A Study on the Impact of Class Distribution on Deep Learning—The Case of Histological Images and Cancer Detection - Extended Abstract

Ismat Ara Reshma[1,*], Josiane Mothe[1,*], Sylvain Cussat-Blanc[1], Hervé Luga[1], Camille Franchet[2], Margot Gaspard[2], Pierre Brousset[2] and Radu Tudor Ionescu[3]

[1]*IRIT, UMR5505 CNRS, Univ. de Toulouse, 118 Route de Narbonne, Toulouse, F-31062 CEDEX 09, France*
[2]*Dept. of Pathology, Univ. Cancer Institute of Toulouse-Oncopole, 1 avenue Irène Joliot-Curie, Toulouse, F-31059 France*
[3]*Univ. of Bucharest, 14 Academiei, Bucharest 010014, Romania*

Abstract

Studies on deep learning tuning mostly focus on the neural network architectures and algorithms hyper-parameters. Another core factor for accurate training is the class distribution of the training dataset. This paper contributes to understanding the optimal class distribution on the case for histological images used in cancer diagnosis. We formulate several hypotheses, which are then tested considering experiments with hundreds of trials. We considered both segmentation and classification tasks considering the U-net and group equivariant CNN (G-CNN). This paper is an extended abstract of another paper published by the authors[1].

Keywords

Computer-aided diagnosis, medical information retrieval, image segmentation and classification, deep learning, class-biased training, class distribution analysis, histological image

The huge success of deep learning models, such as convolutional neural networks (CNNs), in visual recognition has encouraged researchers to explore their use in various domains, including cancer detection from histological images. Histological images or whole slide images (WSIs) are digitalized histological slides. The methods for automatic cancer detection are mainly focused on end-to-end pipeline systems. The success of such systems depends on several hyper-parameters. We hypothesized that one of the most important hyper-parameters is the class distribution of the training set, as it provides the supervision for all learning-based systems.

In machine learning, an imbalanced data distribution has been shown to lead to inferior models compared to a balanced distribution in many domains including biomedical and information

[1]Reshma IA, Franchet C, Gaspard M, Ionescu RT, Mothe J, Cussat-Blanc S, Luga H, Brousset P. Finding a Suitable Class Distribution for Building Histological Images Datasets Used in Deep Model Training—The Case of Cancer Detection. Journal of Digital Imaging. 2022 Apr 20:1-24.

CIRCLE '22: Conference of the Information Retrieval Communities in Europe, July 04–07, 2022, Samatan, France

*Corresponding author.
Ismat-Ara.Reshma@irit.fr (I. A. Reshma); Josiane.Mothe@irit.fr (J. Mothe); Sylvain.Cussat-Blanc@irit.fr (S. Cussat-Blanc); Herve.Luga@irit.fr (H. Luga); Franchet.Camille@iuct-oncopole.fr (C. Franchet); Brousset.p@chu-toulouse.fr (P. Brousset); raducu.ionescu@gmail.com (R. T. Ionescu)
https://www.irit.fr/~Josiane.Mothe/ (J. Mothe)
0000-0002-9917-6668 (I. A. Reshma); 0000-0001-9273-2193 (J. Mothe); 0000-0001-8675-197X (H. Luga); 0000-0002-3214-0142 (C. Franchet); 0000-0002-8629-3291 (P. Brousset); 0000-0002-9301-1950 (R. T. Ionescu)

retrieval[2]. Balanced distribution became the default choice in deep learning state-of-the-art methods[3], although it is not optimal in all cases. There are very few analytical studies on the performance impact of different distributions. They were mainly conducted on toy datasets, even though real datasets may be very different and more complex. There is no evidence that the conclusions of these studies would be appropriate for cancer WSIs.

We present a data-driven analysis which determines the performance impact of different class distributions on training data. We derived several hypotheses with regard to WSIs used for cancer detection. WSIs comprise regions of interest (ROI), where pathologists look for any abnormalities, and the non-ROI. We tested the hypotheses with both image segmentation and classification tasks.

Data imbalance (class bias) is a common problem in machine learning, and many methods have been proposed to make data balanced[4]. A separate analysis is certainly required for each special kind of data following the *No Free Lunch Theorem*[5]: none single model works best for every task.

Moreover, deep CNNs have shown incredible performance levels with regard to cancer detection in WSIs. Bejnordi et al. organised a world-wide challenge known as CAMELYON on cancer detection in WSIs [6]. Most of the proposed methods in the CAMELYON16 challenge were based on deep learning; the variation in the participants' results is induced by hyper-parameter settings and data pre-processing. The winning team[7] trained two 22-layer GoogleNets (V1), one with randomly sampled training patches–probably biased towards negative examples–and another with additional hard negative examples.

