REAL-TIME INFORMATIVE LARYNGOSCOPIC FRAME CLASSIFICATION WITH PRE-TRAINED CONVOLUTIONAL NEURAL NETWORKS

Adrian Galdran[⋆,†], *P. Costa*[⋆], *A. Campilho*[∗,‡]

[⋆] Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal
[†] École de Technologie Supérieure, University of Quebec, Canada
[‡] Faculdade de Engenharia, Universidade de Porto, Portugal

ABSTRACT

Visual exploration of the larynx represents a relevant technique for the early diagnosis of laryngeal disorders. However, visualizing an endoscopy for finding abnormalities is a time-consuming process, and for this reason much research has been dedicated to the automatic analysis of endoscopic video data. In this work we address the particular task of discriminating among informative laryngoscopic frames and those that carry insufficient diagnostic information. In the latter case, the goal is also to determine the reason for this lack of information. To this end, we analyze the possibility of training three different state-of-the-art Convolutional Neural Networks, but initializing their weights from configurations that have been previously optimized for solving natural image classification problems. Our findings show that the simplest of these three architectures not only is the most accurate (outperforming previously proposed techniques), but also the fastest and most efficient, with the lowest inference time and minimal memory requirements, enabling real-time application and deployment in portable devices.

Index Terms— Laryngoscopy, Informative Frame Classification, Convolutional Neural Networks, Real-Time

1. INTRODUCTION

Visual assessment of the laryngeal tract is of key relevance for laryngeal healthcare. For instance, voice disorders can be diagnosed by means of the analysis of vocal fold vibrations on laryngoscopic videos [5]. Also early diagnosis of laryngeal cancer can be supplemented with visual analysis of the larynx, due to the costs associated with extracting tissue samples for further histopathological analysis. For this reason, great efforts have been invested in developing advanced optical exploratory techniques like Narrow-Band Imaging (NBI) endoscopes, which provide an improved view of the laryngeal surface, enabling safer and easier patient examination [9].

However, carefully analyzing an endoscopy is a time-consuming and error-prone operation that is ideally suited for computer-aided systems. A computational technique can draw the attention to relevant parts of the video, automatically find signs of disease that may go unnoticed, or be useful for video stitching purposes [10]. Nevertheless, computational endoscopic video analysis is known to be a challenging task. In particular, laryngoscopic video processing presents relevant difficulties associated to the presence of saliva, the constant movement of vocal folds and swallowing muscles, or a greatly varying illumination [7]. Furthermore, in this context it is of great importance to achieve real-time processing capability, since this could inform the clinician handling the endoscope that the acquisition parameters should be modified, or the endoscope flushed, or even trigger an image quality enhancement technique to improve visibility [6].

Although much research has been devoted to endoscopic video analysis, comparatively few works have focused on laryngoscopies. The work in [7] addressed the identification of blurred frames from endoscopic videos based on simple thresholding, while the technique introduced in [8] designed a machine learning-based classifier from a set manually-engineered visual features. These encompassed edge detector responses, local image entropy and variance, intensity histograms, and amount of detected keypoints on a given frame.

In this paper, we propose to train a Convolutional Neural Network (CNN) for the task of classifying laryngoscopic frames into informative or uninformative, further categorizing uninformative frames into three possible classes, as shown in Fig. 1. In order to extract as much information as possible from little training data while avoiding to overfit it, we analyze the suitability of fine-tuning three state-of-the-art CNN architectures from pre-existing weights. After suitably re-training such models, our findings indicate that, when compared with previous approaches, every architecture results in a superior performance. Interestingly, the simplest of them is the one reaching highest accuracy, resulting in an excellent compromise between classification performance, memory requirements, and inference speed, rendering the approach suitable for real-time predictions in embedded devices.

Corresponding author: Adrian Galdran (adrian.galdran@inesctec.pt). Research funded by ERDF European Regional Development Fund, project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016".

