Comparing Deep Learners with Variability Grading for Cancer Detection on Limited Histopathology Dataset

Pedro Furtado
U. Coimbra / CISUC
DEI - Polo II
3030-290 Coimbra, Portugal
+351-239790000
pnf@dei.uc.pt

ABSTRACT

State-of-the-art deep convolution neural networks (CNN) can be applied to various domains, including the grading of cancers in histopathology images, and are most promising approaches. However, it is well-known that CNNs require huge amounts of tagged images and resources to train and work well, and some prior works on cancer grading also achieved top accuracy by analyzing how cancer affects structures, such as cells, in terms of variability of characteristics. The aim of this work is to compare CNN-based classification of medical images with automated analysis of multiple structures. This is done experimentally, by implementing the alternatives and comparing classification accuracy on a public cancer grading dataset. The results show that a well-designed automated analysis of structures improved accuracy by 4% when compared with the best CNN result, showing that it is worth to study further and establish procedures based on that analysis.

CCS Concepts

Applied computing → Imaging

Keywords

Medical imaging; Supervised machine learning; Deep learning; Convolution neural networks; Classification; Cancer detection; Cancer grading

1. INTRODUCTION

Deep learning is a relatively new paradigm that is amazing for its ability to automatically adapt feature extraction parameters and inner knowledge to build accurate classifiers from any kind of data in very different contexts. This allows deep learning systems to be applied off-the-shelf to most problems directly. In image classification, deep learning based in convolution neural networks (CNN) has shown impressive accuracy results. The totally automated way they learn is a major feature, since a CNN is able to adjust the weights of feature extraction filters and of all other layers based on minimization of the error while classifying

training images. Deep learning also has some drawbacks: it relies on a huge number of weights, consequently it also has huge computational and memory costs, training times are long and it requires a huge amount of training images to reach top accuracy. Finally, one of the reasons for it needing so much training data is that it typically needs to train all kinds of “poses” so that it can match new cases with variability well. More “traditional” methods require customized “hard-coded” feature extraction, although the remaining steps (dimensionality reduction and building of classification model) can be applied in a standard manner and adapt automatically to maximize accuracy.

In a way, when comparing CNNs with traditional methods, CNNs extend the error adaptation capability of classifiers to learn the “best” feature extraction as well, while traditional methods require feature extraction to be customized, therefore they cannot “learn the best feature extraction” by themselves.

Today CNNs have caught all the attention in medical imaging, are seen as state-of-the-art methods and have a promising future, as improved alternatives are investigated. However, some past literature also reports top accuracy for the more traditional approaches. For instance, in [1] (publicly available dataset that can be tested by anyone) we obtained accuracies of 97.5% with a random forest classifier. The approach detects cancer based on measurements of characteristics of the cell nuclei present in the image. In that case the gathering of the data required semi-manual work to choose regions and measurements, but the accuracy results are impressive and at least comparable to CNN results. Therefore we can state the hypothesis that, instead of only brute force low-level feature extraction (e.g. it is not enough just to “throw in” SIFT [10][11], LBP [12] and bag-of-features or so to guarantee good accuracy), generalizations of prior work to identify and characterize structures adequately and to characterize how they vary may be competitive.

In this work we develop CNN variants for the problem of classifying the degree of malignancy, applying different architectures, plus testing data augmentation and the patches-based improvement proposed in [9], to try to maximize CNN accuracy. We compare those alternatives to a more “traditional” approach that identifies structures in images and characterizes their variability to grade cancers (variability grading VG or Var). We take a specific public dataset (Mytos Atypia [3]) and its problem of grading cancers as benign, in-doubt or malignant, to test the hypothesis that a well-designed procedure might even outperform state-of-the-art CNNs for this specific problem. In our experiments VG achieved 86% accuracy, against 82.5% of the best CNN result. To achieve this accuracy in VG we had to fine-tune the first steps of segmentation and structure elicitation, while CNNs are applied directly to any problem without any need to further customize.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ICBBT'19, May 29–31, 2019, Stockholm, Sweden

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6231-3/19/05...\$15.00

DOI:<https://doi.org/10.1145/3340074.3340083>

The paper is structured as follows: section 2 reviews related work. Next three sections review convolution neural networks (CNN) and how variability grading (VG) works. Afterwards, we compare the two paradigms for the cancer grading problem of Mytos Atypia dataset [3]. Finally, we analyze the results and reach conclusions regarding relevant future work.

