Wolfgang Reiter*

Improving endoscopic smoke detection with semi-supervised noisy student models

https://doi.org/10.1515/cdbme-2020-0026

Abstract: Laparoscopic surgery consists of many tasks that have to be handled by the surgeon and the operating room personnel. Recognition of situations where action is required enables automatic handling by the integrated OR or notifying the surgical team with a visual reminder. As a byproduct of some surgical actions, electrosurgical smoke needs to be evacuated to keep the vision clear for the surgeon. Building on the success of convolutional neural networks (CNNs) for image classification, we utilize them for image based detection of surgical smoke. As a baseline we provide results for an image classifier trained on the publicly available smoke annotions of the Cholec80 dataset. We extend this evaluation with a self-training approach using teacher and student models. A teacher model is created with the labeled dataset and used to create pseudo labels. Multiple datasets with pseudo labels are then used to improve robustness and accuracy of a noisy student model. The experimental evaluation shows a performance benefit when utilizing increasing amounts of pseudo-labeled data. The state of the art with a classification accuracy of 0.71 can be improved to an accuracy of 0.85. Surgical data science often has to cope with minimal amounts of labeled data. This work proposes a method to utilize unlabeled data from the same domain. The good performance in standard metrics also shows the suitability for clinical use.

Keywords: computer-assisted interventions; deep learning; endoscopic smoke detection; semi-supervised learning.

Introduction

A surgeon has many tasks which require manual intervention and focused attention. Automation of such tasks can alleviate the burden on surgeons and allow their attention to stay on more important topics. One such task is manual smoke evacuation. Correct handling of surgical smoke consists of multiple steps and would profit highly from automation. Surgical smoke is a byproduct of coagulation tools like the high-frequency electrosurgical unit or coagulating shears. It diminishes the quality of the endoscopic vision with increasing intensity and must be removed from the body. Smoke evacuation must be accompanied by an increase of gas flow into the body to keep the pressure in the cavity constant. Additionally the evacuation is best handled by a specific device to avoid health risks to the surgical staff [3]. Aside from this intraoperative uses of smoke detection, there also exist various postoperative applications for recording analysis. Large amounts of smoke in a recording may indicate surgical errors due to the deteriorated quality of the endoscopic vision. The automatic localization of smoke in recordings can also help to navigate through the vast amounts of recordings that hospitals have to store. A large amount of research on automatic smoke detection focuses on the more general fire and smoke detection in outdoor videos. [4] applies convolutional neural networks (CNNs) for fire classification. The work in [5] uses the Faster R-CNN object detection model to determine the location of smoke and fire in images. The authors also try to overcome the lack of data with synthetic smoke images. In the surgical domain few works exist on the subject of automatic smoke detection. Manually extracted features of laparoscopic recordings lead to very good results with a SVM classifier in [6]. This work evaluates smoke detection on only four recordings and reports metrics on a testset randomly sampled from all data. Another SVM based algorithm is used in [7] with good results. This work also suffers from the very limited dataset of only 76 extracted clips with less than 5,000 frames in total. The work in [8] evaluates a histogram method based on colorspace features as well as a pretrained GoogleNet CNN on a non-public dataset with 30,000 images. The same authors published smoke annotations for parts of the cholec80 dataset [9] together with new evaluations in [1]. To overcome the lack of available training data for surgical data science in endoscopy several works evaluated the use of unlabeled recordings for surgical workflow analysis [10] or instrument segmentation [11]. The work in [2] utilizes large amounts of unlabeled data with a pseudo-labeling approach to improve a noisy student model. A teacher model is trained in a supervised way and used to create pseudo labels for the unlabeled data.

***Corresponding author: Wolfgang Reiter**, Wintegral GmbH, Munich, Germany, E-mail: wolfgang.reiter@wintegral.net

This work evaluates the use of pseudo-labeled data to improve the state of the art in smoke detection with CNNs. Student models are learned iteratively with increased capacity to achieve knowledge expansion. Input and model noise is applied to the student model increasing the range of learnable invariants. This algorithm is used iteratively to get the most out of the available data.