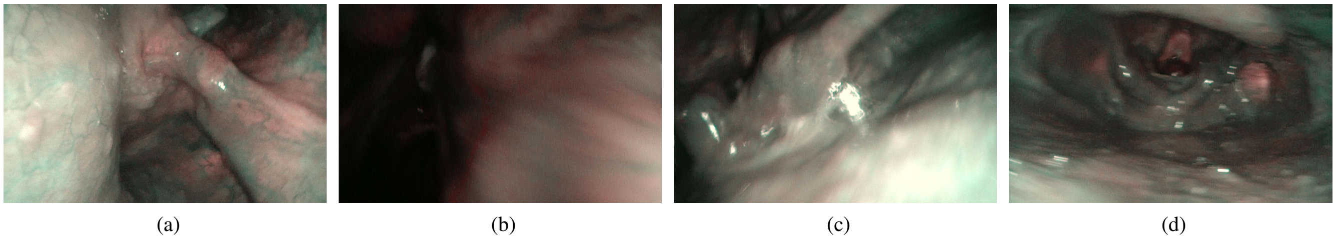

Fig. 1: Laryngoscopic video frames belonging to the four classes of interest: a) Informative (**I**), b) Blurry (**B**), c) Containing Saliva/Specularities (**S**), d) Underexposed (**U**).

2. METHODOLOGY

In this work, we consider three CNN models that have established themselves as a standard in computer vision applications in recent years, namely Inception V2 [11], ResNet 50 [2], and SqueezeNet [3].

Inception V2 is an architecture introduced in [11]. Its main contribution is the reduction in computational burden of previous CNN architectures while reaching improved performances via the introduction of *inception* modules. Inception modules consist of the parallel computation of several data streams built of 1×1 and 3×3 convolutional filters. The key mechanism within inception modules that enables a decreased computational load is the introduction of a *bottleneck layer* on them. A bottleneck layer reduces the depth of the feature maps by simply applying a 1×1 convolution prior to entering the expensive parallel blocks. This can be considered as a dimensionality reduction technique that removes redundancy in the input features, and even if it results in an apparent information loss, Inception V2 was demonstrated to perform at state-of-the-art levels of accuracy [11].

ResNet 50, introduced in [2], is part of the family of residual neural networks, the key contribution of which was the addition of *skip connections*. Skip connections address the well-known vanishing gradient problem: in very deep neural networks, the weights in early layers of the network are not properly updated. This is due to the backpropagation algorithm propagating increasingly smaller error gradient values. Skip connections help in preserving the error gradient by allowing to backpropagate through an identity mapping instead of through standard layers. This is achieved within a residual block through the mapping of input features x to output features $H(x)$ by means of the following formula:

$$H(x) = F(x) + x,$$

where $x \mapsto F(x)$ is a standard neural network layer. In this case, if during the training stage it is found that backpropagating the error signal through F is harmful for the model performance, the training process can be automatically corrected to deviate through the identity mapping $H(x) = x$, which will not modify the error gradient values at all.

By stacking residual blocks, ResNets with up to several hundred convolutional layers can be trained. However, no significant improvement is achieved by using such a large amount of layers. Hence, in this work we restrict ourselves to a 50-layers residual network, which is one of the standard architectures employed nowadays in computer vision.

SqueezeNet was introduced in [3] with the aim of reducing model complexity (amount of learnable weights) as much as possible, while preserving reasonable accuracy in classification tasks. For this purpose, the main novelty was the construction of a special convolutional module, termed *Fire*. The *Fire* layer is based on three objectives, namely 1) favor the use of 1×1 convolutional filters throughout the entire architecture, 2) decrease the number of input channels in volumes that reach convolutional layers composed of 3×3 filters, and 3) differ activation downsampling to later in the network, to produce large spatial resolution in early layers under the hypothesis that this can increase classification accuracy.

In order to achieve the three aforementioned goals, a *Fire* module is built, composed of a so-called *squeeze* convolutional layer containing only 1×1 filters, followed by an *expand* layer composed of a mixture of 1×1 and 3×3 filters, arranged so that 3×3 filters receive feature maps with a limited amount of input channels. Fire modules are combined into a SqueezeNet network, which can be trained to achieve state-of-the-art classification results on ImageNet. A compressed version of this network can be stored in less than 0.5MB while still preserving classification accuracy [3].

2.1. Fine-Tuning Process

All three CNN models used in this work were pre-trained on ImageNet and fine-tuned to the available training data by minimizing the cross entropy loss. We used the Adam [4] optimizer with a learning rate of $1E^{-3}$ and default β_1 and β_2 values ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The learning rate was decayed by 0.1 after every 7 epochs. Early stopping was used with a patience of 5 by monitoring Area Under the ROC Curve (AUC) on a separate validation set. Input images were resized to 299×487 pixels to keep the original aspect ratio. Finally, we used standard dataset augmentation operations such as random translations, scaling and horizontal flips.

