Small Image Training Sets: Exploring the Limits of Conventional and CNN-based Methods

Vassili Kovalev
Biomedical Image Analysis Dept.
United Institute of Informatics Problems
of NAS of Belarus
Minsk, Belarus
vassili.kovalev@gmail.com

Abstract. **This work is dedicated to the problem of image classification under the condition of small image datasets. Both traditional and CNN-based methods are examined and compared based on a benchmark image dataset. The dataset consisted of 12000 routine hematoxylin-eosin stained histological images. They represent the biopsy samples of normal tissue and the malignant tumors caused by breast cancer. The commonly-known image analysis methods which make use of color co-occurrence matrices of images converted to an adaptive 32-color space and the limited number of their principal components (PCA) were used as image features. The features were inputted to SVM and Random Forests classifiers. The original image training set was gradually reduced from 8400 to 840 images with the step of 10%. In addition, the very-small sub-samples of 5% (420), 2.5% (210), 1.25% (105), and 1% (84) of original image dataset were also examined. In its turn, the classical CNN was employed that consisted of only 3 convolutional + MaxPooling layers with 16, 32, and 64 filters respectively. This is because the small image training sets were specifically targeted in this particular study. The convolutional part was followed by a fully connected neural network with 512 intermediate nodes. As a result, it was found that traditional methods outperform the CNN-based image classification technique on the training sets comprised of less than 840 images.**

Keywords: **Image Classification, Benchmarking, Convolutional Neural Networks, Histology images**

I. INTRODUCTION

Nowadays, the Convolutional Neural Networks (CNNs) and Deep Learning (DL) techniques are widely used for solving various image processing, segmentation, classification, clustering, and even realistic image generation problems. These methods have demonstrated tremendous promises in different application domains including medical image analysis, classification, and computerized disease diagnosis [1, 2]. However, training of CNNs with recent architectures requires large amounts of professionally labeled medical images of different classes that could be difficult to collect, laborious to label, and costly.

The histopathology image analysis based on light microscopy has long been recognized as a gold standard in cancer diagnosis. Modern digital pathology which includes whole slide imaging (WSI) scanners and automated image analysis solutions provides a more efficient and cost-effective way of handling, visualization, and analysis of the pathology image data [3]. Although the conventional methods of WSI image analysis based on the extraction of color and morphological features are still in use [4], the new DL approaches often demonstrate better performance and higher tolerance to image variability caused by a number of different factors [5, 6]. In [6] authors have isolated, carefully enumerated, and characterized 10 major challenges of AI in digital pathology which we are currently facing. The challenges that are most relevant to the present study include lack of labeled data, pervasive variability, and so-called realism of DL which is associated here with the available computation power.

Presently, it is commonly understood that the classification results always depend on the degree of representativeness of images included in the training set. Therefore, it is highly desirable that these images should be as "representative" as possible for the classification problem we are dealing with. In terms of the feature space, this means that the training image samples should cover well the regions of feature space that could be potentially populated by the image classes we considering. Such a problem is directly relevant to the following two major factors: the size of the training set and, (b) the variability of images inherent to the classes.

In this paper, we are trying to shed light on the problem of small image training sets and image variability on the typical example of histopathological images used for breast cancer diagnosis. Both traditional and CNN-based methods are examined and compared based on common histological image datasets used as benchmarks.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

II. Materials and Methods

A. Patients and Whole Slide Images

The whole slide histopathological images acquired from biopsy samples of 90 different patients suspicious for breast cancer were used as the source of image data. These WSIs represent a sub-sample of hematoxylin-eosin stained images of lymph node sections used in the Grand Challenge [5]. The challenge was aimed at discovering the best methods and algorithms for detecting breast cancer metastases. A total of 76 WSIs contained metastases of different sizes whereas the other 14 not presented any pathological changes and were considered to be the norm. An example fragment of a WSI image, as well as the high-resolution picture of its inhomogeneous region, are shown in Fig. 1. It should be noted that histological images of biopsy samples may contain both normal and tumor regions simultaneously what is clearly demonstrated by Fig. 1.