2. STATE OF ART

Almost since their inception, Deep Convolution Neural Networks (CNN) have been applied to medical imaging problems. They have been used in experiments to detect cancer and other pathologies, and to classify types of cancer. In [14] the authors apply deep learning over a training set of 571 breast tumors, creating image-based classifiers for tumor grading. Histologic image analysis with deep learning distinguished low-intermediate vs. high tumor grade (82% accuracy), ER status (84% accuracy), Basal-like vs. non-Basal-like (77% accuracy), Ductal vs. Lobular (94% accuracy), and high vs. low-medium ROR-PT score (75% accuracy). The authors in [9] classified benign/malign cancer on the BreakHis dataset [15] by following a strategy of generating image patches and training the CNN with those patches. They achieved best accuracy with 100 random patches sized 64×64 , with accuracies of various alternatives between 80.8 ± 3.0 and $89.6 \pm 6.5\%$ for benign/malign decisions. Other authors [16][17] also obtained high accuracies for the same dataset.

Some “traditional” cancer detection approaches reported accuracies that are also quite high. As examples, in [1] features describe characteristics of the cell nuclei present in the image and resulted in accuracies of 96% to 97.5% using repeated 10-fold cross-validation. The common feature of most successful traditional approaches was that they would usually characterize specific structures in the images and see how those change when there are lesions or cancer features present. This has led us before to work an automated approaches to detect and characterize those structures [18]. Many other authors explored traditional (non-CNN) mechanisms to detect and grade cancers. Just to name a few, Loukas et al. [6] pre-selected 65 regions of interest from 60 images of breast cancer tissue to grade cancer into 3 degrees of malignancy (I-III) from extracted set of features. The neural network classifier achieved 90% accuracy. [8] achieved 80 to 95% accuracy grading prostate and breast cancers using three scales of low-level pixel values, high-level information based on relationships, and structural constraints. Finally, [7] used multi-wavelet grading of prostate pathological images, achieving a precision of 97% discriminating prostate tissue cancer/ no cancer and 81% for different degrees of low and high Gleason.

3. CONVOLUTION NEURAL NETWORKS, PATCHING AND AUGMENTATION

Convolution neural networks are deep neural networks (deep means with “many” hidden layers) with two main “stages”, a stage made of a sequence of convolution layers, and a stage made of fully connected layers. The first convolution layer takes the image itself as input and outputs feature maps. The successive convolution layers transform feature maps. For each convolution layer the programmer specifies a number of filters to use (e.g. around 30 is a typical number), a filter size (e.g. a 3×3 matrix) and a stride. Each filter is used to compute a convolution operation on the input. A “filter” is a (square) matrix, and the convolution [4] operates a sum-product of the filter cells (weights) by the pixels in the vicinity centered at the pixel. This operation captures local characteristics of the input, since it computes over the close neighborhood of each pixel.

By adjusting the matrix cell values (weights of the filter), different details are captured, and the CNN learning algorithm is able to adjust these weights automatically to capture properties that best determine the class of images (e.g. the cancer grade). Posterior convolution layers will be abstracting successively larger regions. The fully-connected neural network at the last stage takes as input a feature vector that is a flattened feature map after the last convolution layer, and outputs a category classifier. The category classifier has as many outputs as there are classes, each displaying a probability of the input image being of that class (sum is 1). A crucial innovation in CNNs is that the learning algorithm – backpropagation – adjusts weights not only over the fully-connected neural network, but also over the convolution layers (filter weights), effectively learning how to extract features and how to abstract them for best accuracy. Training iterates a set of epochs over a dataset of images, at each training step it adjusts a fraction of the error in classification based on a learning rate. As described in [5], successive CNN architectures (AlexNet, VGG, GoogLeNet, Inception-V3 and ResNet) introduced specific modifications that result in improved accuracy in most image classification tasks. The number of weights is quite large in a CNN, reaching 1.4 billion in InceptionV3 and 4 billions in ResNet-101, resulting in high computation, memory and training costs.

