Regression Constraint for an Explainable Cervical Cancer Classifier

Antoine PIROVANO[1,2], Leandro G. ALMEIDA[1], Said LADJAL[2]

[1]Keen Eye, 74 Rue du Faubourg Saint-Antoine, 75012 Paris, France

[2]Laboratoire Traitement et Communication de l'Information, Telecom ParisTech, Universite Paris-Saclay, Paris, France

antoine.pirovano@keeneye.tech, leandro.almeida@keeneye.tech,
said.ladjal@telecom-paristech.fr

Résumé – Cet article s'intéresse à la classification automatique de cellules de l'épithélium pavimenteux pour le depistage cancer du col de l'utérus en s'appuyant sur les outils de l'apprentissage profond. Nous étudions differentes architectures sur un jeu de données public nommé Herlev qui consiste classifier des images de cellules, issues d'un frottis du col de l'utérus, au regard de l'anormalité qu'elles représentent. De plus, nous utilisons et adaptons une méthode d'attribution afin de mettre en lumière les caratéristiques cytomorphologiques discriminantes qui sont utilisées pour la classification. A travers ce papier, nous détaillerons les methodes et architectures qui nous permettent d'atteindre des performances optimisées: 75% de précision pour la classification de la sévérité et 97% pour la classification de la normalité.

Abstract – This article adresses the problem of automatic squamous cells classification for cervical cancer screening using Deep Learning methods. We study different architectures on a public dataset called Herlev dataset, which consists in classifying cells, obtained by cervical pap smear, regarding the severity of the abnormalities they represent. Furthermore, we use an attribution method to understand which cytomorphological features are actually learned as discriminative to classify severity of the abnormalities. Through this paper, we show how we trained a performant classifier: 75% accuracy on severity classification and 97% accuracy on normal/abnormal classification.

1 Introduction

The World Health Organization (WHO) states [1] that around 90% of cervical cancer could be avoided if they were detected and treated earlier. At 500×10^3 new cases at year, screening for cervical cancer needs to be efficient and precise.

With the recent emergence of machine learning using deep Convolutional Neural Networks (CNN) and its success on a large panel of tasks, a lot of work as been done to assist doctors and medical practices [2, 3] using such methods. In the case of cervical cancer, the Herlev public dataset enables to compare different methods on this specific task by providing images of single cells and organizing them into classes regarding the malignancy they represent.

In this paper, we will firstly exploit the ordinal nature of the WHO classification present in the Herlev dataset, by designing a loss function that leads to a training paradigm that closely resembles the medical task at hand. Finally we will apply attribution methods to determine what cytomorphological features are associated with the classification model. This will not only give us confidence in the training process and prove that the model learned relevant features but also show the potential for weak localization tasks.

2 Related Work

Since 2012 and the success of AlexNet on Imagenet Challenge [4], deep CNN have provided high accuracy results in large range of different tasks. Over the years, several architectures have been given a lot of attention. For example, Resnet-101 [5] proposes to use skipped connections over blocks to avoid de-learning on more abstract features.

Previous works have applied CNN models to the Herlev data set using binary normal and abnormal categories. In [11] they reach a 0.78 F1 scoring using a support vector machine. In [12] they use a unsupervisely trained Feature Selection model after a CNN feature extractor to reach a F1 score of 0.90 and an accuracy of 94%. In [6, 7], they used, respectively, an Alexnet-like and a Resnet architecture and trained them on Herlev dataset using normal vs abnormal to provide a model that reaches binary classification accuracy of 98.3%.

3 Herlev Severity Classification using Regression Constraint

3.1 Herlev Dataset

The Herlev Dataset is a cytology image set composed of 917 images gathered in 7 classes : normal columnar, normal intermediate, normal superficial, light dysplastic, moderate dysplastic, severe dysplastic, and carcinoma in situ. The three first

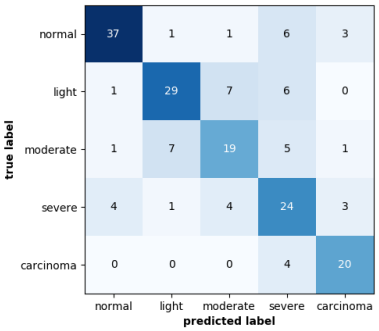

FIGURE 1 – Resnet-101 confusion matrix on Herlev severity test set

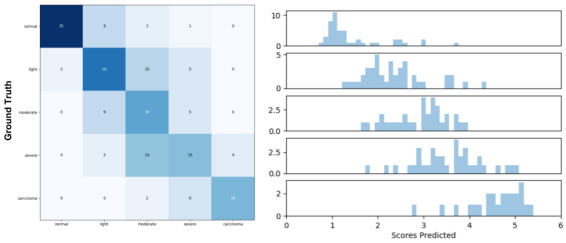

FIGURE 2 – Confusion Matrix (left) and Scores Distribution (right) given by Resnet-101 Regressor regarding Herlev Severity Classes