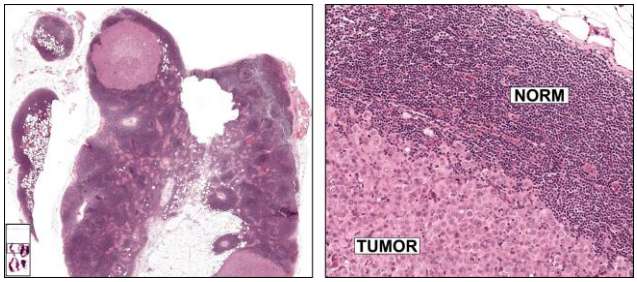

Fig. 1. Example fragment of WSI image and its characteristic region

B. Image datasets

Original professionally-labeled WSI images were partitioned into non-overlapping image sections (image tiles) of 256×256 pixels in size at the highest resolution level that corresponds to the x40 optical microscope magnification. A total of 12000 tiles including 6000 tiles of the norm and 6000 tiles of tumor were randomly sub-sampled from the resultant set of tiles. Examples of the two image classes are given in Fig. 2.

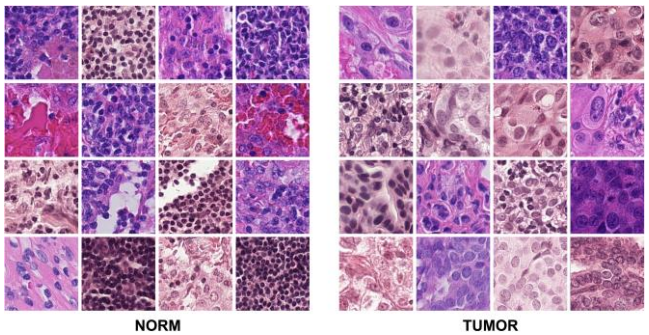

Fig 2. Example image tiles of two classes

The well-balanced train and test image datasets were created following the 70/30 percent proportion. This has resulted in 8400 image tiles included in the training set (4200 tiles of norm plus 4200 tiles representing the tumor) and 3600 images (1800 of norm regions and 1800 of tumor) used for testing. In a similar study [4], we experimentally confirmed that WSI tissue images of each particular patient are holding certain characteristic image patterns (features) which makes them somewhat different from any others. As a result, including image tiles of one single WSI to both training and test sets creates a bias that resulted in an artificial increase of classification accuracy. With this in mind, here, image tiles of any given patient were included in the training or test set only and never in both simultaneously.

For the computational experiments involving the deep learning techniques the image training set was further subdivided into the 5880 training images as such and 2520 validation ones. Again, these particular datasets were well balanced containing exactly 50% of the norm and 50% of images representing tumor regions. The random sub-sampling was preferred on all the occasions where possible.

C. Conventional methods

As usual, the conventional method of binary classification task considered in this study included feature extraction and classification steps. The feature extraction was performed based on color co-occurrence matrices [7]. Given that the hematoxylin-eosin stained histological images are reasonably poor in colors, the original RGB color space was reduced down to the palette of the most common 32 colors. This was accomplished with the help of an adaptive algorithm of reducing color space based on *k*-means clustering as implemented in commonly-known Python PIL library. The inter-pixel spacings were selected to be 1, 2, and 4 pixels. As a result, the co-occurrence image descriptors had the form of 3D arrays of "colors-colors-spacing" type with a dimensionality of 32×32×3.

It is known (and it is very natural) that elements of co-occurrence matrices are highly correlated and therefore they are too redundant to be utilized as image features directly. For this reason, their principal components (PCs) derived with the help of the Principal Component Analysis (PCA) method were used instead. The advantage of such features is that they are compact, linear, and mutually uncorrelated.

The image features were inputted into the Support Vector Machine (SVM) and Random Forests (RF) classifiers. These classifiers were selected because they typically provide competitive results and their software implementations are available broadly. The relatively low computational expenses required by the classifiers allow to subsample given amount of image data from the whole dataset of 8400 training images and repeat the subsampling-training-prediction loop 100 times for obtaining reliable estimates of classification accuracy.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

D. CNN-based methods

In this study, the simple and well-known CNN architecture was employed that consisted of only 3 convolutional + *MaxPooling* layers with 16, 32, and 64 filters respectively (Fig. 3). This is because the assessment of the use of limited image training sets was particularly targeted in this study. In addition, such a decision enables other researchers to easily reproduce results of the computational experiments reported in this paper. The convolutional part of CNN was followed by a fully connected neural network with 512 intermediate nodes.

