Lymphocytosis Classification Using End-to-End Attention-Based Multiple Instance Learning Model

Roman Castagné
roman.castagne@eleves.enpc.fr

Nathan Godey
nathan.godey@eleves.enpc.fr

Abstract

Lymphocytosis is a condition that consists in an increase of the proportion of lymphocytes in the blood. To identify the disease that causes this condition, a differential diagnosis is performed based on images of blood smears taken from the patient, along with clinical annotations such as age and lymphocytes concentration. Using a dataset of labelled samples made of bags of blood smear images, along with clinical data for each patient, we propose an attention-based classification model able to identify probably diseased cells and to produce a diagnosis for unseen patients.

1. Introduction

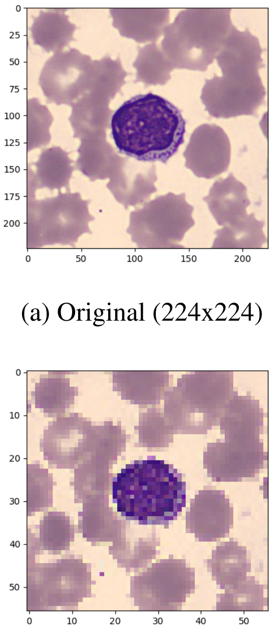
1.1. Problem

Lymphocytosis is a blood disease that can be caused by leukemia and lymphomas for the elderly, or by blood infections for the new-born. It is described as an increase in the count of lymphocytes in the blood. More precisely, in adults, absolute lymphocytosis is declared when the lymphocyte count in the blood rises above 4000 units/μL [1].

The diagnosis of Lymphocytosis and its cause can be done by different measurements and observations, including:

- A direct count of the lymphocytes contained in the patient's blood.
- A slide review of the blood smears can allow identification of the disease causing the lymphocytosis, whether it is leukemia, lymphomia or others [2].

We are provided with a dataset made of slide reviews as bags of images, made of the blood smears of each of 204 patients presenting a case of Lymphocytosis (lymphocyte count above 4000 units/μL). For 162 of these patients, which we refer to as the training set, we have access to clinical annotations regarding their date of birth, gender, lymphocyte count, and a label indicating whether the Lymphocytosis is caused by a tumoral disease (positive class 1) or by another factor (negative class 0). For the remaining 42 patients, which we refer to as the test set, we have access to their date of birth, gender, lymphocyte count, but their class label remains unknown.



(a) Original (224x224)

(b) Downsampled (56x56)

Figure 1: Blood smear images taken from a bag.

1.2. Data and Feature selection

First of all, we notice that the training dataset is imbalanced, and contains about 69% of positive samples. This might lead to local optimas while training our model, such as predicting positive labels only.

Moreover, the blood smears images are already framed properly, showing most of the time a lymphocyte in the center of the image. These images are fairly high-dimensional, with dimension 224x224. We hypothesize that we can reduce the size of these images to 56x56 pixels to reduce time and memory consumption of our models without losing much visual information.

Finally, it can be argued that these lymphocyte pictures don't carry as much information as other types of medical images (such as MRI or lung radiographies for instance), and don't belong in the range of natural images either. This motivated us to use our own model for visual feature extraction rather than using a pre-trained (or fine-tuned) state-of-the-art architecture such as ResNet for instance.

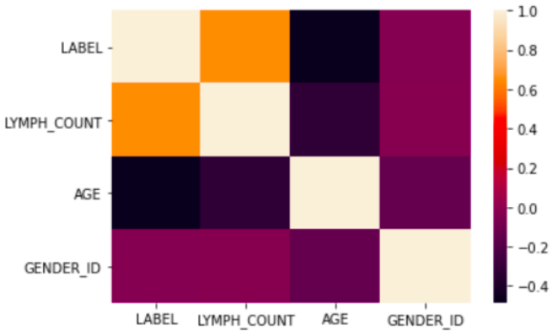


Figure 2: Correlation matrix between features (training set).

When it comes to feature selection, we observed that the gender has no substantial impact on the given label by looking at the correlation matrix, whereas lymphocyte count and age are correlated to the classification. Hence, we decided to use this clinical information (age and lymphocyte count) along our image features.

2. Architecture and Methodology

The problem we have to tackle belongs to the class of multiple instance learning. Namely, the labels we have to predict apply to the bag level, meaning we don't know which lymphocytes are malignant *per se*, but we know that a set of lymphocytes contains some malignant lymphocytes. A patient with at least one malignant lymphocyte is treated as a positive example, while a patient with no malignant lymphocyte is treated as a negative example. We can divide a multiple-instance learning in three distinct components:

- A feature extractor: we want to extract non-local features from our images, to get a low-dimensional representation that will allow us to compare and classify more easily.
- A bag-level aggregator: we want to aggregate our set of low-dimensional representations of the images in a given bag into a single vector representing the whole bag.
- A bag-level classifier: we want to predict labels from the aggregated bag-level representation.

