenge.

1 Architecture and methodological components

1.1 Architecture

After some research about Multi-Instance Learning, we decided to use the approach proposed in the paper Attention-based Deep Multiple Instance Learning by Ilse et al. The authors propose an Attention-based MIL pooling that we will present.

The idea is to use an attention mechanism. If we take $H = h_1, ..., h_K$ a bag of K embeddings, the MIL pooling is defined as :

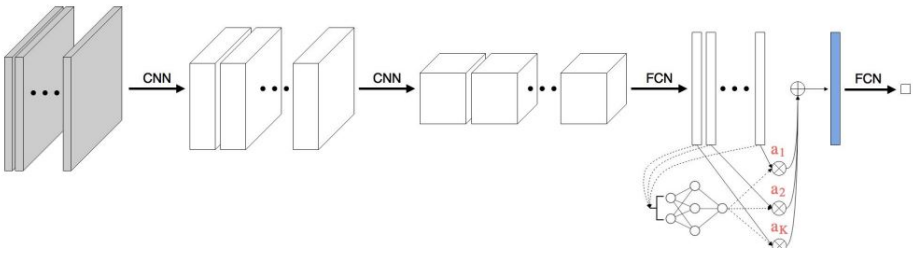
$$\mathbf{z} = \sum_{k=1}^{K} a_k \mathbf{h}_k \text{ where } a_k = \frac{\exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_k^\top)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_j^\top)\}}.$$

The weights a_k are determined by a two-layered neural network. The idea is that key instances (images labelled at 1) will have a higher weight. The embeddings are obtained after two layers of convolution and two fully connected layers. We can visualize the complete architecture in the following figure :

Layer	Type
1	conv(4,1,0)-36 + ReLU
2	maxpool(2,2)
3	conv(3,1,0)-48 + ReLU
4	maxpool(2,2)
5	fc-512 + ReLU
6	dropout
7	fc-512 + ReLU
8	dropout
9	mil-max/mil-mean/mil-attention-128
10	fc-1 + sigm

Figure 1: Architecture

This architecture was proposed by Sirinukunwattana et al. and used for the experiments on histopathology in the paper of Ilse et al.



The convolutional layers allow to extract features from the images. Then, the MIL pooling gives a weight to the obtained embeddings and outputs a bag score. Finally, the sigmoid layer allows the attention activation.

1.2 Class imbalance

1.2.1 Loss function

The first problem we had to solve was the class imbalance. Indeed, the provided train dataset contains approximately 70% of sick people. To address this issue, we decided to modify the loss so that the weight that the model gives to an error depends on the class distribution. In other words, the model penalizes more the mistakes on the class 0 (i.e false positives) than on the class 1. For that purpose, we modified the binary crossentropy loss using weights to balance the two classes.

1.2.2 Balanced accuracy

In order to avoid interpretation mistakes due to class imbalance, we used balanced accuracy defined as : $\frac{1}{2}$ (sensitivity + specificity).

1.3 Experiments

Using the MIL pooling described above, we first obtained scores between 0.496 and 0.58 on Kaggle. This improvement was due to the optimization of some hyperparameters but we felt that it was not sufficient to significantly improve our score.

In their paper, the authors mentioned a gated attention mechanism. The idea was to add a learnable non-linearity to the weights a_k as follows :

$$a_k = \frac{\exp\left\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\mathbf{h}_k^\top\right) \odot sigm\left(\mathbf{U}\mathbf{h}_k^\top\right)\right)\right\}}{\sum_{j=1}^{K} \exp\left\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\mathbf{h}_j^\top\right) \odot sigm\left(\mathbf{U}\mathbf{h}_j^\top\right)\right)\right\}} \tag{1}$$

This modification had a great impact on our results and we obtained a score of 0.789. We decided to use this submission for the final score on Kaggle. The reason that can explain this improvement is that in the previous definition of the weights, the tanh non-linearity is approximately linear for x in [-1, 1]. The introduction of the sigmoid non-linearity then allowed to avoid this 'linearity'.

2 Model tuning and comparison

2.1 Preprocessing

One great limitation we had to train the model was the RAM. The authors of the paper obtained great results by training their model on 100 epochs but we could not do it. In order to reduce the memory, we resized the images from 128 x 128 to 64 x 64, but we were still limited.

2.2 Validation procedure

For the validation procedure, the authors of the paper used a 10-fold cross validation. After a lots of runs, and due to our limitation in terms of memory, we found that a 4-fold cross validation with 40 epochs led to good results. For the split between train and validation sets, we set at 0.8 the proportion of train data.