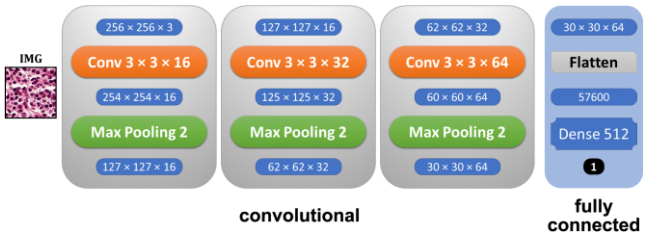

Fig 3. The CNN architecture being employed

Note that despite the simple architecture, under condition of 256×256 pixels of input image size the CNN contains 29,515,809 trainable parameters (weights). It was found that the use of *GlobalMaxPooling* instead of a straightforward *Flattening* of the output of the convolutional part of neural network reduces the number of parameters down to 57,377 with no accountable reduction of the classification accuracy.

E. Experimental arrangements

In order to obtain a relatively complete picture of the influence of training set size on the classification results, the original image training set was gradually reduced from 8400 to 840 images with the step of 10%. In addition, the very-small sub-samples of 5% (420), 2.5% (210), 1.25% (105), and even 1% (84) of original image dataset were also examined where possible. In all the occasions the test set was kept exactly the same and consisted of 3600 images.

The pipeline of computational experiments included the major steps given below.

(a) Initial preparations. They included converting original color RGB images into the reduced paletted representation with 32 colors as well as calculation of color co-occurrence matrices for each of 12000 images.

(b) Creating a data table by way of storing vectorized versions of co-occurrence matrices into 12000 different rows. Performing PCA on the resultant data table for obtaining a concise feature representation of every image involved in the experiments.

(c) Splitting the whole image dataset into the train and test subsets by 70/30 rule. In the case of CNN-based classification (not applicable to conventional SVM and RF) the train set was further split by the same rule into the part used for training as such and the validation.

(d) Cary out the conventional part of classification experiments in 14 steps by way of step-by-step reduction of training set size from the original 8400 images down to 84 ones as described above. At every classification step except for the first one, the training+prediction procedure repeated 100 times on varying image training sets obtained by a random sub-sampling from the original 8400 items. As a result, the total number of training and prediction steps was amounted up to 13*100+1 = 1301. This was to account for the inhomogeneity of original image classes as well as for the variability of images the training set is made of.

(e) The CNN-based experiments were done in a similar manner. However, in this case, one more key parameter came into the way what is the number of training epochs that need to be performed. Also, due to known fluctuations of the training process, the exact measurement of classification accuracy often includes performing a safe, i.e., over-rated number of epochs in order to identify the best one. There are some more control parameters such as random seeds different values of which may lead to slightly different results. These parameters increase potential computational expenses even further. For estimation purposes let us simply suppose that we need only 10 additional exploratory runs due to these factors specific for CNNs. Then the number of repetition loops of type {sub-sampling} {training} {prediction} {adjusting control parameters} increases up to approximately 13,000 what is going beyond the reason.

Thus, in order to make conventional and CNN-based results comparable, at each step of experiments we selected the training dataset that provided the best classification accuracy by SVM classifier and repeat it on exactly the same set of training and test images using CNNs.

III. Results

Results of classification experiments are given below and itemized in the same way as their description presented in the previous section.

(a) Color co-occurrence matrices were computed using a fast algorithm based on indexing arrays that implemented in R language [8]. The elements below the leading diagonal of square-shaped Color-Color slices of resultant 3D arrays were summed up to the corresponding elements situated above the leading

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

diagonal to avoid dependence of results on the rotation and reflection of original images as described in [7].

(b) The image features, i.e., principal components were selected using values of 0.9, 0.95, and 0.98 as thresholds for cumulative variance. These resulted in 21, 221, and 370 components respectively. The value of 0.95 was finally selected as the basic and used in all the computational experiments.