In the following sections, we will expand our implementation of each of the components mentioned above into an end-to-end model training all these sub-models altogether.

2.1. Feature extraction

As mentionned above, we found it inappropriate to start from a pre-trained feature extraction model such as ResNet because the image data is specific (different from a general domain dataset, e.g. ImageNet) and clean. To that extent, we build a classical Convolutional Neural Network (CNN) allowing to encode the images into a low-dimensional representation. The architecture of our CNN consists in:

1. A 2D Convolution layer made of 16 3x3 kernels
2. A 2D Max-Pooling layer of size 2
3. A 2D Convolution layer made of 32 3x3 kernels
4. A 2D Max-Pooling layer of size 2
5. A 2D Convolution layer made of 64 3x3 kernels
6. A 2D Max-Pooling layer of size 2
7. One Dense layer from 3136 (64x7x7) input features into $3 \times d_{emb}$ features

8. One Dense layer from $3 \times d_{emb}$ input features into d_{emb} features

In order to train this CNN, we designed an auto-encoder, using a decoder with a perfectly symmetrical architecture. We took advantage of the indices extracted in the pooling layers from the encoder to unpool the feature maps in the decoder. The auto-encoder was trained to produce embeddings of size d_{emb}, using a batch size of 5 and an Adam optimizer of learning rate 10^{-4} combined with a Mean Squared Error loss. The training stopped when the leave-p-out validation loss started increasing.

The d_{emb} parameter needs to be tuned properly, since a too small embedding won't be able to capture enough information but a too large embedding could carry localized information.

2.2. Attention-based bag-level aggregation

Once the embeddings corresponding to each image in a bag are computed, we have to produce a bag-level compact representation of the image data. To that end, a basic approach would be to compute the mean of the embeddings along each coordinate, which would provide us with a d_{emb}-dimensional representation of each bag. Another approach would be to apply a max-pooling over all these embeddings to extract information.

However, we also need to take into account that some images might be more important than others in the diagnosis. Our first approach was to simply average embeddings using the following :

$$R(b) = \sum_{k=1}^{|b|} \frac{1}{\sum_j \mathcal{W}(e_j)} \mathcal{W}(e_i) e_i$$

Attention based Multiple Instance Learning [3] aims at reweighting each embedding using attention, a probability score on each instance. Formally, to get a representation $R(b)$ of the bag $b = (e_i)_{i \in [1,|b|]}$ where the e_i are the embeddings, we compute:

$$R(b) = \sum_{k=1}^{|b|} a_i e_i$$

and

$$a_i = \frac{\exp(\mathcal{W}(e_i))}{\sum\limits_{j=1}^{|b|} \exp(\mathcal{W}(e_j))}$$

such that $\sum_{i=1}^{|b|} a_i = 1$ and for all $i,\ a_i \geq 0$. $\mathcal{W}$ can be any function. In practice, we use a Multi Layer Perceptron that we can train simultaneously with the rest of the model, since the attention operation is fully differentiable.

This results in an attention-based aggregation of the visual features into a single vector representation of a bag in $\mathbb{R}^{d_{emb}}$.

The MLP $\mathcal{W}$ is designed as follows:

1. One dense layer with d_{emb} input features and 1024 output features

2. One hidden layer with 256 features

3. One hidden layer with 128 features

4. An output layer returning the weight that the image embedding should have in the softmax average

2.3. Bag-level classifier

The bag-level classifier represents the final component of our model. It will map bag representations to a binary label, corresponding to a patient with or without Lymphocytosis. Since we already aggregated features for all images in the bag using attention and only possess a global bag representation, the classifier only has to map this representation to a single scalar in $[0, 1]$, corresponding to the probability associated to the positive label.

In practice, we used the following architecture for this classifier :

1. One dense layer with d_{emb} input features and 1024 output features

2. One hidden layer with 256 features

3. One hidden layer with 128 features

4. An output layer returning the probability that the input bag belongs to the positive class

3. Model tuning and comparison

In this section, we present the performance of our model in the light of different architectural choices, and interpret possible gains in performances.

3.1. Metrics

Our dataset being imbalanced, using appropriate metrics is particularly important for model selection and validation. The validation scores of models are computed using the balanced accuracy and F1 scores for each model. Balanced accuracy is defined as the arithmetic mean of specificity and sensitivity, while the F1 score is defined as the harmonic mean of precision and recall. We additionally computed the ROC curve of our best model as well as its AUC (Area Under Curve) which we present in the next section.

3.2. Results

All results presented in this section were obtained on a validation set constructed from 20% of the labelled data. This task being part of a challenge, the final test set is not available until the last moment.

