Deep Learning Medical Imaging : Kaggle Challenge report
Team: IBS/ABM

Iheb BEN SALEM
ENS PARIS SACLAY
iheb.bensalem@telecom-paris.fr

Anis Yassine BEN MABROUK
ENS PARIS SACLAY
anis.benmabrouk@telecom-paris.fr

Abstract

The problem consists of trying to distinguish lymphoproliferative disorder cases from common cases of Lymphocytosis that are reactive through automatic and accurate methods based on various blood smears and patient attributes. A common theme in the listed approaches is the usage of Multi-instance-Learning (due to the nature of the blood smears)

1. Introduction

In current-day clinical practice, diagnosis of Lymphocytosis is either reactive or tumoral.this diagnosis relies usually on visual microscopic examination of the blood cells (texture and size of the lymphocytes) while also taking into account clinical attributes such as age and lymphocyte count which leads to determining the subtype of lymphoid malignancy. This method, despite being fast and affordable, is not easily reproducible. Hence the interest in deep learning assisted methods for conducting said diagnosis to overcome those shortcomings.

2. Data

The dataset can be seperated into two parts .For each patient,we have a folder that contains different images of blood smear.The number of these images varies from a person to an other .The second part of data is in the csv files in the form attributes : date of birth,gender and lymphocyte count.Visualizing the distribution of data across the labels (see figure 1), our early hypothesis is that the gender most likely have no positive impact on the predictions.We confirmed this in our models later.

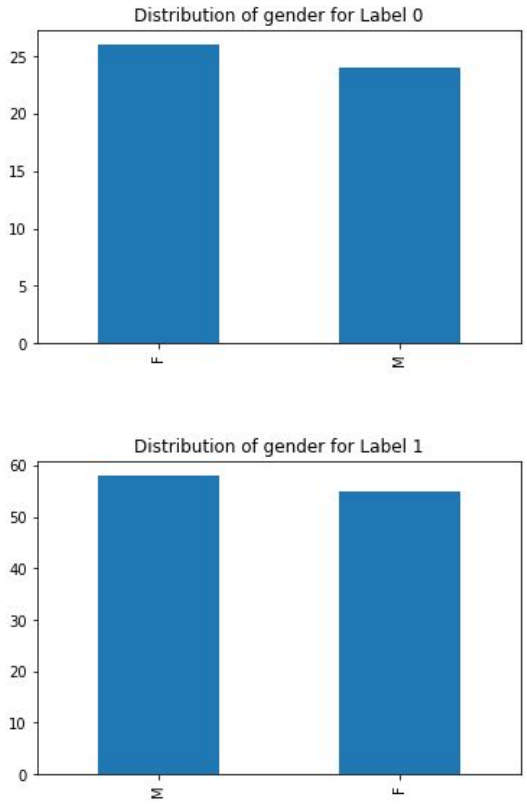

Figure 1. Distribution of gender across the labels

3. Architecture and methodological components

3.1. Attention-based Deep Learning

3.1.1 Multiple Instance Learning based:

The MIL approach is a variation of supervised learning where a single class label is assigned to a bag of instances. The main goal is to predict a bag label (diagnosis) while coping with the ambiguity of not knowing which

of the patterns in a positive bag are the actual positive examples(key examples) and which ones are not. the bag label is modeled using the Bernoulli distribution and the training is done by optimizing the log-likelihood function. Individual labels are presumed unknown. the architecture used is the one mentioned in the article [1].It also exploits the Interpretability and learnability of the attention based MIL pooling to achieve better results since usually, In order to solve the primary task of a bag classification, methods like utilizing similarities among bags, embeddings instances to a compact low-dimensional representation combined with a classifier or combining responses of an instance-level classifier were not enough due to their lack of interpretability. The attention-based MIL pooling bridges the instance-level approach and the embedding-level approach. The method also proposes the usage of a "Gated attention mechanism"[3] to allow for a better description and learning for complex relations since it introduces a learnable non-linearity that removes expressiveness problems related to the linearity in the tanh function as follows:

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

where $\mathbf{U}$ are parameters , $\odot$ is an element wise multiplication and sigm is the sigmoid non linearity.