(c) In CNN-based classification with 14 different training set sizes the images for training and validation were selected at random by 70/30 proportion. The amounts of norm and tumor images were kept equivalent in both.

(d) Results of image classification using conventional methods are summarized in Fig. 4. As it can be seen from the top panel of Fig. 4, the difference between the mean accuracy values provided by SVM and RF classifiers is reasonably low with a maximum mutual deviation of about 1%.

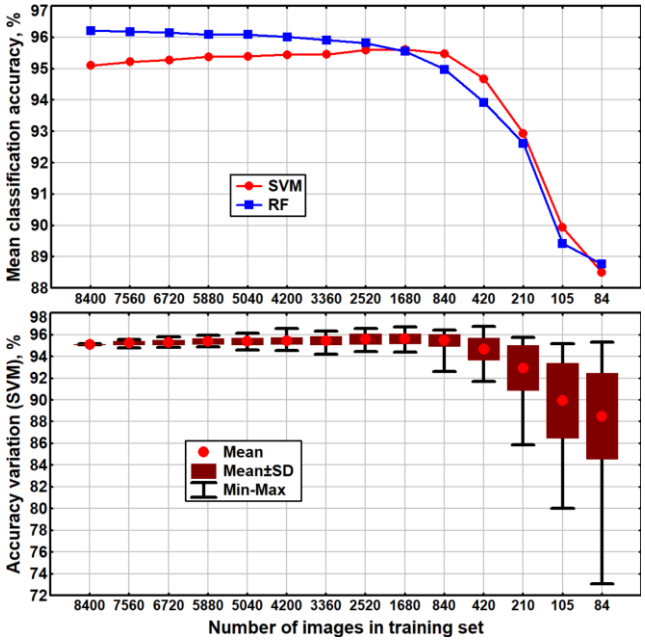

Fig. 4. The mean classification accuracy achieved across 100 replications by SVM and RF methods (top) and its variation in case of SVM (bottom)

Next, the shape of plots suggests that the mean accuracy is keeping almost constant for all training set sizes reduced from 8400 down to 840 images. Then it drops quickly from 95.1% (SVM) and 96.2% (RF) down to 88.5% and 88.8% respectively when reaches the smallest training set of 84 images.

Interestingly, the most fortunate combination of 84 training images among 100 randomly sampled ones for SVM-based classification provided 95.3% of the classification accuracy on the balanced test set consisting of 3600 images. This is slightly better than the worst results on all the repetitions of experiments and all tested training set sizes (see corresponding whiskers of the box-and-whiskers plots of Fig. 4).

The pattern of variability of classification results is somewhat more interesting (see SVM as an example on the bottom panel of Fig. 4). While the training sets remain relatively large, the standard deviation keeping small and ranged from STD=0.172% for 7560 images and going up to STD=0.545% for 840 images. Then it increases significantly and achieves STD=3.946% in the case of 84 images randomly chosen for training. The extreme values ranged even more substantially. For instance, in case of SVM and 84 training images, the classification accuracy varied in 100 repetitions from 73.1% to 95.3%. The described behavior can be explained by the following two reasons:

- the large portions of training images represent better the whole population (general regularity),
- the histological images used in this study are very heterogeneous (see Fig. 2) and vary significantly depending on the patient, biopsy techniques, sample preparation and staining protocols, image acquisition devices used in different hospitals, and some other factors.

(e) Results of CNN-based classification are given in Fig. 5. From a first glance, it becoming clear that results produced by CNN are comparable with the ones obtained using color co-occurrence features followed by SVM and RF (see bars for 8400, 420, and 840 images in the training set). However, once the training set is reduced further, the popular nowadays DL-based approach starts to lose completely against classical methods. This is especially obvious when the CNN results are compared to the ones produced by SVM. For making this fact easier to capture, the bottom panel of Fig. 5 provides two plots that compare results produced by CNN and the maximum accuracy achieved either by SVM or RF classifiers for each step of the experiments.

It is clear that due to the low computational expenses both of them can be comfortably run in parallel and the best result can be taken as the final solution. Note that such results are not surprising at all because it is commonly known that CNNs are hardly usable in the circumstances when only a few tens or hundreds of images are available for training (the use of possible benefits provided by augmentation and other similar techniques should be discussed separately).