In this work, we consider four categories of patches: ROI categories, *cancer* ($\mathbb{C}$), *non-cancer* ($\neg\mathbb{C}$), or multi-label *mixed* ($\mathbb{C}\&\neg\mathbb{C}$) and the *other* ($\mathbb{O}$), non-ROI category. We make several hypotheses and design several experiments with the relevant class distributions to be able to test the proposed hypotheses. The total number of patches in the training set of each experiment is kept the same to ensure fair comparison but their distribution differs. We introduce $\mathbb{U}$ to denote a unit (fixed number) of patches. Here, the results for segmentation are reported while in the initial paper we considered both binary classification and segmentation.

At the training step, we generate different class distributions in the training set (See Table 1). The generated training set is used to train a fully convolutional neural network (FCNN) U-net[8].

[2]Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002 Jun 1;16:321-57.

[3]Halicek M, Shahedi M, Little JV, Chen AY, Myers LL, Sumer BD, Fei B. Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. Scientific reports. 2019 Oct 1;9(1):1-1.

[4]Prati RC, Batista GE, Silva DF. Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowledge and Information Systems. 2015 Oct;45(1):247-70.

[5]Wolpert DH. The lack of a priori distinctions between learning algorithms. Neural computation. 1996 Oct 1;8(7):1341-90.

[6]Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, Van Der Laak JA, Hermsen M, Manson QF, Balkenhol M, Geessink O. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama. 2017 Dec 12;318(22):2199-210.

[7]Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718. 2016 Jun 18.

[8]Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InInternational Conference on Medical image computing and computer-assisted intervention 2015 Oct 5 (pp. 234-241). Springer, Cham.

Table 1
Experiment settings. E1 setting is designed to test H1. E2, E3, and E4 test H2 in different settings with single-label, multi-label, and non-ROI patches, respectively. Comparison between E2 and E3 tests H3. Comparison between E3 and E4 tests H4. There is a total of 9 units ($\mathbb{U}$) of training patches in E1 and 4$\mathbb{U}$ in the other settings. Here, 1 $\mathbb{U}$=5000 patches.

Experiment ID	Distribution	Patch ratio ($\mathbb{O} : \mathbb{C} : \neg\mathbb{C} : \mathbb{C}\&\neg\mathbb{C}$)
E1.a	Balanced	$3 : 3 : 3 : 0$
E1.b	Over-represented $\mathbb{O}$ (natural)	$7 : 1 : 1 : 0$
E2.a	Balanced	$0 : 2 : 2 : 0$
E2.b	Over-represented $\neg\mathbb{C}$	$0 : 1 : 3 : 0$
E2.c	Over-represented $\mathbb{C}$	$0 : 3 : 1 : 0$
E3.a	Balanced	$0 : 1.5 : 1.5 : 1$
E3.b	Over-represented $\neg\mathbb{C}$	$0 : 0 : 3 : 1$
E3.c	Over-represented $\mathbb{C}$	$0 : 3 : 0 : 1$
E4.a	Balanced	$1 : 1 : 1 : 1$
E4.b	Over-represented $\neg\mathbb{C}$	$1 : 0 : 2 : 1$
E4.c	Over-represented $\mathbb{C}$	$1 : 2 : 0 : 1$

During inference, the trained model is employed to predict the patches extracted from unseen test WSIs. Since false positive (FP) is still an ongoing problem in cancer detection in WSI, we focus on minimizing FPs and utilize FP-based evaluation metrics, although false negative (FN)-based metrics are also considered. Specifically, we test our hypotheses by employing receiver operating characteristic (ROC) curve, precision-recall (PR) curve, precision, and false positive rate (FPR) curves, although here, because of the page limit, we present the latter curves only.

To generate the result, we used the Metastatic Lymph Node dataset from the University Cancer Institute of Toulouse-Oncopole, which is abbreviated as MLNTO. We extracted 127,898 (15,328 belong to $\mathbb{C}$) and 101,262 (17,351 belong to $\mathbb{C}$) patches from the training and test sets, respectively (see our original paper[1] for detail.) There is no duplicate patches, but homogeneous and heterogeneous patches occur.

H1: Balanced distribution is optimal for training a model. To test H1, we designed two experiments: E1.a and E1.b. In E1.a we consider the same number of patches in each of the three classes ($\mathbb{C}$, $\neg\mathbb{C}$, $\mathbb{O}$), whereas in E1.b the training examples are highly biased (7 times) towards class $\mathbb{O}$ (similar to the natural distribution) as presented in Table 1. To test H1, a total of 9$\mathbb{U}$ of patches is used to create both the natural and balanced distributions.

According to the result (see Figure 1), the natural distribution (blue curve) is better than the balanced one (green curve). The same result holds when considering the ROC and PR curves.

H2: Over-representing the $\neg\mathbb{C}$ class in the training set reduces false positives during cancer detection. In experiment E2.a (see E2 settings in Table 1), we consider the balanced case between $\mathbb{C}$ and $\neg\mathbb{C}$, while E2.b over-represents $\neg\mathbb{C}$ and E2.c over-represents $\mathbb{C}$.