Three common issues discussed in CNNs are the need of a huge training dataset to learn adequately, the possible advantage of feeding image details instead of the whole image, and data augmentation. Patching [9] refers to dividing the images of the dataset into multiple smaller images (e.g. 64×64 patches) to be fed to the CNN. Patches can better individualize structures and may provide a convenient degree of detail about regional structures, while also augmenting the dataset by dividing an image into many patches. A stride (e.g. start a patch every 10 pixels) allows to scan the whole image while obtaining the patches with possibly overlapping parts. Data augmentation is a different technique designed to increase the size and variability of the training dataset based on simple operations such as scaling, rotation, shearing or translation. It can contribute to increase variability of training images, resulting in more training images and more capability to deal with variability.

4. VARIABILITY GRADING

As we mentioned earlier, successful prior work on more traditional image classification pipelines for the cancer problem picked instances of specific structures and measured how those structures vary with cancer using a classifier (e.g. [1] achieves 96-97.5% accuracy on the cancer-no cancer problem). By completely automating the procedure of obtaining regions representing specific structures and generalizing that to a set of pre-defined structures, it is possible to characterize more completely how those structures are affected with degree of cancer. This requires segmentation, identification and characterization of pre-defined image structures. The approach, which is described in more detail in [18], segments and identifies pre-defined object types, obtains statistical descriptions of those types and how they vary with cancer grades. The following steps were implemented:

(1) Segmentation and identification – segmentation separates images into histopathology structures (e.g. mammarian cells, leucocytes, vacuoles, stroma, ducts, others). This step must be developed and tested separately to ensure that structures are well separated and identified correctly. We applied hardcoded thresholding based on intervals of intensity, followed by postprocessing with analysis of connected regions, merging of

outliers (few pixels in large regions) and structure identification based on intensity interval and region size. This approach worked well to separate the basic parts quickly, while still identifying structures with high accuracy. Future work should deal with more complex structures;

(2) Determination of statistical features – Generalizing the approach in [1], each type of structure identified in step 1 must be described statistically in terms of geometry, shape, contour, counts, texture and color;

(3) Feature reduction - given this characterization of all training cases and the corresponding class (e.g. cancer grade), a dimensionality reduction step is ran, keeping only a pre-defined number of features;

(4) Model building – model building can use any standard classifier algorithm, but given the complexity and nonlinearity of the cancer grading problem, we found that the best classifiers in our experimental work were Random Forests and Multilayer Perceptron.

5. EXPERIMENTAL SETUP

The Mytos Atypia dataset [3] is an annotated set of breast cancer biopsy slides. The dataset concerns detection of mitosis and evaluation of nuclear atypia score in breast cancer histological images. As part of the dataset, a score for nuclear atypia is provided for each of the 20x frames (grades 1=benign, 2=in-doubt, 3=malignant). The score was assigned by two pathologists plus, when in conflict, a third pathologist. The task is to create a model and guess the scores for test images. We compare state-of-the-art CNNs (Resnet-101 and Inception-V3) to VG. To maximize CNN accuracy, we produce three main variations of the CNN setups: 1) Base dataset: after balancing and dividing images into 4 quadrants, the base dataset has 390 training images, 60 validation and 60 independent test images; 2) Augmented dataset: For CNN-based approaches, data augmentation is applied based on rotations, translations, shear, to generate a large number of distinct batches for training; 3) Patched dataset: the approach in [9] is followed, creating a large number of patches (128x128 pixels) from all images. This resulted in 7560 training and 1890 test images. We verified the code by running it for BreakHis dataset [9], achieving accuracy of 96.3% cancer/no cancer.

