The effect of blurring on lung cancer subtype classification accuracy of convolutional neural networks

Tejal Nair, Ali Foroughi pour, Jeffrey H. Chuang
The Jackson Laboratory for Genomic Medicine
Farmington, CT, USA
{tejal.nair,ali.foroughipour,jeff.chuang}@jax.org

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313192

***Abstract*—Deep learning models are extensively used for analyzing hematoxylin and eosin stained whole slide images. They enjoy high prediction accuracies, but may suffer large performance drops when applied to out-of-sample data. Here we systematically investigate how resolution differences between train and test sets may affect lung cancer subtype predictions from whole slide images using a transfer learning model based on the Inception V3 network. We observe models trained on blurred images perform well when applied to test sets with similar blurrings, but suffer poor predictions when applied to images with large blurring differences. In particular, we observed low area under curve values when models trained on blurred images were applied to non-blurred images.**

***Index Terms*—deep learning, convolutional neural networks, digital pathology, image analysis**

I. Introduction

Deep learning has become a popular methodology for analyzing cancer whole slide images. It has been used to study pan-cancer morphological features in hematoxylin and eosin stained pathology images [1]. While deep learning models tend to perform well on hold-out test sets, they may suffer from large performance drops when applied to external or out-of-sample datasets (see [2], [3] for examples). This suggests a deep learning model may be dataset specific or sensitive to dataset specific artifacts. The effect of artifacts such as blurring, noise, and lossy image compressions on classification of natural images is studied in [4]. Additionally, neural networks can encode features due to various blurring modalities, and detect the blur-specific artifacts in natural images [5]. Sensitivity of deep learning models to data artifacts and quality issues in the context of automated pathology is underappreciated in the literature. Here, we investigate if the inception V3 convolutional neural network can differentiate lung cancer subtypes from blurred whole slide images. Lung cancer subtype classification is an interesting case example as it requires expert pathology knowledge for reliable subtype detection [6]. The need for expert knowledge suggests morphological features that differ across lung cancer subtypes may be weak. Stability and reliability of neural networks in separating robust and biologically meaningful morphological features from image artifacts is an important issue, and a major concern in translational lung cancer research [6].

We evaluated how classification accuracy decreases as slides are blurred using a sequence of Gaussian kernels, increasing the blurring standard deviation from 0 (no blurring) to 5 in steps of 0.25 pixels. Fresh frozen lung adenocarcinoma and lung squamous cell carcinoma whole slide images were used. The networks showed strong robustness to blurred images when train and test sets were blurred similarly: the area under curve (AUC) averaged over 100 train-test splits was 0.92 and 0.90 for non-blurred slides and slides blurred using a Gaussian kernel with a standard deviation of 5, respectively. On the other hand, the AUC decreased drastically when train and test slides were blurred to a different extent: the network trained on non-blurred images had an AUC of 0.57 on slides blurred with a kernel standard deviation of 5. Similarly, the network trained on slides blurred with a kernel standard deviation of 5 had an AUC of 0.63 on non-blurred images. Similar relations were observed for other blurring standard deviations. For example, the network trained on images blurred with a kernel standard deviation of 2.5 achieved AUCs of 0.92, 0.62, and 0.76 on slides blurred with kernel standard deviations of 2.5, 0, and 5, respectively. The results suggest blurring differences between train and test sets might have a large impact on accuracy. Therefore, it would not be surprising for a network to suffer low prediction accuracy on datasets with different resolutions, or blurred regions of a dataset where such regions comprise only a small portion of data.

II. Methods

A. Data Acquisition

The fresh-frozen 20X hematoxilyn and eosin stained slides of the lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) of the cancer genome atlas (TCGA) were used in the study. We borrowed the pre-processed tile level images of [1]. Here, we briefly describe the processing steps of [1]: Whole slide images were tiled using non-overlapping 512-by-512 (size in pixels) tiles. Otsu's thresholding was used to separate tissue from background, and tiles with less than 50% tissue were removed. The tiles with sufficient tissue were saved. The borrowed curated data contained 819 LUAD and 753 LUSC slides.

JHC acknowledges support from NCI grant R01CA230031. Copyright: 978-1-7281-6215-7/20/$31.00 ©2020 IEEE.

Authorized licensed use limited to: The Ohio State University. Downloaded on April 09,2021 at 14:53:44 UTC from IEEE Xplore. Restrictions apply.

B. Blurred Image Construction

The tiles were blurred by 21 Gaussian kernels, with standard deviations ranging from 0 pixels to 5 pixels in steps of 0.25 pixels. Note zero blurring denotes the original images without any blurring performed. The Gaussian blurring step was performed using the `open cv` package in python. The images were then passed through the Inception V3 network using the `TensorFlow 2.2` package. `Keras` API was used to implement this pipeline. The 2048 global average pooled features were then saved.

