

Classification of Lymphocytosis from Blood Cells

Deep Learning for Medical Imaging

Clement Grisi

cole des Ponts ParisTech

grisi.clement@gmail.com

Marius Schmidt-Mengin

cole des Ponts ParisTech

marius.schmidt.mengin@gmail.com

Abstract

In this report, we emphasize our work for the data challenge organized as part of the Deep Learning for Medical Imaging class. The goal of the challenge is to classify lymphocytosis from blood cells. Through iterations in model definition and data representation, we managed to get 0.96623 classification accuracy on the academic leaderboard (rank 1).

1. Introduction

Lymphocytosis is an increase in the number or proportion of lymphocytes in the blood. It's a common finding, which can be either a reaction to infection or acute stress (reactive), or the manifestation of a lymphoproliferative disorder – a type of cancer of the lymphocytes (tumoral). In clinical practice, the diagnosis as either reactive or tumoral is performed by trained pathologists who visually inspect blood cells under a microscope. The final decision also takes into consideration clinical attributes such as age, gender, and lymphocyte count. In spite of being relatively fast and affordable, lymphocytosis diagnosis lacks reproducibility between experts. Additional clinical tests are often required to confirm the malignant nature of the lymphocytes. However, this analysis is relatively expensive and time-consuming, and therefore is not performed for every patient in practice. In this context, automatic classification has the potential to provide accurate and reproducible diagnosis, saving precious time and resources by quickly identifying which patient should be referred for flow cytometry analysis.

2. Problem Definition

2.1. Dataset

Blood smears and patient attributes were collected from 204 patients from the routine hematology laboratory of the Lyon SudUniversity Hospital. All included patients have a lymphocyte count above $4 \times 10^9/L$. The blood smears

were automatically produced by a Sysmex automat tool, and the nucleated cells were automatically photographed with a DM-96 device. The training set consists of 142 patients (44 reactive and 98 malignant cases), and the testing set of 42 patients. For each patient, we have access to dozens of images of lymphocytes, as well as the following clinical attributes: gender, age, and lymphocyte count.

2.2. Evaluation Metric

This challenge is evaluated on balanced accuracy (BA), which normalizes true positive (TP) and true negative (TN) predictions by the number of positive and negative samples, respectively. In particular, if one denotes false positives as FP and false negatives as FN, we have:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP}$$

$$BA = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

2.3. Weak Supervision through Multiple Instance Learning

The idea of weakly supervised learning is to exploit coarse-grained annotations to automatically infer finer-grained information. Coarse-grained information is often readily available in the form of patient level labels, but finer grained annotations are more difficult to obtain. Without precise local annotations, classification models cannot be trained in a fully supervised manner. Therefore, various weakly supervised techniques have recently been developed to overcome this issue. One of these techniques relies on multiple instance learning (MIL), an existing framework largely used in classic computer vision that has recently showed state-of-the-art results in several medical imaging tasks [1].

Babenko [2] gives a good example to understand multiple instance learning. Imagine several people, each of them

having a key chain that contains a few keys. Some of these people are able to enter a certain room, and some aren't. The task is to predict whether or not a given key chain can open the door of that room. To solve this problem, one needs to find the exact key that is common to all the *positive* key chains. If one can correctly identify this key, one can also correctly classify an entire key chain - *positive* if it contains the required key, or *negative* if it doesn't.

Hence, the multiple instance learning framework allows the training of a classifier from weakly labeled data: instead of providing input-label pairs, labels ℓ_b are assigned to *sets* or *bags* of instances. In this setting, the true instance labels ℓ_i can be thought of as latent variables, as they are not known during training. In our case, we can assume that only a subset of the images available for a patient with tumoral lymphocytosis do carry the information necessary to correctly classify that patient. Identifying these special instances perfectly fits in the multiple instance learning framework.

3. Related Work

4. Methodology

4.1. Standard MIL Assumption

top- k = 1

4.2. Extended MIL Assumption

top- k = hyperparameter

4.3. Towards a Custom Aggregation Function

top- k & bottom- k

4.4. Custom Aggregation Function

taking all instances into consideration

5. Architecture

We start from a simple baseline and we evaluate different modifications. For each configuration, we run 10 trainings. Each training has its own seed for initializing the model and splitting the dataset (50/50 split). The 10 seeds are kept the same for each configuration. Each model is trained for 40 epochs with a batch size of 16, resulting in 5 batches per epoch. Evaluation is performed at the end of every epoch. We always optimize with Adam and a learning rate of 10^{-4} . When we report a metric for any configuration, we average the 10 best values of each training, and again average this over all 10 trainings (the average is thus taken over 100 values). We always report 95% confidence intervals.

5.1. Baseline

We adopt a simple and explainable architecture composed of a ResNet [3] backbone followed by an aggregation module. For a given batch of patients, we take all their images and stack them into a batch. This batch is passed independently to the backbone, without global pooling. A 1×1 convolution with one output channel is then applied to the resulting feature maps to obtain pixelwise prediction logits. These logits are rearranged into bags of scores, each bag corresponding to all prediction for one patient. Finally, the aggregation module transforms each bag of scores into a final patient prediction.