The experiments were all setup and ran in Matlab 2018a. The CNNs tested were Matlab 2018 implementations of state-of-the-art CNNs, including InceptionV3 and Resnet-101, we report the best results obtained. InceptionV3 and Resnet-101 are pre-trained implementations that were trained on more than a million images from the ImageNet database to classify images into 1000 object categories. Resnet-101 is 101 layers deep Resnet architecture [19], Inception-v3 is 48 layers deep Inception architecture [20]. In all cases, transfer learning was used, by replacing the fully connected layer for one that classifies three degrees of atypia of the Mytos classification alternatives, depending on the experiment, and with a softmax activation function to output probabilities of each class. The maximum number of epochs was configured to 600, with initial learning rate 0.05, a stochastic gradient descent optimizer and validation frequency of 5 iterations.

Experiments were ran in a PC with an intel core i5 at 3.4 GHz, with 16 GB RAM and an SSD disk of 1TB running windows. The PC had an NVIDIA GeForce GTX 1070 GPU installed (Pascal architecture with 1920 cores, 8 GB GDDR5, 8 Gbps memory speed). The metrics were Accuracy or $acc = \text{nr. correct} / \text{nr. of}$

classifications and $\text{Precision}(Ci) \text{ or } p(ci) = \text{nr. correctly classified as class } Ci / \text{nr. classified as class } Ci$.

The following cancer grading approaches were tested: (a) Var = variability-based (VG); Iv3 = Inception-V3; Iv3 augment = Inception-V3 with data augmentation; R101 = Resnet-101, R101patch= Resnet-101 with patching; Hexpert = grading by human experts (medical doctors).

In all approaches an independent test dataset was obtained by separating a randomly obtained fraction of 20% from the full dataset. A validation dataset was also extracted for the training of the DNN approaches by separating another 20% of the data. The validation data was used by the approaches during iterations to assess accuracy. After the approaches converged over a sufficient number of training epochs (300), the training would be over and the test set was used to obtain the accuracy results reported here.

For illustration purposes, Figure 3 (at the end of the paper) shows an example of a plot of the run of a DNN. It shows the validation accuracy (black) and the accuracy for each image (blue) during training. The particular example shown is for inception V3 with data augmentation, the blue curve converges to nearly 100% as the DNN learns, but with significant per-image oscillations not seen when augmentation is not used (rotations, shearing, cropping and so on make classification potentially more challenging).

5.1 Results

Figure 1 compares accuracy of the variability-based approach VG with that of DNN alternatives. These results were obtained after running the various alternative DNN approaches described in the experimental setup section (InceptionV3 and Resnet-101 with normal dataset, augmented dataset and also patched dataset). For comparison purposes, we can see that human experts (HExpert) agreed on 84% of the scores (fraction of times the two pathologists agreed), a value that is very close to accuracy of the best-scoring approaches.

Figure 2 shows per-class precision, revealing how the classification accuracy boils down into classes. Note for instance that a certain approach might exhibit a very good precision for some class(es) but at the expense of other class(es).

Figure 1. Accuracy comparison

Figure 2. Per-class precision comparison

5.2 Analysis and Discussion

All tested approaches achieved good accuracy (73-86%). But the variability-based approach was most accurate, even compared with CNNs with data augmentation and patching. Notice also that Iv3 (Inception-V3) obtained 100% precision classifying cancer cases as cancer correctly (Figure 2), but this was at the expense of also classifying as cancer cases that were not cancer (much lower precision for no-cancer and in-doubt). In other words, it had bad specificity. These comparative results confirm the ability already shown in approaches such as [1] that analyze how structures vary with and without cancer. It means that a well-designed technique based on prior individualization of regions, statistical characterization of those regions and characterization of variability with respect to classes can be competitive, possibly even better than state-of-the-art deep learning methods, at least for some problems.

6. CONCLUSIONS AND FUTURE CHALLENGES

For the problem of detecting degree of malignancy from histopathology images from limited training data, we compared deep learning to a variability-based approach that characterizes each type of object and detects variability in their characteristics. After describing the approaches, we engaged in experimental validation. The results reveal that carefully designed variability-based techniques can achieve comparable or even higher accuracy than state-of-the-art deep learning (86% vs 82.5% in this case). On the other hand, DNNs can be applied off-the-shelf and process many thousands of images to create a good classifier effortlessly, while a variability-based approach must be hand-tailored and is dependent on the problem, since it is necessary to tune segmentation and feature extraction to fit specific types of objects in images. In the future we wish to analyze how the two seemingly-different approaches (DNNs and variability analysis) can be integrated.