3. EXPERIMENTAL EVALUATION

3.1. Data

For benchmarking the three models described in the previous section, we employ the recent introduced NBI-InfFrames dataset [8]. This dataset contains 720 video frames obtained from 18 NBI laryngoscopic videos of 18 different patients. All of them were affected by laryngeal spinocellular carcinoma, as confirmed by further histopathological examination. Video acquisition was performed with a Narrow Band Imaging endoscope at a frame rate of 25 frames per second and a resolution of 1920 × 1072 pixels. From the 720 available samples, 180 were Informative (**I**), 180 were considered as Blurred (**B**), 180 were declared as containing Saliva or Specular reflections (**S**), and 180 were deemed as underexposed (**U**) by two different medical experts, see Fig. 1.

The NBI-InfFrames dataset comes already divided into three different folds, carefully constructed to separate frames patient-wise into different folds. In our case, for each considered model we performed three different training stages. In each stage, one fold was employed for training the model, another one for validation purposes, and the third fold was used for testing the performance of the model. This was repeated three times, suitably varying the corresponding test fold.

3.2. Quantitative Evaluation

For a numerical evaluation of the performance achieved by each model, we computed True Positive Rate and False Positive Rate averaged over the three experiments described above. From this, and given that the dataset was balanced, we built macro-average ROC curves for each of them. The resulting curves are shown in Fig. 2, together with the corresponding Area Under the Curve (AUC) values.

From the previous experiment, we can observe that the three models achieved a high performance in the task of discriminating among the four classes of interest. Interestingly, the fine-tuned SqueezeNet model obtained the largest performance, with a nearly perfect AUC. For this reason, we selected this model as the best-performing one and further studied its performance in each class of interest, with a similar analysis as the one offered in [8]. For this, after thresholding predictions at $t = 0.5$, we computed True Positives (TP_j), False Positives (FP_j), and False Negatives (FN_j) for each class $j \in \{1, 2, 3, 4\}$. With this, per-class Precision, Recall, and F1 scores were computed as follows:

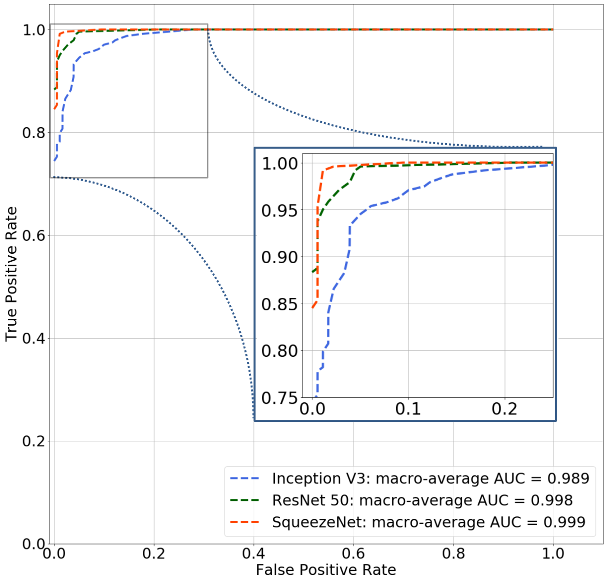

Fig. 2: ROC curves for each of the three considered models.

$$\text{Precision}_j = \frac{\text{TP}_j}{\text{TP}_j + \text{FP}_j}, \qquad \text{Recall}_j = \frac{\text{TP}_j}{\text{TP}_j + \text{FN}_j},$$

$$\text{F1}_j = 2 \cdot \frac{\text{Precision}_j \cdot \text{Recall}_j}{\text{Precision}_j + \text{Recall}_j}.$$

Table 1 shows the result of computing the above performance measures for the SqueezeNet-based model, with results reported in [8] for the same task and equal experimental setting.

In addition to analyzing how accurate predictions were for each class, we were also interested in studying the time required for each model to produce a prediction given an input frame. These inference times are shown in Table 2 for each of the three considered techniques, together with the execution time reported in [8]. It should be noticed that the latter was not obtained by running the method in a computer with the same specifications as the first three. In particular, inference times in this paper are reported for a GPU-based computation with a NVIDIA GeForce GTX 1060, by testing each model with a batch size of 1. Increasing the batch size further decreases the mean inference time due to a better exploitation of the parallel computing capabilities of GPUs.

Table 1: Performance comparison between the technique from [8] and fine-tuned SqueezeNet. **IQR**: Inter-Quartile Range.