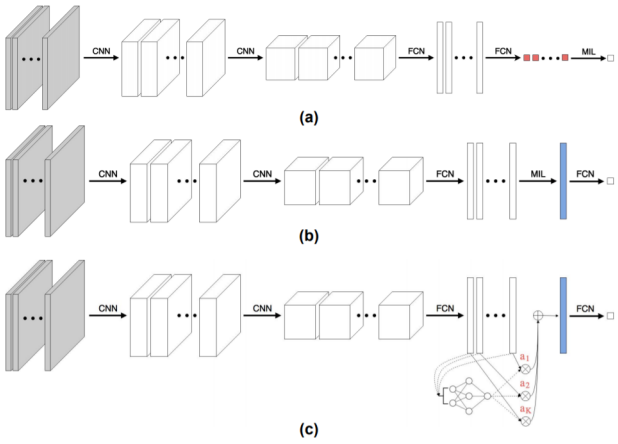

Figure 2. MIL approaches

The attention based MIL pooling mentioned above allows the assignment of different weights to instances within a bag and hence the final representation of the bag could be highly informative for the bag-level classifier (which would lead to a better discovery of the key instances). The general architecture [4] consists of a feature extractor (split into 2 parts with the first part consisting of 2 convolutional layers each coupled with a Relu activation function and a max-pooling layer and the second part being a simple fully connected layer with also a Relu activation function), an attention mechanism (based on the Tanh) and a classifier. for this approach , the bags given needed to be of the same size hence we proceeded to re-sample the images per patient to keep around 50 per patient (we increased to 100 later on since we'd have less of a chance of removing the key instances)

3.1.2 Classical supervised learning:

Due to the need to have "balanced" bag sizes (which wasn't the case), we decided to modify the attention method to learn on each image separately (through looping on all images per patient) and using the bag label as individual labels and then combining the result by taking the mean of the obtained probabilities to have a final prediction for the bag, we preserved the rest of the architecture as the one mentioned above.

3.2. XGB

XGBoost is library that contains decision-tree-based ensemble Machine Learning algorithms. They are based on the idea of using gradient boosting with weak learners being trees, For simple data classification task,we learnt from our previous challenges/experiences that it has very good performances.So, it was one of the earliest algorithm that we tried. We used the patient attributes (year of birth and LYMPH COUNT) as input .

3.3. ResNet and SVM MIL

In our search about multi-instance approaches,we found about SVM MIL [5] .We created a pipeline that enable us to use both images and attributes in the model. First,we use pretrained resnet50 to extract features from each bag of images. We chose resnet since it seemed like it had a potency for properly distinguishing features linked to main aspects in images (edges and borders and outlines etc). We ended up with 2048 features.We add the attributes to the feature vectors and we use Principal component analysis (PCA) to get rid of noisy features and avoid overfitting . We retain 95% of the variance information (correspond to the first 159 component). Finally we can use SVM-MIL (this implementation [6]) on this data.This method is also a multiple-instance learning approach based on two alternative generalizations of the maximum margin idea (maximum Pattern and maximum Bag) used in SVM classification.
Followed by local optimization heuristics(alternating between, for a given integer variable solving the associated QP and finding the optimal discriminant function and for a given discriminant, updating one, several, or all integer variables in a way that minimizes the objective), the method

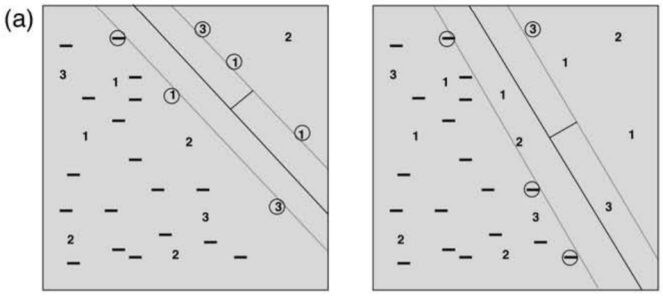

Figure 3. mi-SVM solution and MI-SVM solution

yields sufficient results. This turned out to be our best algorithm in term of test accuracy.