Figure 3. Inception-V3, data augmentation, 300 epochs

7. REFERENCES

- [1] Wolberg W.H., W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology, *Human Pathology*, 26:792--796, 1995.
- [2] Haykin, Simon S. *Neural Networks: A Comprehensive Foundation*. 1994.
- [3] Mytos Atypia ICPR-2014 contest using breast cancer histological images. URL: <https://mitos-atypia-14.grand-challenge.org/> [accessed on 2017-01-01].
- [4] R. C. Gonzalez and P. Wintz, Addison-Wesley, *Digital Image Processing (Second Edition)*, 1987, 503.
- [5] Alom, Md.Zahangir, et al. The history began from AlexNet: A comprehensive Survey on deep learning approaches. ArXiv Preprint ArXiv: 1803.01164, 2018.
- [6] Loukas, C., S. Kostopoulos and D. Cavouras. Breast cancer characterization based on image classification of tissue sections visualized under low magnification. *Computational and Mathematical Methods in Medicine*, 2013.
- [7] K. Jafari-Khouzani and H. Soltanian-Zadeh. Multiwavelet grading of pathological images of prostate. *IEEE Transactions on Biomedical Engineering*, 50:pp. 697-704, June 2013.
- [8] Naik, Shivang, et al. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology." *Biomedical Imaging: From Nano to Macro*, 2008. ISBI 2008. 5th IEEE International Symposium on. IEEE, 2008.
- [9] Spanhol, Fabio Alexandre, et al. Breast cancer histopathological image classification using convolutional neural networks. *Neural Networks (IJCNN)*, 2016, pp. 2560-2567.
- [10] U.S. Patent 6,711,293, Method and apparatus for identifying scale invariant features in an image and use of same for locating an object in an image, David Lowe's patent for the SIFT algorithm, March 23, 2004.
- [11] Lowe, David G., 1999. Object recognition from local scale-invariant features (PDF). *Proceedings of the International Conference on Computer Vision*. 2. pp. 1150-1157. doi:10.1109/ICCV.1999.790410.
- [12] L. Wang and DC. He, 1990, Texture classification using texture spectrum, *Pattern Recognition*, Vol. 23, No. 8, pp. 905 - 910.
- [13] Couture, H. D., Williams, L. A., Geradts, J., Nyante, S. J., Butler, E. N., Marron, J. S., ... & Niethammer, M., 2018. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *NPI breast cancer*, 4(1), 30.
- [14] Wei, B., Han, Z., He, X., & Yin, Y., Apr 2017, Deep learning model based breast cancer histopathological image classification. In *Cloud Computing and Big Data Analysis (ICCCBDA)*, 2017 IEEE 2nd International Conference on (pp. 348-353). IEEE.
- [15] Spanhol F A, Oliveira L S, Petitje C, et al. A dataset for breast cancer histopathological image classification, *[J]. IEEE Transactions on Biomedical Engineering*, 2016, 63(7): 1455-1462.

- [16] Wang, D., Khosla, A., Gargeya, R., Irshad, H. & Beck, A. H. Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606. 05718, 2016.
- [17] Motlagh, N. H., Jannesary, M., Aboulkheyr, H., Khosravi, P., Elemento, O., Totonchi, M., & Hajirasouliha, I., 2018. Breast Cancer Histopathological Image Classification: A Deep Learning Approach. bioRxiv, 242818.
- [18] Furtado, P. Objects characterization-based approach to enhance detection of degree of malignancy in breast cancer histopathology. In SPIE Medical Imaging Conference, San Diego, USA, 2019.
- [19] L. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z., 2016. Rethinking the Inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826.
- [20] He, K., Zhang, X., Ren, S., & Sun, J., 2016. Deep residual learning for image recognition, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.