	Feature-Based + SVM [8]						**Proposed (SqueezeNet-based)**					
	I	**B**	**S**	**U**	**Median**	**IQR**	**I**	**B**	**S**	**U**	**Median**	**IQR**
Precision	0.91	0.76	0.78	0.76	0.77	0.09	0.97	0.94	0.93	0.97	0.95	0.03
Recall	0.91	0.83	0.62	0.85	0.84	0.16	1	0.94	0.91	0.94	0.94	0.02
F1-Score	0.91	0.79	0.69	0.80	0.80	0.12	0.98	0.94	0.91	0.95	0.95	0.03

Table 2: Inference time per frame for the technique from [8] and the three considered models.

	SVM [8]	Inception V3	ResNet 50	SqueezeNet
Time (s)	3.00E^{-2}	1.70E^{-2}	8.57E^{-3}	4.27E^{-3}

4. DISCUSSION AND FUTURE WORK

From Fig. 2, we observe that the three fine-tuned CNNs tested in this work exceed previously reported performance in the NBI-InfFrames dataset by a wide margin. Performance metrics demonstrate that the weights learned for the task of natural image classification are indeed useful for identifying informative laryngoscopic frames. Interestingly, the simplest of the three employed models delivered the highest performance, revealing that too much complexity may lead to a slight overfitting, and it may be preferable to opt for moderately expressive architectures to increase generalizability in this setting.

The analysis reported in Table 1 shows that both the approach introduced in [8] and SqueezeNet suffer when assigning the correct category to uninformative frames, whereas the classification into informative frames was almost perfect for the case of the CNN. When classifying uninformative frames into blurry, containing specularities/saliva, and underexposed, the fine-tuned SqueezeNet achieved a slightly lower performance, although the overall median accuracy was high enough so as to consider this simple CNN as an excellent baseline for further research on laryngoscopic frame classification, with a 94% median recall and a 95% median precision. In addition, the inter-quartile range was also lower than in [8] for every measure, pointing to a great robustness of the proposed approach in this problem.

Our findings indicate that a simple CNN like SqueezeNet suffices to successfully solve the task of informative frame classification in laringoscopies. In addition, the moderate complexity of SqueezeNet turns this approach into an ideal candidate for its introduction in clinical workflows. Its lightweight memory requirements (less than 0.5 MB) enable its embedding even in portable devices, and its fast inference time (one order of magnitude lower than previously reported times) allows for the addition of further image processing or computer vision post-processing modules (*e.g.* abnormality detectors or image quality enhancement techniques) without hindering the potential for real-time execution.

The next step will consist of considering full video processing, where the temporal dimension of the data could be also taken into account by means of Recurrent Neural Networks coupled with CNNs. This may further increase the consistency of predictions, and enable an improved analysis of laryngoscopic frame visual content. Image processing techniques for restoring visibility on un-informative frames based on different exposures [1] will also be considered.

5. REFERENCES

[1] A. Galdran. Image dehazing by artificial multiple-exposure image fusion. *Signal Processing*, 149:135–147, Aug. 2018.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[3] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5mb model size. Feb. 2016. arXiv: 1602.07360.

[4] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, volume 5, 2015.

[5] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Döllinger. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. *Medical Image Analysis*, 11(4):400–413, Aug. 2007.

[6] X. Luo, A. J. McLeod, S. E. Pautler, C. M. Schlachta, and T. M. Peters. Vision-Based Surgical Field Defogging. *IEEE Transactions on Medical Imaging*, 36(10):2021–2030, Oct. 2017.

[7] S. Moccia, V. Penza, G. O. Vanone, E. D. Momi, and L. S. Mattos. Automatic workflow for narrow-band laryngeal video stitching. In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1188–1191, Aug. 2016.

[8] S. Moccia, G. O. Vanone, E. D. Momi, A. Laborai, L. Guastini, G. Peretti, and L. S. Mattos. Learning-based classification of informative laryngoscopic frames. *Computer Methods and Programs in Biomedicine*, 158:21–30, May 2018.

[9] C. Piazza, F. Del Bon, G. Peretti, and P. Nicolai. Narrow band imaging in endoscopic evaluation of the larynx:. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 20(6):472–476, Dec. 2012.

[10] M. Schuster, T. Bergen, M. Reiter, C. Münzenmayer, S. Friedl, and T. Wittenberg. Laryngoscopic Image Stitching for View Enhancement and Documentation – First Experiences. *Biomedical Engineering / Biomedizinische Technik*, 57(SI-1 Track-H):704–707, 2012.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.