ScienceDirect

Medical Image Analysis

Volume 58, December 2019, 101544

Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology

David Tellez [a], Geert Litjens [a], Péter Bándi [a], Wouter Bulten [a], John-Melle Bokhorst [a], Francesco Ciompi [a], Jeroen van der Laak [a b]

https://doi.org/10.1016/j.media.2019.101544

Get rights and content

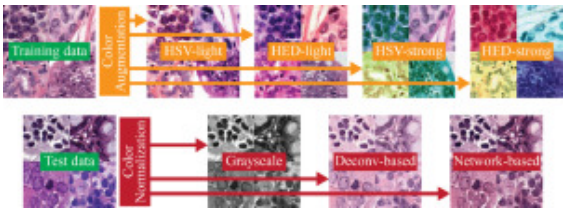
Highlights

- Combining color augmentation and color normalization achieves the best performance.
- Using color augmentation is essential to reduce the generalization error.
- The specific type of color augmentation (HSV or HED) and its strength is irrelevant.
- Color normalization based on neural networks is superior to more traditional methods.
- Skip color normalization to save computational resources at a negligible performance cost.

Abstract

Stain variation is a phenomenon observed when distinct pathology laboratories stain tissue slides that exhibit similar but not identical color appearance. Due to this color shift between laboratories, convolutional neural networks (CNNs) trained with images from one lab often underperform on unseen images from the other lab. Several techniques have been proposed to reduce the generalization error, mainly grouped into two categories: stain color augmentation and stain color normalization. The former simulates a wide variety of realistic stain variations during training, producing stain-invariant CNNs. The latter aims to match training and test color distributions in order to reduce stain variation. For the first time, we compared some of these techniques and quantified their effect on CNN classification performance using a heterogeneous dataset of hematoxylin and eosin histopathology images from 4 organs and 9 pathology laboratories. Additionally, we propose a novel unsupervised method to perform stain color normalization using a neural network. Based on our experimental results, we provide practical guidelines on how to use stain color augmentation and stain color normalization in future computational pathology applications.

Graphical abstract



Download: Download high-res image (306KB)

Download: Download full-size image

Introduction

Computational pathology aims at developing machine learning based tools to automate and streamline the analysis of whole-slide images (WSI), i.e. high-definition images of histological tissue sections. These sections consist of thin slices of tissue that are stained with different dyes so that tissue architecture becomes visible under the microscope. In this study, we focus on hematoxylin and eosin (H&E), the most widely used staining worldwide. It highlights cell nuclei in blue color (hematoxylin), and cytoplasm, connective tissue and muscle in various shades of pink (eosin). The eventual color distribution of the WSI depends on multiple steps of the staining process, resulting in slightly different color distributions depending on the laboratory where the sections were processed, see Fig. 1 for examples of H&E stain variation. This inter-center stain variation hampers the performance of machine learning algorithms used for automatic WSI

analysis. Algorithms that were trained with images originated from a single pathology laboratory often underperform when applied to images from a different center, including state-of-the-art methods based on convolutional neural networks (CNNs)(Goodfellow, et al., 2016, Komura, Ishikawa, 2018, Veta, Heng, Stathonikos, Bejnordi, Beca, Wollmann, Rohr, Shah, Wang, Rousson, et al., 2019, Sirinukunwattana, et al., 2017). Existing solutions to reduce the generalization error in this setting can be categorized into two groups: (1)*stain color augmentation*, and (2)*stain color normalization*.

Stain color augmentation, and more generally data augmentation, has been proposed as a method to reduce CNN generalization error by simulating realistic variations of the training data. These artificial variations are hand-engineered to mimic the appearance of future test samples that deviate from the training manifold. Previous work on data augmentation for computational pathology has defined two main groups of augmentation techniques: (1) morphological and (2) color transformations(Liu, Gadepalli, Norouzi, Dahl, Kohlberger, Boyko, Venugopalan, Timofeev, Nelson, Corrado, et al., Tellez, Balkenhol, Otte-Höller, van de Loo, Vogels, Bult, Wauters, Vreuls, Mol, Karssemeijer, et al., 2018). Morphological augmentation spans from simple techniques such as 90º rotations, vertical and horizontal mirroring, or image scaling; to more advanced methods like elastic deformation(Simard et al., 2003), additive Gaussian noise, and Gaussian blurring. The common denominator among these transformations is the fact that only the morphology of the underlying image is modified and not the color appearance, e.g. Gaussian blurring simulates out of focus artifacts which is a common issue encountered with WSI scanners. Conversely, color augmentation leaves morphological features intact and focuses on simulating stain color variations instead. Common color augmentation techniques borrowed from Computer Vision include brightness, contrast and hue perturbations. Recently, researchers have proposed other approaches more tailored to mimic specific H&E stain variations, e.g. by perturbing the images directly in the H&E color space(Tellez et al., 2018), or perturbing the principal components of the pixel values(Bug et al., 2017).