It is common that the features are aggregated before being linearly mapped to the final prediction score. In contrast, we apply the aggregation after reducing each feature to a single prediction score. We did not find this to impact the results (and in the case where the aggregation function is linear, both methods are equivalent). Doing so makes the model more interpretable as each pixel of each image is given a score.

As the number of images passed through the backbone is often large, we keep the gradients for only a subset of these images. Pseudo-code for this is provided by Algorithm 1.

```
B, 3, H, W = image_batch.shape
max_forward_size = 512
max_backward_size = 512
features = []
i = 0
# forward images without keeping gradients
with torch.no_grad():
    while i < B - max_backward_size:
        j = min(i+max_forward_size,
                num_images-max_backward_size)
        features.append(backbone(images[i:j]))
        i = j
# forward image with gradients
features.append(backbone(images[i:]))
```

5.2. Augmentations

As the images are centered on a lymphocytes, we always apply a center crop of size 112 to all images (training, validation and testing). Initial experiments showed that this significantly improves the accuracy and reduces the computational resources. We use vertical and horizontal random flipping. We tried to use more aggressive augmentations such as affine transforms and color jittering but this heavily impaired the accuracy.

5.3. Backbone Choice

We compared three backbones: ResNet18, ResNet34 and ResNet50 [3]. We also tried a more recent backbone, EfficientNet [4], but did not obtain good results. The results are summarized in Figure 4. ResNet18 achieves the best validation balanced accuracy. We use it for all our subsequent experiments.

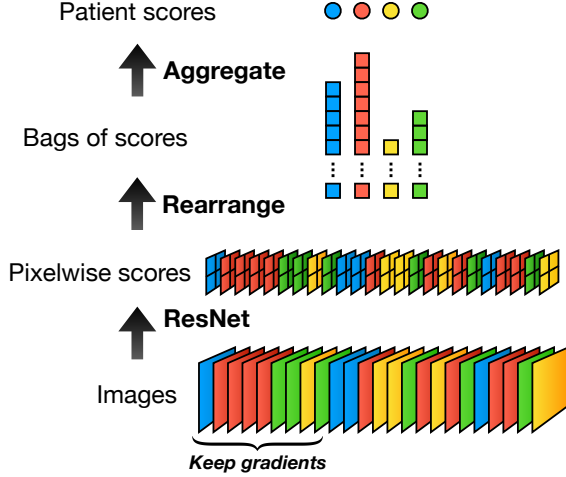


Figure 1: Baseline architecture.

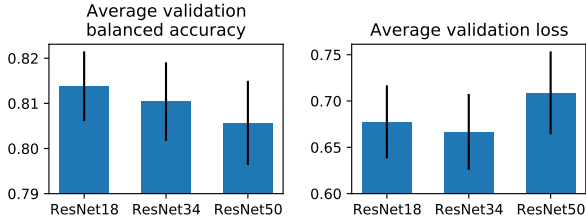


Figure 2: Impact of ResNet depth on validation loss and balanced accuracy. The black bars indicate 95% confidence intervals. Shallower seems to be better but the results are not very significant.

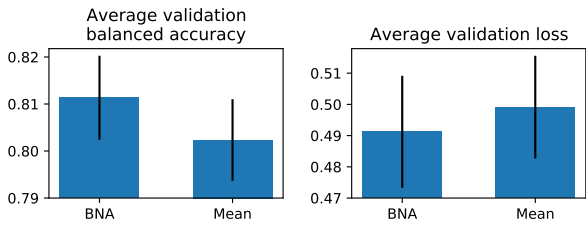


Figure 3: Impact of our BN aggregation function on validation loss and balanced accuracy. The black bars indicate 95% confidence intervals.

5.4. Number of images

6. Results

7. Conclusion

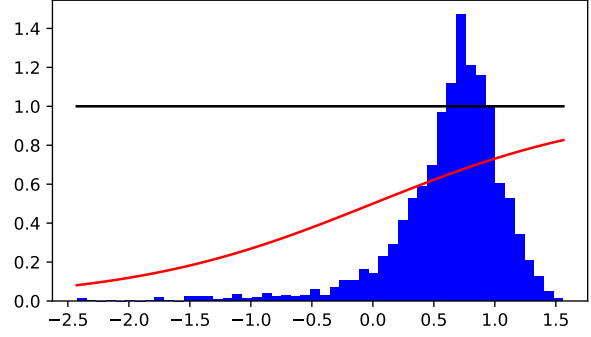


Figure 4: Impact of our BN aggregation function on validation loss and balanced accuracy. The black bars indicate 95% confidence intervals.

References

- [1] Le Hou, Dimitris Samaras, Tahsin M. Kurç, Yi Gao, James E. Davis, and Joel H. Saltz. Efficient multiple instance convolutional neural networks for gigapixel resolution image classification. *CoRR*, abs/1504.07947, 2015. 1
- [2] Boris Babenko. Multiple instance learning: algorithms and applications. 2008. 1
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 2
- [4] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. 2