Learning Robust Representation for Laryngeal Cancer Classification in Vocal Folds from Narrow Band Images

Debayan Bhattacharya[*1,2] DEBAYAN.BHATTACHARYA@TUHH.DE
Finn Behrendt[1] FINN.BEHRENDT@TUHH.DE
Axelle Felicio-Briegel[3] AXELLE.FELICIO@MED.UNI-MUENCHEN.DE
Veronika Volgger[3] VERONIKA.VOLGGER@MED.UNI-MUENCHEN.DE
Dennis Eggert[2] D.EGGERT@UKE.DE
Christian Betz[2] C.BETZ@UKE.DE
Alexander Schlaefer[1] SCHLAEFER@TUHH.DE

[1] *Institute of Medical Technologies and Intelligent Systems, Hamburg University of Technology, Hamburg, Germany*

[2] *Clinic for Ears, Nose and Throat, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*

[3] *Clinic for Ears, Nose and Throat, Ludwig Maximilian University, Munich, Germany*

Editors: Under Review for MIDL 2022

Abstract

Narrow Band Imaging (NBI) is increasingly being used in laryngology because it increases the visibility of mucosal vascular patterns which serve as important visual markers to detect premalignant, dysplastic and malignant lesions. To this end, deep learning methods have been used to automatically detect and classify the lesions from NBI endoscopic videos. However, the heterogeneity of the lesions, illumination changes due to phlegm on the mucosa and imaging artifacts such as blurriness make inter-patient endoscopic videos exhibit diverging image distributions. Therefore, learning representations that are robust to image distribution changes can be beneficial and improve the generalizing capability of the convolutional neural network (CNN). To this end, we propose a dual branch CNN that learns robust representations by combining deep narrow band features and wavelet scattering transform features of the narrow band images to classify vocal cord NBI images into malignant and benign classes. We show the generalizing capability of our learnt representation by training our neural network using two different losses: cross-entropy (CE) loss and supervised contrastive (SupCon) loss. Our source code is provided here.

Keywords: Narrow Band Imaging, Laryngeal Cancer, Cross Entropy Loss, Supervised Contrastive Loss

1. Introduction

Narrow Band Imaging is an emerging optical technique to detect and classify lesions in the upper respiratory tract as it enhances the contrast of vascular patterns compared to standard white light endoscopy. Physicians analyse the vascular patterns to diagnose the severity of the pathology (Arens et al., 2016). To this end, deep neural networks have been used to leverage these vascular patterns to detect and classify laryngeal squamous cell carcinoma (LSCC) from NBI video laryngoscopies (Azam et al.). However, there has been little investigation on the properties that features must have to improve the laryngeal

[*] First Author

cancer classification from NBI images of vocal folds. Typically, the vocal folds appear at different scales, have noise such as blurriness, reflection and illumination changes and appear in various orientations in the endoscopic videos. Therefore, learning representations that are scale, noise and rotation invariant could be beneficial. While data augmentations are typically employed to learn invariant representation, it is not guaranteed especially with limited ground truth data. To this end, we use wavelet scattering transform to extract features with defined properties such as scale, rotation, translation and deformation invariance (Mallat, 2012). We incorporate wavelet scattering transform features through a dual branch CNN with one branch extracting deep features from raw NBI images and a second branch extracting deep features from wavelet scattering transform features of NBI images. We show the improvement in generalization by evaluating on test set images that are grouped by patients (patient-stratified). Furthermore, we show the generalizing capability of the learnt representations by comparing our model trained on two losses (CE loss and SupCon (Khosla et al., 2020)) with and without the wavelet scattering features.

2. Methods

The dataset for this study consists of NBI videos of the larynx of 14 patients that were recorded using a flexible laryngoscope (ENF-VH, Olympus) attached to an NBI capable endoscopy system (Evis Exera III, Olympus). The study was approved by the local ethics committee (Project number 44-15, Ludwig Maximilian University of Munich). Written informed consent was gained from all patients undergoing this study. Images showing the vocal cords were extracted from NBI videos of 14 patients. 6 patients were diagnosed with LSCC (positive class) based on histological analysis and 8 were diagnosed with no LSCC. In total, there are 270 images of vocal cords with LSCC lesions and 553 images of vocal cords with no LSCC. The mean and standard deviation of the number of vocal cord images per patient was 50 ± 12. We performed a 6-fold cross validation approach where we left one patient with LSCC and one patient with no LSCC out for cross-validation and test set. The images of the remaining 10 patients were used in the training set. Data augmentations such as random rotation, color jitter, grayscale, random affine, vertical and horizontal flip were used. All are images were resized to 128×128. A batch size of 32 was used for all the experiments.