Materials

This work uses the following datasets: The cholec80 dataset [9] with the public smoke annotations from [1] referred to as dataset I, the part of cholec80 without annotations as dataset II and a non-public unlabeled dataset showing similar laparoscopic procedures designated dataset III. The available smoke annotations have been created on a subset of the cholec80 dataset extracted at 25 frames per second. The roughly 100,000 annotated frames of this dataset show minimal variation in parts due to this high framerate. Extracting test and validation datasets with random sampling would therefore lead to information leakage and unrealistic high metrics. This makes it necessary to choose the validation and test subsets from different videos to show a more realistic generalization error. Sixty videos are used for the training set, and 10 videos each for validation and test. This results in 75,622 images for training and the rest for validation and test. For the unlabeled part of the cholec80 dataset pseudo labels are created at one frame per second. This gives 159,440 additional images. Pseudo labels for the non-public dataset are created in the same way for 257,611 images.

A rectangular area is cropped from the center of the cholec80 images to reduce the influence of the dark circle as used in [1]. The images of both datasets are cropped to the same aspect ratio, resized to square images with size 256 by 256 pixels. Finally data augmentations are applied and the image data is normalized to the range [–1, 1].

Methods

We compare the two best performing methods, Saturation Peak Analysis (SPA) and classification with a CNN, from [1] which represent the state of the art with different noisy student models.

State of the art

For the SPA a histogram of the saturation channel of HSV images is calculated. Peaks in the histogram that stretch over multiple histogram bins and reach over an empirically determined threshold are used as features in the analysis. A classification threshold separating high from low saturation values, again found empirically, is then used to classify the images in the classes smoke and non-smoke [8].

The best performing method presented in [8], is a CNN classifier based on the GoogleNet-Architecture. It is trained on 80% of the class-balanced dataset starting with pretrained weights. Evaluated inputs are RGB images and saturation-channel only HSV images, where the former perform better by a small margin.

Noisy student model

The self-training of a noisy student model as introduced in [2] is based on the idea of utilizing large amounts of unlabeled data to augment a smaller labeled dataset. As a first step a teacher model is trained on the labeled data and used to create predictions for the unlabeled data. These pseudo labels are then subsequently used to train a student model on the combined datasets, hand-labeled and pseudo-labeled. This training-procedure successfully improved the state of the art in the ImageNet challenge using 300 million unlabeled images showing improvements especially in the robustness of the classifier.

We apply this self-training approach to the problem of smoke-detection. DenseNet is chosen as model architecture, since ImageNet pretrained weights for various model sizes are available. This ability to scale the network size is needed to apply the concept of knowledge expansion: the student model, trained on more data, is larger than the teacher. The student is thought to achieve better results through more training data and a more difficult target function due to the applied noise [2]. For the teacher model we choose DenseNet121. The student is evaluated with network sizes DenseNet121 and DenseNet169. Classification heads of pretrained models are replaced with a global average pooling layer, followed by a fully connected linear layer with dimensionality 4096 and another one with dimensionality two. The linear layers are separated by ReLu non-linearities. Model noise is added with dropout functions between the newly added linear layers. Aside from its ability to improve generalization, this noise improves local smoothness in the decision function when used together with unlabeled data. This enables a more coherent clustering of the inputs with the hand-labeled inputs providing for the correct cluster assignment of the output vectors [12]. Additionally the input of the model is noised with data augmentation acting as invariance constraints for small perturbations. Random rotations, center crops and color perturbations of the HSV-image channels are applied with a probability of 50%. The created pseudo labels are reduced with confidence based subsampling. The softmax-activated outputs of the teacher model are interpreted as confidence values and all samples with a confidence lower than 0.75 are discarded from the training. Overrepresented classes in the pseudo-labeled dataset are further reduced to achieve class balance. To allow the model to better fit the pseudo-labeled training data, the ratio of pseudo-labeled to hand-labeled data is increased in favor of the former. This training procedure is then repeated 3 times. Each iteration leading to larger datasets due to increased confidence on the unlabeled data. Datasets II and III are subsampled to approximately 50 and 62% in the first two iterations. In the third iteration 75% of dataset II and 80% of dataset III are used for training.