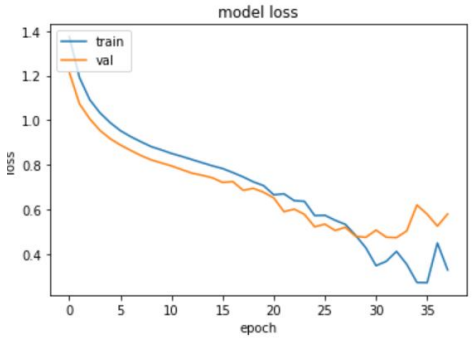
2.3 Hyperparameters

In order to obtain the best results depsite the limitation of RAM, we tried to optimize some hyperparameters, like the number of folds in the validation procedure. For the early stopping, we set the patience at 10 since we do only 40 epochs per fold. We did not change hyperparameters such as the learning rate, the weight decay and the momentum (lr = 1e-4, weight_decay = 5e-4, momentum = 0.9) because modifying them led to bad results.

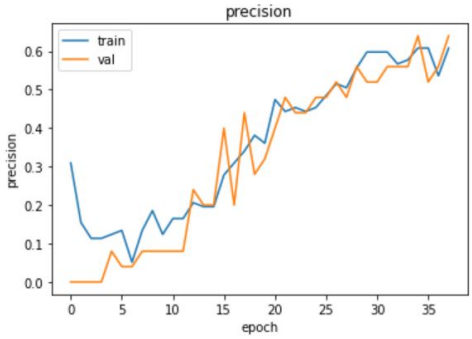
2.4 Metrics & Evaluation

Since the data is very imbalanced, with a proportion of 70% of sick people, we adapted the loss. It had a real impact on the results since before this modification, the network was predicting only 1 on the test set. In order to rigorously prove it, we used the precision metric. A high precision means a low false positive rate.

Let's take a look to some graphs :



Loss



Precision

These two figures come from the same fold. We can see that the loss decreases while the precision increases, both for training and validation data.

Balanced accuracy was also used to evaluate the accuracy without having a wrong interpretation due to the class distribution.

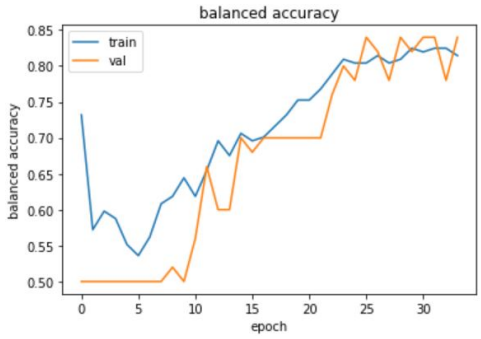
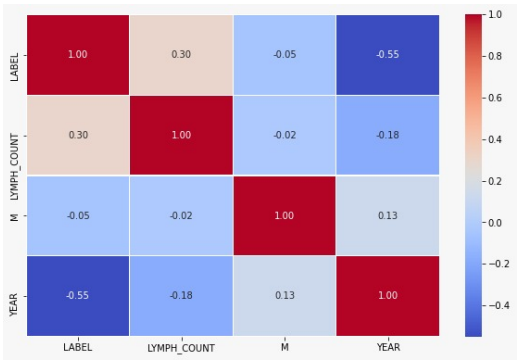


Figure 3: Balanced accuracy

2.5 Use of clinical data

To have an idea about the importance of the clinical data, we took a look to the correlation matrix.



We can see that the variables 'YEAR' ('DOB') and 'LYMPH_COUNT' are highly correlated with the label of the patient. By curiosity, we performed a logistic regression with these two variables and submitted the results on Kaggle (obviously, we didn't choose this submission for the final score). The obtained score was surprising (0.83) and we had a long debate about how to use these data.

We then had the occasion to ask to a doctor in what way these attributes are used. He answered us that in that case, and for a lot of other diseases, these attributes were considered after the analysis of the images in order to know if the diagnosis makes sense.

That's why, it seemed to us more coherent to use this clinical data outside (and after) the MIL. Another reason for that is that if we include these attributes in the MIL network, we wouldn't know to what extent the images will have a weight in the final prediction.

We then thought about a way to introduce the clinical attributes in a second step. We propose the following scheme :

1. MIL with the gated attention mechanism in a first step.
2. Logistic regression only for the patients for whom the network predicted the label with a probability between 0.4 and 0.6. (We first tried 0.45 and 0.55 but the score did not improve).

Conclusion

As a conclusion, this task was a very interesting and challenging work. We were quite satisfied with the results we obtained despite the limitations about the RAM. We can naturally think that a better memory would lead to better results.

The model proposed in the paper we used was quite efficient for this task. We particularly noticed the great impact of the gated attention mechanism, together with the modification of the loss. These two elements allowed us to perform quite well for this challenge. The use of appropriate metrics like balanced accuracy and precision allowed us to objectively validate the modifications and improvements we made on the model.

Moreover, we saw that clinical data have an important part in the diagnosis of Lymphocytosis. That's why, we proposed to use some attributes in a second step.