classes belong to the category of normal cells and the last four are abnormal ones (in order of severity, carcinoma in situ hinting at the presence of an actual cancer). Images are between 50 and 400 pixels wide. Previous work processed the set in a binary classification problem of normal vs abnormal classes. Here, we merged normal images into a single class in order to study the medical severity only, thus building a 5 classes dataset, we call Herlev severity consiting of : normal, light dysplastic, moderate dysplastic, severe dysplastic, and carcinoma in situ.

3.2 Herlev Severity

This section descibes three pipelines that can be used to train a severity model and the motivation that led to them.

3.2.1 Classification Pipeline

We started by retraining a Resnet-101 model pretrained on ImageNet [10] on Herlev severity dataset. The computed performances were a mean AUC of 0.9, with the highest AUC being 0.95 on the carcinoma in situ class and lowest being 0.87 onsevere dysplastic with an overall accuracy of 70%, a binary (normal/abnormal) accuracy of 91% and a binary F1 score of 0.94.

From the confusion matrix shown in Figure 1, we see that the model tends to misclassify images from the normal and carcinoma in situ classes. This was already reported in [6] and identified to be due to the visual similarities between normal columnar and carcinoma in situ cells. Obviously, missing a potential highly abnormal diagnonis is something to avoid. Similarly, due to the fact that 93% of pap smears are normal during routine diagonis, misclassifying normal cells would require a additional action by the attending cytotechnicians.

3.2.2 Regression Pipeline

Since the WHO classification used in the Herlev set have an order of severity, this task can be interpreted as a regression problem. regression loss will oblige the network to focus on how to differentiate normal samples from malignant ones. We relabel Herlev samples using a score from 1 (for normal ones) to 5 (for carcinoma ones) and use a mean square error as loss to optimize. Thus, we retrain the exact same Resnet-101 architecture replacing, to have a single score output, the softmax layer by a fully connected layer.

Figure 2 shows the distribution of scores predicted on the test set and highlights that the model succeeded in assigning scores regarding maligancy. Most importantly, it does not misclassify any normal samples or carcinoma in situ samples with each other. A further point to note from the confusion matrix deriving from this distribution, this model does more misclassifications than the categorical model, with an accuracy of 60.3%, however these misclassifcation are less severe in the scope due to their relative prognosis distance. This is can be more easily displayed by the overall MSE of 0.58 over the test set. The binary accuracy was of 92% and the F1 score was 0.95.

3.2.3 Classification + Regression Pipeline

While the regression loss was more adapted than a classification (cross entropy) loss to the severity task, it nonetheless did not improve the performances per class. In this section we combine the strength of both approaches into a single architecture.

Figure 3 shows the additional layer to the classification architecture. We simply sum the cross entropy loss and the MSE loss. This would be equivalent to weighting loss regarding the distance between the ground truth class index and the predicted class index. We turn probabilities given by the softmax layer into a score using a fixed weights fully connected layer corresponding to the class score (or class index).

Noting $p = (p_1, \ldots, p_5)$ these class probability neurons, our loss finally reads $\mathcal{L}(x) = \mathcal{CE}(p; y_x) + (y_x - \sum_{i=0}^{4}(i+1).p_i)^2$ where x is an image, y_x the label (one hot for cross-entropy and score for the regression constraint) and $\mathcal{CE}$ is the cross-entropy loss.

On Figure 4, we can see that our Resnet-101, Classifier + Regressor, makes less misclassifications than the classifier and lower MSE than the regressor. Thus, we have an architecture performing on classification task (mean AUC = 0.94) and on scoring severity task (average MSE = 0.51). What is particularly appreciated here is that the 'extreme' classes ('normal'