We found that $\neg\mathbb{C}$-biased distribution (blue curve) is better than the two other distributions:

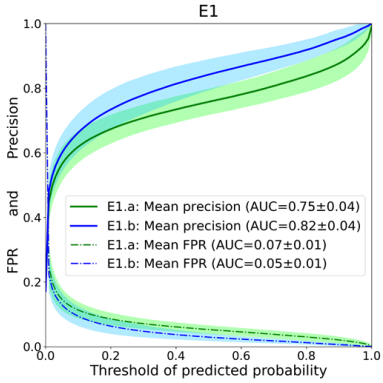

Figure 1: Natural distribution (E1.b, blue) is better than balanced distribution (E1.a, green) (¬H1). Mean precision and FPR curves for 10 runs for balanced (E1.a) and over-representation of $\mathbb{O}$ (E1.b) distributions; color shading is the standard deviation.

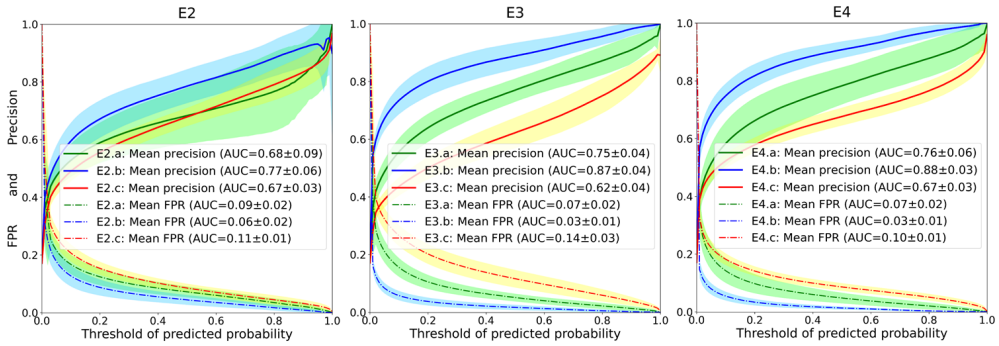

Figure 2: All hypotheses tested true. Here, E2 (with single-label only), E3 (with multi-label), and E4 (with non-ROI) test H2 in different settings; comparison between E2 and E3 tests H3; comparison between E3 and E4 tests H4. Same notion as Figure 1.

the balanced (green curve) and the $\mathbb{C}$-biased (red curve) ones. H2 is true according to both precision and FPR curves (see Figure 2).

H3: Multi-label examples are more useful than single-label examples as training data. We design three experiments (E3 settings in Table 1). First, in E3.a, we considered a balanced case between $\mathbb{C}$ and $\neg\mathbb{C}$. Then, similarly to E2, in E3.b and E3.c we considered over-represented $\neg\mathbb{C}$ and over-represented $\mathbb{C}$ cases.

Experiments with multi-label examples (E3) are better than the ones with single-label (E2) with an exception for the $\mathbb{C}$-biased case (E3.c) (see E2 and E3 in Figure 2). The exception occurs because of increasing the $\mathbb{C}$ bias in E3.c than in E2.c (see Table 1). H3 is thus true according to the precision and FPR curves. When comparing the $\neg\mathbb{C}$-biased case in the current setting (E3.b) with the balanced (E3.a) and $\mathbb{C}$-biased (E3.c) cases, $\neg\mathbb{C}$-biased produces less false positives, i.e., H2 is thus also true in this setting (see Figure 2, E3).

H4: Non-ROI data are useful for training. We designed three experiments denoted as E4.* in Table 1. The first purpose is to test H4 by comparing E4 with E3; the second is to re-test H2 with the current E4 settings.

When comparing the precision and FPR curves of the experiments with non-ROI data (E4) with the ones without non-ROI (E3), H4 is true (see E3 and E4 in Figure 2). When comparing the $\neg\mathbb{C}$-biased case in the current setting (E4.b) with the balanced (E4.a) and $\mathbb{C}$-biased (E4.c) cases, $\neg\mathbb{C}$-biased produces less false positives (see Figure 2, E4); H2 is thus true here as well.

To conclude, in this research which was published in details in another paper of ours[1], we performed a data-level analysis to determine the optimal distribution of the classes in the training set for WSIs when using deep learning. In natural distribution, the WSI data is highly biased towards the non-ROIs. Common practice is to artificially balance the classes while there is no evidence this is accurate. To the best of our knowledge, our analysis is pioneering in the case of class distribution analysis of WSI data for deep learning models; previous research has focused on end-to-end pipeline development for cancer detection. We show that non-ROI easy to annotate patches help the model training. This result will be helpful for researchers who are building a training dataset of WSIs or other applications in which annotation is costly. Such analyses could also help in other real-world problems where data have a complex historyregarding the importance of building a training set with proper distribution.