Results and discussion

The top two rows of Table 1 show the metrics for the state of the art, taken from [1]. SPA in the third line shows the

results on the testset for our own reimplementation of this algorithm: results are very near to the original, higher precision suggests a reduced number of false positives, the lower recall an increased number of false negatives. Row four shows the testset metrics for the teacher model (D121). It shows an increase in all metrics compared to GoogleNet (GLN RGB) due to the modernized neural network architecture.

To evaluate the noisy student algorithm the datasets were used as follows. Dataset I, the part of the cholec80 dataset with public smoke annotations is used to train the teacher model. The student models were trained with datasets I, II and III. Dataset II, the unlabeled part of cholec80 with pseudo labels. And dataset III, the non-public dataset with pseudo-labels.

Table 2 shows student models, of both sizes, trained on two of the three datasets: dataset I and II or dataset I and III. It can be seen that using any one of both pseudo-labeled datasets improves performance compared to the teacher model by approximately the same margin. The numbers also show, that larger student models trained on two of the datasets perform slightly worse than a same-sized student.

Finally Table 3 shows the results for student models trained on all three datasets. Utilising more pseudo-labeled data shows an improvement for the larger models (D169) over the small students (D121) by a small margin. The ratio of pseudo-labeled samples is further increased for the models in the last two rows. The D121 model (penultimate row) shows almost no change in metrics after this change of ratios, except a small substitution of false negative and false positive samples. This leads to an increase of precision and a decrease of the recall metric. The D169 model (last row) on the other hand shows distinct improvements. Accuracy, precision and f1 metrics show an increase, with only recall being in the same range as the others or slightly lower than the first model. A nearly constant recall metric is an indication that the number of false negatives for this best model stays approximately the same. Meaning that the recognition of the most difficult examples is not improved by much. Whereas the growth of the other metrics points to a decrease in the number of false positives. In other words, the reduction of false positives suggests less overfitting, meaning better generalization.

Figure 1 depicts results for the best model. False negatives in the first row containing low amounts of smoke, making it hard to recognize even for the human eye. The second row shows smoke-free images that were wrongly determined to contain smoke. The image on the left shows the typical color distribution of laparoscopic images, the classifier might have been fooled by the white discoloration on the tool dominating the bottom right. The image on the right with a specimen retrieval bag and slightly blurred vision is most likely misclassified due to these greyish features. The last row depicts two correctly classified smoke images. The amount of smoke is high enough that the human eye can distinguish the smoke even in this small image size.

Table 1: Results showing the accuracy, precision, recall and F1 metrics for the baseline models. Metrics for DenseNet121 teacher model are listed in row D121 (DenseNet121). Rows 1 and 2 show results on DS C.2 from [1]. Models in rows 3 and 4 are trained and evaluated on dataset I.

Model	Acc	Prec	Rec	F1
SPA [1]	0.697	0.771	0.560	0.649
GLN RGB [1]	0.711	0.771	0.600	0.675
SPA	0.688	0.812	0.490	0.612
D121	0.787	0.768	0.785	0.776

Table 2: Student model results for D121 and D169 (DenseNet169) utilising one pseudo-labeled dataset. Dataset-ratios are stated after the model-name in the order: I/II/III.

Model	Acc	Prec	Rec	F1
D121 0.2/0.8/0	0.814	0.882	0.767	0.820
D121 0.2/0/0.8	0.804	0.849	0.767	0.806
D169 0.2/0.8/0	0.810	0.840	0.781	0.809
D169 0.2/0/0.8	0.808	0.896	0.751	0.817

Table 3: Student model results utilising both pseudo-labeled datasets. Dataset-ratios are again stated after the model-name.