3.4. Pure Neural Networks approaches

3.4.1 ANN

We tried to train a neural network using mainly the numerical attributes (year of birth and LYMPH COUNT) since those are the aspects that are taken into account in building a prior idea over the probability of being sick or not . Using validation results , we shaped the structure to have a rather stable convergence as the number of epochs varied greatly for same structures between different executions .We ended up with this sequence :[Linear(2, 4), Relu, Linear(4, 3), Relu, Linear(3, 1), sigmoid]

3.4.2 MLP / CNN Hybrid

We tried to recreate the deep learning model proposed in the paper [2] "Deep Multi-Instance Learning Using Multi-Modal Data for Diagnosis of Lymphocytosis" . The proposed pipeline is composed from 3 parts:

A MLP(Multi-layer Perceptron) for Clinical Data that is composed from this sequence : [Linear(2, 2), Linear(2, 2), sigmoid, Linear(3, 1), sigmoid] .
This model use negative log-likelihood loss described in the following expression:

$$\mathcal{L}_{\text{MLP}} = -\ln \sigma\left(\hat{y}_i^{\text{MLP}}\right)\Big|_{y_i=1} - \ln\left(1-\sigma\left(\hat{y}_i^{\text{MLP}}\right)\right)\Big|_{y_i=0}$$

Second part is a CNN for blood smears that is composed from a :
-Feature extractor that is analogue to ResNet
-A pooling operation that combines the features of images across the bag (we have chosen the averaging operation as it was recommended in the paper)
-A linear classifier that we represented by a linear layer with a sigmoid. The article use the same loss for part 1 and 2 of the model.
In our case we used K=16 for the feature extractor(see figure 4) which we judged to be a good compromise between

Layer	Layer Name	Output Size	Residual Blocks
1	conv1	112×112	-
2	conv2	56×56	$\begin{bmatrix} 3\times 3, K \\ 3\times 3, K \end{bmatrix}$
3	conv3	28×28	$\begin{bmatrix} 3\times 3, 2K \\ 3\times 3, 2K \end{bmatrix}$
4	conv4	14×14	$\begin{bmatrix} 3\times 3, 4K \\ 3\times 3, 4K \end{bmatrix}$
5	conv5	7×7	$\begin{bmatrix} 3\times 3, 8K \\ 3\times 3, 8K \end{bmatrix}$
6	flatten	$7 \cdot 7 \cdot 8K$	-

Figure 4. Feature extractor structure

speed and good results with the hardware/limits we are using(we are running notebooks on kaggle).

The third and final part is combining both results. In our case,we wanted to go with the average score.

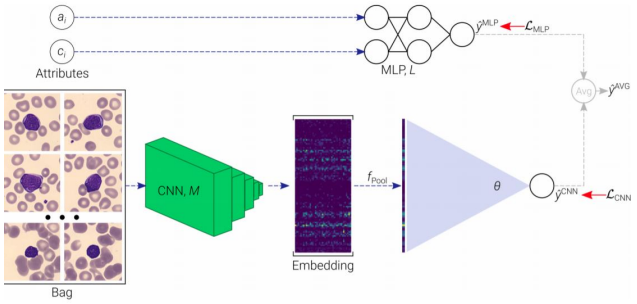

Figure 5. CNN, MLP, and AVG models

We aimed to recreate the model described in figure 5. We put all the parts inside the same model and since the cnn and mlp parts use same loss . We do the average on the score before calculating the loss to have a consistent training joint between the two models.

4. Model tuning and comparison

4.1. Preprocessing

For MIL attention, Instead of using fixed bag sizes we tried to learn from each picture on its own (as in traditional Deep learning approaches) in order to use the full-resolution pictures ,Hence the need for re-sampling (which was done at random) . we tried to resample to a fixed bag size of 50 then of a 100 but results were lacking in both cases due to the fact that we can ignore the key instances during re-sampling. We also tried downsizing the images to

apply the MIL approaches (with an even bigger bag size as mentioned before combined with upsampling).