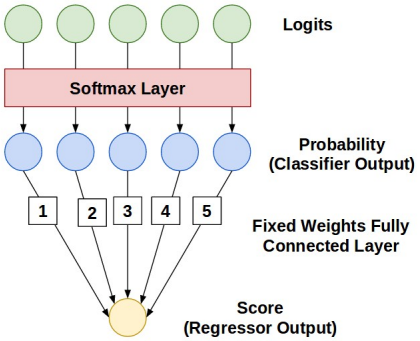

FIGURE 3 – Regression Constraint applied to Classifier

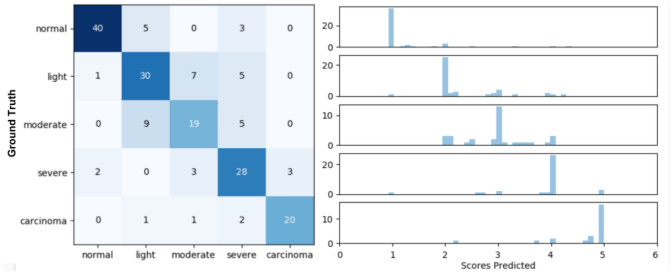

FIGURE 4 – Resnet-101 Classifier + Regressor Scores Distribution and Confusion Matrix on Herlev Severity Test Set

and 'carcinoma in situ') have the best AUC (respectively 0.98 and 0.97). The overall accuracy of the order of 74.5% and the binary accuracy was 94% with an F1 score of 0.96.

3.2.4 Pipeline Comparisons

Figure 5 shows the AUC distribution per class obtained training the classifier pipeline and the classifier + regressor pipeline on 4 random folds. It brings to the fore how the regression loss does not change much on the light dysplastic, moderate dysplastic and severe dysplastic classes but improves 'extreme' cases especially 'normal' samples that were really impacted by the ressemblance between 'normal columnar' and 'carcinoma in situ' samples.

4 Explainability / Interpretability

Understanding how our model arrives to the severity of cancer progression is an important step in validating its use. Besides giving the user assurances of its performance, it allows us to understand and possibily build stronger models. We need a method that provides meaningful explanations, which ideally are related to the cytomorphological features and used by cytotechnians and doctors during day-to-day routine. Gradient based methods give the attribution to the classification associated with each input feature given to the model, in the case of digital images of cytology slides, the image pixels. This allows us to

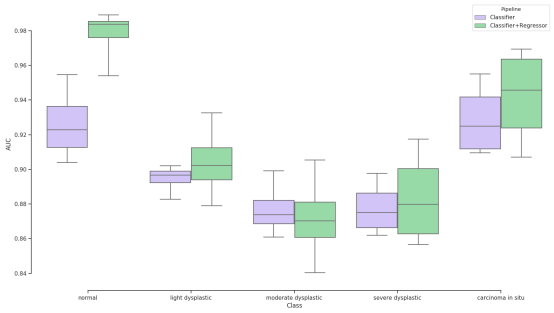

FIGURE 5 – AUC Distribution for Classifier & Classifier + Regressor Pipeline Comparison using 4 random folds

identify and localized regions that contribute to the severity of the diagnosis. Integrated Gradient [8] is of particular interest due to its model agnosticity and its baseline comparision.

In this section we are going to use an attribution method to understand what has been learned by our models and on what cytomorphological features it relies to assign a degree of malignancy. The Bethesda guidelines [9] states that the main cytomorphological features used to determine the severity are mostly based on nucleus, we thus would expect the attribution to be in the nucleus region.

Integrated Gradient For the attribution we utilize a model agonistic methods, the integrated gradient. As with most attribution methods it relies on the comparision between the image and baseline (that is representative of the absence of the class of the image) and computation of the gradient to the image. The attribution map, $Am(x_i; F, x')$ for an image x gives the contribution of $i-$ th pixel given a model F and baseline image x', $Am(x_i; F, x') = (x_i - x'_i).\sum_{k=0}^{m} \frac{\delta F(x' + \frac{k}{m}.(x - x'))}{\delta x_i} . \frac{1}{m}$.

Baseline Design What we are interested in here is how our model predicts the malignancy (i.e. regression result), this is why we will try the Integrated Gradient method on malignant samples i.e. dysplastic and carcinoma in situ samples. An obvious abscence of object in Pap tests context is a white image (since background of pap smears slides is white).

Qualitative Results Figure 6 shows examples of the attribution map from integrated gradient method, along with the annotated cytology features of the associated the images. This highlights that the malignancy scoring seem to be mainly due to the nucleus.

Quantitative Results Here we make use of the annotation masks present in the Herlev set to create specific attribution metrics. Given their role in the different consensus and guidelines, we measure the amount of attribution within the nucleus and cytoplasm compared to total attribution (respectively denoted as At_N and At_C), these contributions are given by,

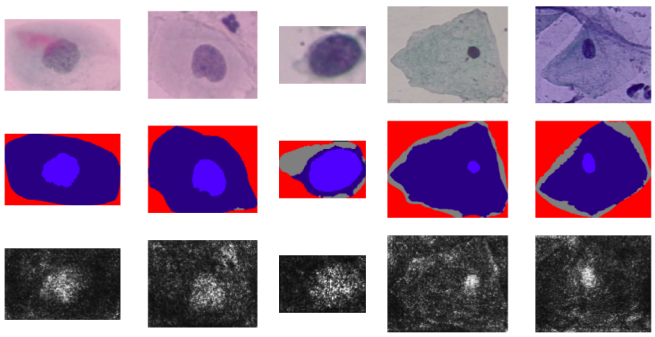

FIGURE 6 – Integrated Gradient Result on 5 images Herlev set (top) using a white baseline and cytological feature masks present in the dataset (middle). Attribution maps are shown showing "activated" pixels mostly in nucleus (bottom).