The following ablation study highlights how our architecture performed depending on its pretraining (with an autoencoder, or without pretraining) as well as the method used for the bag level aggregation. Those results are presented in table 1. Weighted average and softmax attention refer to the approaches described in section 2.2, in that order. Average refers to a simple unweighted average of the embeddings.

Finally, we try different activation functions for the last layer before the softmax attention, to optimize the distribution of pre-softmax weights $W(e_i)$. For instance, when using a Leaky ReLU activation function, it is harder for the model to completely ignore an embedding since $exp(W(e_i))$ will less likely come close to 0. All results were obtained with $d_{emb} = 60$.

One important hyperparameter in our model is the embedding dimension d_{emb} used to represent individual images and then the bag of images. A large value for d_{emb} represents a more informative embedding, but might be too expressive for the bag-level classifier. A small value could lack expressiveness. We present in table 2 the performance of our model for different values of d_{emb}.

Embedding dim	**F1**	**Balanced Accuracy**
40	0.85	0.8695
60	**0.93**	**0.945**
80	0.90	0.9130
100	0.878	0.8043
120	0.757	0.8913

Table 2: Results of our model depending on the dimension of the image/bag embeddings

Using embedding dimension 60 for the autoencoder seems to be the best choice in our model, and allows to obtain a balanced accuracy of 94.5%.

Finally, we plot the Receiving Operator Characteristic (ROC) curve and compute the AUC of our best model in figure 3.

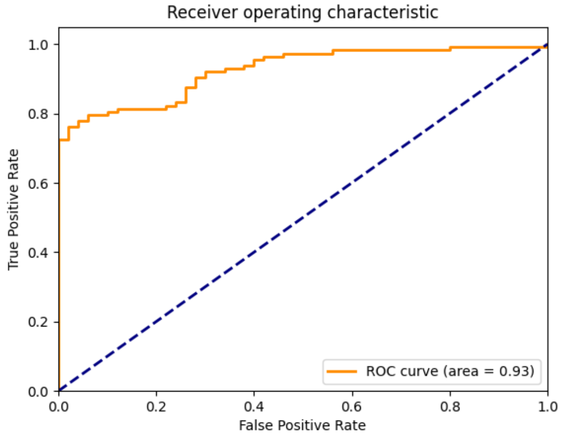


Figure 3: ROC curve of a model pretrained as an autoencoder with Softmax Attention.

The result is close to an ideal ROC curve, with an AUC of 0.93 (a perfect AUC would be 1). Moreover, we can clearly observe a gap around 25-30% of false positive predictions, with a sudden increase of the true positive rate. Picking a threshold right of this sudden increase is interesting as the True Positive Rate is maximal for only a slightly worse False positive rate. An optimal threshold can also be found by maximizing the difference between the True positive rate and the False positive rate. We found that these optimal thresholds were usually found around 0.6-0.7.

Aggregation	Pretrained Encoder	F1	Balanced Acc
Weighted average	No	0.8	0.798
Average	Yes	0.857	0.841
Weighted average	Yes	0.93	0.945
Softmax Attention + linear layer output	Yes	0.884	0.863
Softmax Attention + LeakyReLU	Yes	**0.955**	**0.957**

Table 1: Ablation results of different components of our model.

3.3. Discussion

The first observation from the ablation study is the effectiveness of using an autoencoder to pretrain our feature extractor. Pretraining gains up to 13 points of F1 score and more than 14 points in balanced accuracy. Indeed, training our feature extractor with an autoencoder helps build useful representations of images. Furthermore, images form a strong signal in our dataset, unlike labels that are very scarce.

Using attention to extract representation is particularly useful, with gains from 3 to 10 points in F1 depending on the type of attention compared to an average pooling of image embeddings. Indeed, not all images are of equal importance in Multiple Instance Learning, and a single image of an abnormal lymphocyte is enough to attribute a positive label to the bag. This motivates the fact that not all images should be treated with equal weight in the prediction.

4. Conclusion

In this project, we designed a model to perform Lymphocytosis Classification. This task is framed as Multiple Instance Learning, in which we possess labels only for bags of instances instead of individual instances directly. We divided our model architecture into three parts : a feature extractor that extract embeddings for our images, a bag representation using attention to reweight our embeddings, and a bag level classifier that maps those embeddings to a final vector.

In future developments, one possibility would be to study further the use of a pretrained model as a feature extractor, for instance by trying models pretrained on completely different data (e.g. ImageNet) and compare the performance with our autoencoder.

References

[1] Wikipedia The Free Encyclopedia. Lymphocytosis.

[2] Tracy I. George. Diagnostic approach to lymphocytosis. *The Hematologist*, 12(6), 2015.

[3] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.