Stain color normalization reduces stain variation by matching the color distribution of the training and test images. Traditional approaches try to normalize the color space by estimating a color deconvolution matrix that allows identifying the underlying stains(Reinhard, Adhikhmin, Gooch, Shirley, 2001, Macenko, et al., 2009). More recent methods use machine learning algorithms to detect certain morphological structures, e.g. cell nuclei, that are associated with certain stains, improving the result of the normalization process(Khan, et al., 2014, Bejnordi, Litjens, Timofeeva, Otte-Höller, Homeyer, Karssemeijer, van der Laak, 2016). Deep generative models, i.e. variational autoencoders and generative adversarial networks(Kingma, Welling, 2013, Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, Bengio, 2014), have been used to generate new image samples that match the template data manifold(Cho, Lim, Choi, Min, 2017, Zanjani, Zinger, Bejnordi, van der Laak, de With, 2018). Moreover, color normalization has been formulated as a style transfer task where the style is defined as the color distribution produced by a particular lab(Bug et al., 2017). However, despite their success and widespread adoption as a preprocessing tool

in a variety of computational pathology applications(Clarke, Treanor, 2017, Albarqouni, Baur, Achilles, Belagiannis, Demirci, Navab, 2016, Janowczyk, Basavanhally, Madabhushi, 2017, Bándi, Geessink, Manson, van Dijk, Balkenhol, Hermsen, Bejnordi, Lee, Paeng, Zhong, et al., 2019), they are not always effective and can produce images with color distributions that deviate from the desired color template. In this study, we propose a novel unsupervised approach that leverages the power of deep learning to solve the problem of stain normalization. We reformulate the problem of stain normalization as an image-to-image translation task and train a neural network to solve it. We do so by feeding the network with heavily augmented H&E images and training the model to reconstruct the original image without augmentation. By learning to remove this color variation, the network effectively learns to perform *stain color normalization* in unseen images whose color distribution deviates from that of the training set.

Despite the wide adoption of *stain color augmentation* and *stain color normalization* in the field of computational pathology, the effects on performance of these techniques have not been systematically evaluated. Existing literature focuses on particular applications, and does not quantify the relationship between these techniques and CNN performance(Komura, Ishikawa, 2018, Wang, Foran, Ren, Zhong, Kim, Qi, 2015, Zhu, Zhang, Liu, Metaxas, 2014, Veta, Heng, Stathonikos, Bejnordi, Beca, Wollmann, Rohr, Shah, Wang, Rousson, et al., 2019). In this study, we aim to overcome this limitation by comparing these techniques across four representative applications including multicenter data. We selected four patch-based classification tasks where a CNN was trained with data from a single center only, and evaluated in unseen data from multiple external pathology laboratories. We chose four relevant applications from the literature: (1) detecting the presence of mitotic figures in breast tissue(Tellez et al., 2018); (2) detecting the presence of tumor metastases in breast lymph node tissue(Bándi et al., 2019); (3) detecting the presence of epithelial cells in prostate tissue(Bulten et al., 2019); and (4) distinguishing among 9 tissue classes in colorectal cancer (CRC) tissue(Ciompi et al., 2017). All test datasets presented a substantial and challenging stain color deviation from the training set, as can be seen in Fig. 1. We trained a series of CNN classifiers following an identical training protocol while varying the *stain color normalization* and *stain color augmentation* techniques used during training. This thorough evaluation allowed us to establish a ranking among the methods and measure relative performance improvements among them.

Our contributions can be summarized as follows:

- We systematically evaluated several well-known *stain color augmentation* and *stain color normalization* algorithms in order to quantify their effects on CNN classification performance.
- We conducted the previous evaluation using data from a total of 9 different centers spanning 4 relevant classification tasks: mitosis detection, tumor metastasis detection in lymph nodes, prostate epithelium detection, and multiclass colorectal cancer tissue classification.