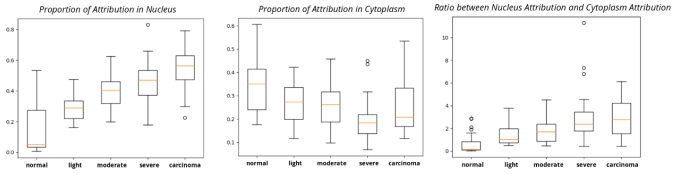

FIGURE 7 – Integrated Gradient Result on Herlev Test set using a white baseline and Associate Mask

$$At_N = \frac{\sum_{i \subset \mathcal{N}} Am(x_i)}{\sum_i Am(x_i)}, \quad (1)$$

$$At_C = \frac{\sum_{i \subset \mathcal{C}} Am(x_i)}{\sum_i Am(x_i)}, \quad (2)$$

where $\mathcal{N}$ and $\mathcal{C}$, refer to the nucleus and cytoplasm pixels respectively and Am is the attribution map defined before. In order to understand how much each region contributes to the model's prediction, we also compute the ratio of nucleus and cytoplasm attribution.

Figure 7 shows the distribution for At_N, At_C, and their ratio for each severity class. It emphasises the relevance of the nucleus over the cytoplasm for the model as the severity increases. Particularly, in the case of carcinoma in situ, the nucleus contributes 2 times more than when classifying a normal case.

5 Conclusion

In this work, we have shown that a proper loss design, based on the final goal of the medical exam under study, one can construct a model that differentiates properly between normal and abnormal cells reaching a severity accuracy of 74.5%, a binary accuracy of 96.7% was achieved along wiht a F1 score of 0.95. Furthermore, we adapted an attribution method that can be used by doctors to check the relevance of the network's decision. These two contributions are essential in the construction of an automatic diagnostic assistance method that can be trusted and accepted by doctors.

Acknowledgements

We are grateful to our colleagues at Keen Eye and Telecom ParisTech (LTCI) for many valuable discussions, in particular Isabelle Bloch for advice and encouragment. This work was supported by ANRT.

Références

[1] World Health Organization *Comprehensive cervical cancer control : a guide to essential practice*. 2006.

[2] Yaniv Bar and Idit Diamant and Lior Wolf and Hayit Greenspan *Deep learning with non-medical training used for chest pathology identification*. Medical Imaging 2015 : Computer-Aided Diagnosis, 2015.

[3] Olaf Ronneberger and Philipp Fischer and Thomas Brox *U-Net : Convolutional Networks for Biomedical Image Segmentation*. Medical Image Computing and Computer-Assisted Intervention MICCAI 2015.

[4] Alex Krizhevsky and Ilya Sutskever and Geoffrey E. Hinton *ImageNet Classification with Deep Convolutional Neural Networks*. Advances in Neural Information Processing Systems 25, 2012.

[5] Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun *Deep Residual Learning for Image Recognition*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[6] Le Lu and Ling Zhang and al *DeepPap : Deep Convolutional Networks for Cervical Cell Classification*. IEEE Journal of Biomedical and Health Informatics, 2017.

[7] G. Forslid and H. Wieslander and al *Deep Convolutional Neural Networks for Detecting Cellular Changes Due to Malignancy*. IEEE International Conference on Computer Vision Workshops, 2017

[8] Mukund Sundararajan and Ankur Taly and al *Axiomatic Attribution for Deep Networks*. ICML 2017.

[9] D. Solomon and D. Davey and al *The 2001 Bethesda System : Terminology for Reporting Results of Cervical Cytology*. 2001

[10] J. Deng and W. Dong and al *A Large-Scale Hierarchical Image Database*. CVPR 2009.

[11] Jonghwan Hyeon and Ho-Jin Choi and al *Diagnosing cervical cell images using pre-trained convolutional neural network as feature extractor*. IEEE International Conference on Big Data and Smart Computing, 2017

[12] Kangkana Bora and Manish Chowdhury and al *Pap Smear Image Classification Using Convolutional Neural Network*. Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, 2016