Our proposed model consists of two parallel branches. Each branch is a ResNet18. The first branch receives the NBI images of vocal cord. The second branch receives the wavelet scattering transform feature maps of the NBI images of vocal cord. We use a wavelet scattering transform with order $J = 2$ resulting in 81 feature maps per image channel. The two branches produce latent vectors with dimension D_1 and D_2 respectively. In our experiments, we keep the latent vector dimension of the ResNet that takes NBI images constant at $D_1 = 128$. We empirically found $D_2 = 32$ to be most beneficial for our model trained with CE loss based on our cross-validation experiments. Similarly, $D_2 = 8$ showed the best performance for our model trained with SupCon loss. The latent vectors of the two branches are concatenated and a linear classifier is employed to classify into one of two classes. We follow the training strategy described by (Khosla et al., 2020) for SupCon loss.

3. Results and Discussion

The results of our experiments are reported in Table 1. We observe that specificity, sensitivity and F1 weighted metrics are higher for models with wavelet scattering features.

Table 1: Results of our experiments

Method	**Loss**	$\mathbf{D_2}$	**Specificity**	**Sensitivity**	**F1 weighted**
w/o wavelet scattering	CE	-	0.58 ± 0.39	0.42 ± 0.44	0.40 ± 0.22
w/ wavelet scattering	CE	32	$\mathbf{0.74 \pm 0.33}$	$\mathbf{0.59 \pm 0.36}$	$\mathbf{0.62 \pm 0.33}$
w/o wavelet scattering	SupCon	-	0.52 ± 0.46	0.59 ± 0.43	0.42 ± 0.19
w/ wavelet scattering	SupCon	8	$\mathbf{0.58 \pm 0.37}$	$\mathbf{0.63 \pm 0.29}$	$\mathbf{0.56 \pm 0.26}$

The classification performance is better for models trained with CE loss than SupCon loss. Our results indicate that incorporating wavelet scattering features improve the generalizing capability of the CNN. The performance improvements indicate that features maps with defined properties such as translation, rotation, noise and deformation invariance are important towards laryngeal cancer classification on our dataset of NBI images. However, we also observe a relatively large standard deviation in our reported metrics. An extended dataset containing more patient laryngoscopies may reduce the standard deviation of the observed metrics. The large standard deviation also highlights the image distribution divergence from patient to patient video laryngoscopies. Multiple factors such as heterogeneity of pathology, blurriness, laryngeal content such as phlegm and reflections account for the image distribution divergence in inter-patient NBI images. Therefore, it becomes critical to evaluate a network using patient-stratified split to make a fair assessment of its generalizing capability. Furthermore, patient-stratified evaluation better reflects the clinical use-case where an inference must be made on an unseen endoscopic video of a new patient. In conclusion, we propose a model that combines wavelet scattering features and deep features. We show improved generalizing capability due to wavelet scattering features by training our model on two losses.

References

Christoph Arens, Cesare Piazza, Mario Andrea, Frederik G. Dikkers, Robin E. A. Tjon Pian Gi, Susanne Voigt-Zimmermann, and Giorgio Peretti. Proposal for a descriptive guideline of vascular changes in lesions of the vocal folds by the committee on endoscopic laryngeal imaging of the european laryngological society. *European Archives of Oto-Rhino-Laryngology*, 273(5):1207–1214, May 2016. ISSN 1434-4726. doi: 10.1007/s00405-015-3851-y. URL https://doi.org/10.1007/s00405-015-3851-y.

Muhammad Adeel Azam, Claudio Sampieri, Alessandro Ioppi, Stefano Africano, Alberto Vallin, Davide Mocellin, Marco Fragale, Luca Guastini, Sara Moccia, Cesare Piazza, Leonardo S. Mattos, and Giorgio Peretti. Deep learning applied to white light and narrow band imaging videolaryngoscopy: Toward real-time laryngeal cancer detection. *The Laryngoscope*, n/a(n/a). doi: https://doi.org/10.1002/lary.29960. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/lary.29960.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10): 1331–1398, 2012. doi: https://doi.org/10.1002/cpa.21413. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21413.