- We formulated the problem of *stain color normalization* as an unsupervised image-to-image translation task and trained a neural network to solve it.

The paper is organized as follows. Sections 2 and 3 describe the materials and methods thoroughly. Experimental results are explained in Section 4, followed by Sections 5 and 6 where the discussion and final conclusion are stated.

Section snippets

Materials

We collected data from a variety of pathology laboratories for four different applications. In all cases, we used images from the Radboud University Medical Centre (Radboudumc or *rumc*) exclusively to train the models for each of the four classification tasks. Images from the remaining centers were used for testing purposes only. We considered RGB patches of 128x128 pixels extracted from annotated regions. Examples of these patches are shown in Fig. 1. The following sections describe each of the …

Methods

In this study, we evaluated the effect in classification performance of several methods for *stain color augmentation* and *stain color normalization*. This section describes these methods. …

Experimental results

We conducted a series of experiments in order to quantify the impact in performance of the different *stain color augmentation* and *stain color normalization* methods introduced in the previous section across four different classification tasks. We trained a CNN classifier for each combination of organ, color normalization and data augmentation method under consideration. In the case of *grayscale* normalization, we only tested *basic, morphology* and *BC* augmentation techniques. We conducted 152 …

Discussion

Our experimental results indicate that *stain color augmentation* improved classification performance drastically by increasing the CNN's ability to generalize to unseen stain variations. This was true for most of the experiments regardless of the type of *stain color normalization* technique used. Moreover, we found *HSV* and *HED* color transformations to be the key ingredients to improve

performance since removing them, i.e. using *BC* augmentation, yielded a lower AUC under all circumstances; …

Conclusion

For the first time, we quantified the effect of *stain color augmentation* and *stain color normalization* in classification performance across four relevant computational pathology applications using data from 9 different centers. Based on our empirical evaluation, we found that any type of *stain color augmentation*, i.e. *HSV* or *HED* transformation, should always be used. In addition, color augmentation can be combined with neural network based *stain color normalization* to achieve a more robust …

Declaration of Competing Interest

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript: …

Acknowledgments

This study was supported by a Junior Researcher grant from the Radboud Institute of Health Sciences (RIHS), Nijmegen, The Netherlands; a grant from the Dutch Cancer Society (KUN 2015–7970); and another grant from the Dutch Cancer Society and the Alpe d'HuZes fund (KUN 2014–7032); this project has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825292. The authors would like to thank Dr. Babak Ehteshami Bejnordi for …

Recommended articles

Research data for this article

Open Data

Supplementary Data S1
(XLSX, 119KB)

References (37)

A. Gertych *et al.*
Machine learning approaches to analyze histological images of tissues from radical prostatectomies
Comput. Med. Imaging Graph. (2015)

J.N. Kather *et al.*
Multi-class texture analysis in colorectal cancer histology
Sci. Rep. (2016)

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P....

J.T. Springenberg *et al.*
Striving for simplicity: The all convolutional net
Proceedings of the International Conference on Learning Representations(2014)

S. Van der Walt *et al.*
Scikit-image: image processing in python
PeerJ (2014)

S. Albarqouni *et al.*
Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images
IEEE Trans. Med. Imaging (2016)

P. Bándi *et al.*
From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge
IEEE Trans. Med. Imaging (2019)

B.E. Bejnordi *et al.*
Stain specific standardization of whole-slide histopathological images
IEEE Trans. Med. Imaging (2016)

B.E. Bejnordi *et al.*

Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer

JAMA (2017)

D. Bug *et al.*

Context-based normalization of histological stains using deep convolutional features

Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (2017)

View more references

Cited by (433)

Transformer-based unsupervised contrastive learning for histopathological image classification

2022, Medical Image Analysis

Show abstract

Deep neural network models for computational histopathology: A survey

2021, Medical Image Analysis

Show abstract

Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study

2021, Lancet Oncology

Show abstract

The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis

2021, Computers in Biology and Medicine

Show abstract

Deep learning in histopathology: the path to the clinic ↗

2021, Nature Medicine

WILDS: A Benchmark of in-the-Wild Distribution Shifts ↗

2021, Proceedings of Machine Learning Research

View all citing articles on Scopus ↗

View full text