For hybrid model data,we added also data augmentation with random horizontal flips and random rotation of 15 degree since the training data is not very large.

For both hybrid model and especially SVM+Resnet(since it is pretrained),we added the usual ImageNet normalization.

4.2. Internal Validation

We split our data into 80% train and 20% validation. As validation metric,we use the balanced accuracy measure same as the one employed for Kaggle test data.

$$\text{Balanced Accuracy} = \frac{1}{2}(TPR + TNR)$$

$$TPR = \frac{TP}{TP + FN}, \quad TNR = \frac{TN}{TN + FP}$$

4.3. Ablation study

Table 1. Ablation study for the final solution

Model	Validation	Test
SVM-MIL +ResNet with PCA	81.30%	83.11%
SVM-MIL +ResNet without PCA	64.78%	85.71%

The PCA aims to reduce the risk of overfitting and have a more generalizable model.

4.4. Final Results

Table 2. Models Results

Model	Validation	Kaggle Test
SVM-MIL +ResNet	81.30%	83.11%
XGB	83.47%	78.96%
ANN	76.30%	59.74%
Hybrid	70.43%	63.11%

After a lot of tuning ,we ended up with the following final parameters:

-SVM-MIL + ResNet: Parameter C= 80 (C parameter tells the SVM optimization how much you tolerate misclassifying each training example) , max iteration = 150

-XGB:number of estimators = 1000 , maximum depth(of the trees) = 100

-ANN: Adam optimizer with learning rate = 0.0005 , weight decay = 10e-5 , scheduler multiplying lr by gamma = 0.9 every 300 epochs ,nb epoch = 5000 , batch size = 32

-Hybrid: We used mostly the same parameters as the experiment in the paper.
Adam optimizer with learning rate = 0.0001, weight decay = 0.0005, scheduler multiplying lr by gamma = 0.9 every 96000 iteration , nb epoch = 200, batch size = 1

4.5. Dropped models

4.5.1 Attention mechanism

Despite the model being very interesting and fitting to our case, it suffered from long training cycles (which made playing with the architecture itself a lot more tricky) and required much more training data than what we possessed (despite data augmentation efforts). Moreover, the bag size "balancing" efforts we tried weren't fruitful (we tried resampling but this either had us settling for a small bag size and possibly losing the key instances or oversampling such that we wouldn't have sufficient memory)

4.5.2 ANN

ANN works pretty fine on its own, we "dropped" the approach as in we dropped using it on its own since we figured using the images would obviously allow for a better result .

5. Conclusion

Multi-instance classification is an interesting type of problem especially in this context when we have also attributes data. We tried different approaches focusing either on these attributes exclusively or combined with the bags of images for each patient. Our best result lies in the SVM-MIL+ResNet (both with or without PCA) as it goes beyond the methods that didn't employ the images(XGB and ANN) .We believe that hybrid have the potential as indicated in the paper to give also great results but it requires a very long training period (around 800 epochs were shown in the graphs of the article) .This was a great opportunity to experiments and learn a lot of new things about machine learning.

References

[1] Maximilian Ilse et al. *Attention-based Deep Multiple Instance Learning* . DOI: https://arxiv.org/pdf/1802.04712.pdf.

[2] Mihir Sahasrabudhe et al. *Deep Multi-Instance Learning Using Multi-Modal Data for Diagnosis of Lymphocytosis* . DOI: https://hal.archives-ouvertes.fr/hal-03032875/document.

[3] Yann N. Dauphin et al. *Language Modeling with Gated Convolutional Networks* . DOI: https://arxiv.org/abs/1612.08083.

[4] AMLab-Amsterdam. *AttentionDeepMIL* . DOI: https://github.com/AMLab-Amsterdam/AttentionDeepMIL/blob/master/model.py.

[5] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. *Support Vector Machines for Multiple-Instance Learning* . DOI: https://proceedings.neurips.cc/paper/2002/file/3e6260b81898beacda3d16db379ed329-Paper.pdf.

[6] *miSVM implementation* . DOI: https://github.com/garydoranjr/misvm.