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

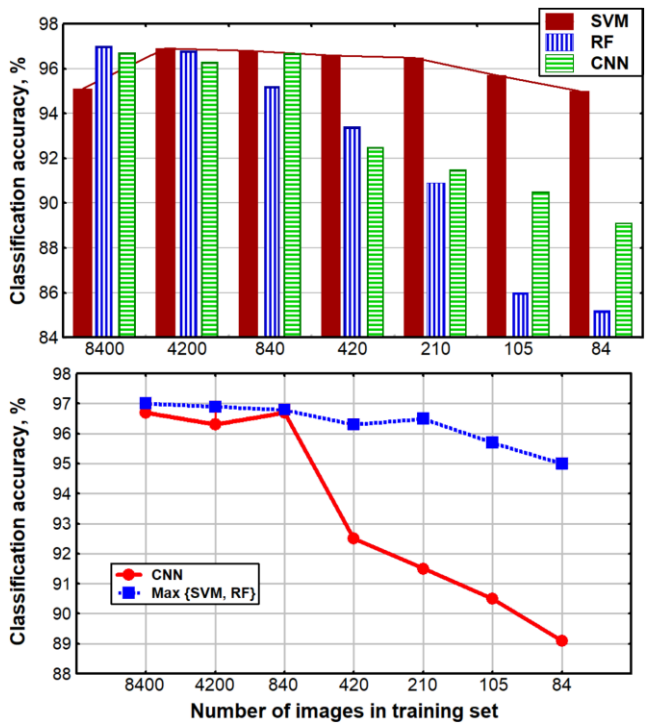

Fig. 5. The best classification accuracy achieved on large (8400–840) and small (420–84) training images by SVM, RF, and CNN methods

VI. Conclusions

Results reported with this study allow drawing the following conclusions.

(1) Conventional and CNN-based methods produce similar classification accuracy on relatively large training sets (from 840 to 8400 images). However, on smaller training sets containing 84-420 images, conventional methods reliably outperform the results demonstrated by CNN.

(2) Under the condition of the high variability of the content of original images and small training sets the classification results may vary substantially depending on the images used for training. For instance, in this particular study, the classification accuracy varied in a wide range from 73.1% to 95.3%. In the case of conventional methods, this problem can be resolved by multiple re-sampling training images and re-running the training for obtaining a reliable estimate of the accuracy. However, with CNNs such a solution can be not feasible due to much higher computational expenses.

(3) The use of recent large and heavy CNN architectures with small image datasets is questionable. However, a separate investigation is necessary for quantitative assessment.

Acknowledgment

The author wishes to thank his employer for providing all the necessary conditions for conducting this exploratory research work.

References

[1] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, R. M. Summers. “A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies with Progress Highlights, and Future Promises,” in Proceedings of the IEEE, vol. 109, no. 5, pp. 820-838, May 2021, doi: 10.1109/JPROC.2021.3054390.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sanchez, “A survey on deep learning in medical image analysis.” Medical Image Analysis, vol. 42, pp. 2017.

[3] F. Aeffner, M.D. Zarella, N. Buchbinder, M. M. Bui, M. R. Goodman et al. “Introduction to digital image analysis in whole-slide imaging: A white paper from the digital pathology association.” Journal of Pathology Informatics, vol. 10, no. 9, 2019.

[4] M. Veta, J. Y. Heng, N. Stathonikos, E. B. Bejnordi, F. Beca, et al. “Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge,” Medical Image Analysis, vol. 54, pp. 111-121, 2019.

[5] B. E. Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, et al. “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer,” Journal of the American Medical Association, vol. 318, no. 22, pp. 2199-2210, 2017.

[6] H. R. Tizhoosh, L. Pantanowitz. Artificial intelligence and digital pathology: Challenges and opportunities. Journal of Pathology Informatics, vol 9, no. 38, 2018.

[7] V. Kovalev and S. Volmer. ”Color co-occurrence descriptors for querying-by-example,” in International Conference on Multimedia Modelling, Oct. 12-15, Lausanne, Switzerland, IEEE Comp. Society Press, pp. 32-38, 1998.

[8] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.