C. Slide Representation and Classification Algorithm

Slides are associated with different number of tiles. A typical approach is to train and predict labels at a tile level, and combine tile-level predictions to obtain slide-level predictions as a post-processing step (see [1] for examples). Here we take a different approach. The median value of each of the 2048 Inception V3 features is used to encode each slide by a 2048-dimensional feature vector. Initial evaluations suggest such approach results in higher AUCs for blurred slides, and was hence adopted as the methodology for the current study.

The Inception V3 architecture trained on image-net data directly applies the globally averaged features to the classification layer with softmax activation. This corresponds to logistic regression in a binary classification task. As each slide corresponds to a 2048 dimensional vector, the feature matrix (the matrix in which each row is a sample point and each column is a feature) was small enough to be fully loaded into memory. Therefore, instead of a batch-based approach we opted for the classical approach of training a logisitc regression model, where model parameters are updated using the full set of trainng data. We used the `sklearn` package to implement the classification step. We used the `saga` optimizer with elastic net penalty, $C = 1000$, and $l1$ ratio was set to 0.5. No hyper-parameter optimization was performed.

III. Results

A. Data Visualization

Figure 1 provides examples of an LUAD tile blurred with various kernel standard deviations. Note blurring with a standard deviation of 5 heavily degrades the image quality. Similar quality reductions and blurring impacts was observed among LUSC tiles.

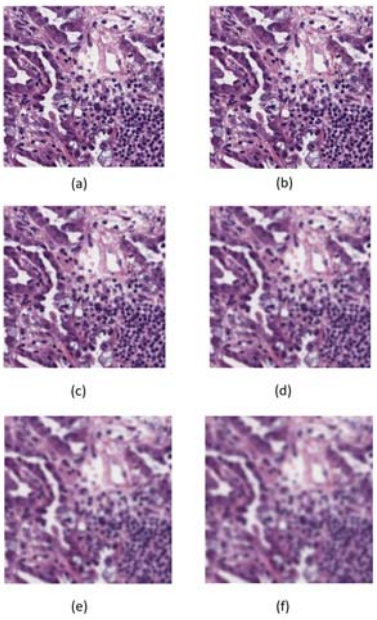

Fig. 1. The effect of different blurring standard deviations on a LUAD tile: (a) no blurring, and blurring with standard deviations of (b) 0.5, (c) 1.5, (d) 2.5, (e) 3.5, and (f) 4.5 pixels.

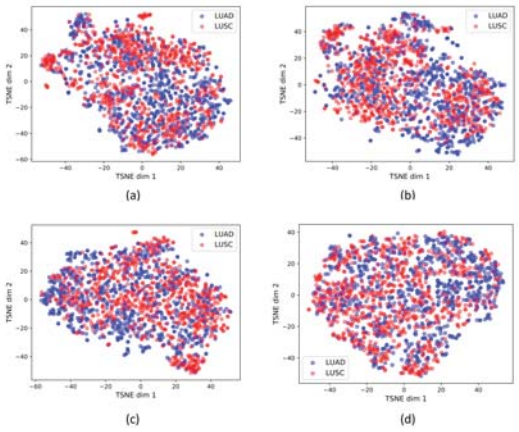

Fig. 2. TSNE plots depicting the separation between LUAD and LUSC classes at different blurring values: (a) no blurring, and blurring with standard deviations of (b) 0.5, (c) 2.5, and (d) 4.5 pixels.

B. Feature Representation Visualization

Figure 2 provides 2D TSNE plots (produced using the `sklearn` package) for various blurring standard deviations, suggesting the two classes only marginally separate in the TSNE plots. Although we don't observe strong class separations in the plots, [1] reports an AUC of $> 85\%$ using a tile-level approach based on the Inception V3 features. We observed similar issues using UMAP. Therefore, classification AUC is our only reference for assessing class separations.

Figures 3 and 4 provide 2D TSNE plots comparing non-blurred and blurred LUAD and LUSC slides, respectively. For both classes we observed marginal differences between TSNE plots for blurring standard deviations ≤ 0.5; however, larger standard deviations result in separations of the blurred and non-blurred slides in the TSNE plots. While blurring with a standard deviation of 1 is sufficiently strong to separate slides in the 2D TSNE space, the differences between non-blurred LUAD and LUSC deep learning features do not fully separate them in 2D TSNE plots. These results suggest blurring may have a stronger impact on the deep feature representation than the subtype differences. The strong effect of blurring on the deep feature representations are in-line with the results of [5] in blurred image detection.

Authorized licensed use limited to: The Ohio State University. Downloaded on April 09,2021 at 14:53:44 UTC from IEEE Xplore. Restrictions apply.

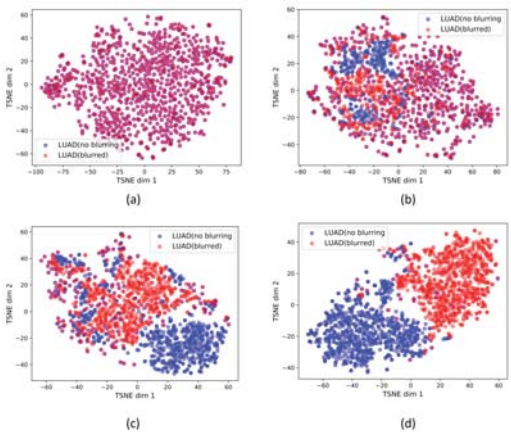

Fig. 3. TSNE plots depicting the separation between the original (non-blurred) LUAD and blurred LUAD slides using a Gaussian kernel with standard deviations of (a) 0.5, (b) 0.75, (c) 1, and (d) 1.25 pixels.

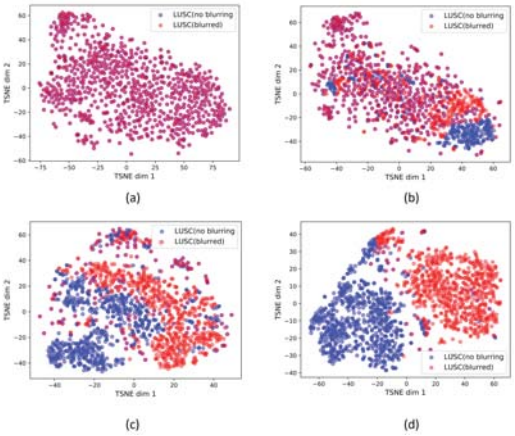

Fig. 4. TSNE plots depicting the separation between the original (non-blurred) LUSC and blurred LUSC slides using a Gaussian kernel with standard deviations of (a) 0.5, (b) 0.75, (c) 1, and (d) 1.25 pixels.

C. Classification Performance Evaluation

Figure 5 provides the AUCs of the trained models averaged over 100 train-test splits, where in each split 70% of slides in each class were used for training. LUSC is considered the positive label for AUC computation. The model trained on each set of blurred slides is applied to all blurred representations of test slides. As the Figure suggests, models trained on slides blurred with a Gaussian kernel with some standard deviation tend to perform well on test set slides are blurred with similar kernels. On the other hand, the classifiers suffer low AUCs when there are large differences between the blurring standard deviations of the train and test sets. Interestingly, models trained on moderate and large blurrings tend to suffer low AUCs when applied to non-blurred images. These results suggest that not only do the models suffer low AUCs when the test slides have lower resolutions (more blurring), but also when they have much higher resolution (less blurring). We observed a monotone decrease in AUC across the diagonal in Figure 5, but

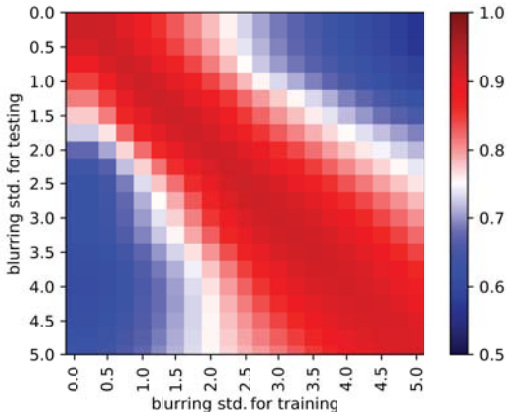

Fig. 5. The test AUC averaged over 100 train-test splits.

the reduction in AUC was small throughout. Gaussian kernels with standards deviations of 0 and 5 resulted in average AUCs of 92% and 90% on the diagonal, respectively. We observed a monotone decrease in AUC as the blurring difference between train and test slides increased.

IV. Conclusion

Deep learning models have drastically increased prediction accuracies in biomedical image analysis; however, they are black boxes and difficult to interpret. The current study suggests not only test slides with lower resolution compared with training data may result in low prediction accuracies, but also slides with higher resolutions may suffer similar issues. Models trained on blurred images tend to work well on slides with similar blurring. Future work entails similar analyses on other classification tasks to assess the extent the current observations generalize.

References

[1] Javad Noorbakhsh, Saman Farahmand, Ali Foroughi pour, Sandeep Namburi, Dennis Caruana, David Rimm, Mohammad Soltanieh-ha, Kourosh Zarringhalam, and Jeffrey H. Chuang. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *bioRxiv*, 2020.

[2] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, 1(8):800–810, 2020.

[3] Andrew J Schaumberg, Mark A Rubin, and Thomas J Fuchs. H&e-stained whole slide image deep learning predicts spop mutation state in prostate cancer. *BioRxiv*, page 064279, 2018.

[4] Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2016.

[5] Rui Wang, Wei Li, Runnan Qin, and JinZhong Wu. Blur image classification based on deep learning. In *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–6. IEEE, 2017.

[6] Shidan Wang, Donghan M Yang, Ruichen Rong, Xiaowei Zhan, Junya Fujimoto, Hongyu Liu, John Minna, Ignacio Ivan Wistuba, Yang Xie, and Guanghua Xiao. Artificial intelligence in lung cancer pathology image analysis. *Cancers*, 11(11):1673, 2019.

Authorized licensed use limited to: The Ohio State University. Downloaded on April 09,2021 at 14:53:44 UTC from IEEE Xplore. Restrictions apply.