Model	Acc	Prec	Rec	F1
D121 0.30/0.35/0.35	0.840	0.852	0.821	0.836
D169 0.30/0.35/0.35	0.843	0.895	0.801	0.845
D121 0.10/0.45/0.45	0.840	0.904	0.792	0.845
D169 0.10/0.45/0.45	0.852	0.910	0.807	0.855

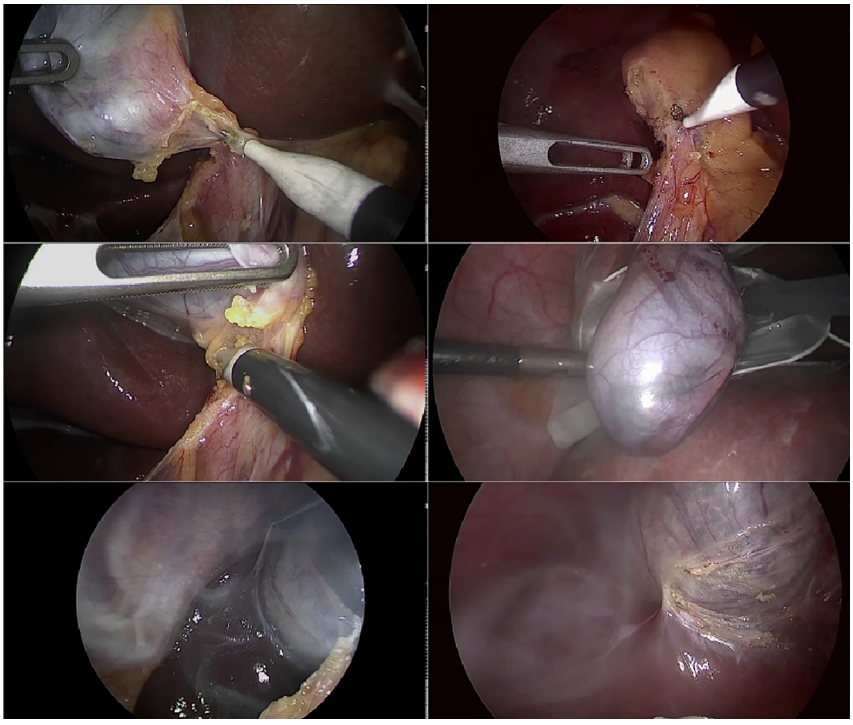

Figure 1: Examples of model predictions. The rows contain from top to bottom: false negative, false positive and in the third row true positive smoke detections.

Conclusion

A method for utilising unlabeled endoscopic recordings to improve smoke detection was proposed. In a first step the state of the art is improved upon by modernizing the employed model architecture. The results show that further improvements can be reached by utilising increasing amounts of pseudolabeled data and by scaling the network size accordingly. The prediction of less false positives also shows that the robustness of the classifier is improved. While the results are promising there are still many false negatives in the predictions. Performance on these harder examples, showing only small amounts of smoke when it is emerging or during the evacuation might be improved by using more training data with such edge cases.

Research funding: The author state no funding involved.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Informed consent: Informed consent has been obtained from all individuals included in this study.
Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

Competing interest: My employer is Wintegral GmbH and the parent company is Richard Wolf GmbH.

References

1. Leibetseder A, Primus MJ, Petscharnig S, Schoeffmann K. Real-time image-based smoke detection in endoscopic videos. In: Proceedings of the on thematic workshops of ACM multimedia 2017; 2017:296–304 pp.
2. Xie Q, Luong MT, Hovy E, Le QV. Self-training with noisy student improves ImageNet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020: 10687–98 pp. https://doi.org/10.1109/CVPR42600.2020.01070.
3. Takahashi H, Yamasaki M, Hirota M, Miyazaki Y, Moon J, Souma Y, et al. Automatic smoke evacuation in laparoscopic surgery: A simplified method for objective evaluation. Surg Endosc 2013;27: 2980–7. Publisher: Springer. https://doi.org/10.1007/s00464-013-2821-y.
4. Sharma J, Granmo OC, Goodwin M, Fidje JT. Deep convolutional neural networks for fire detection in images. In: International conference on engineering applications of neural networks. Springer; 2017:183–93 pp.
5. Zhang Q, Lin G, Zhang Y, Xu G, Wang J. Wildland forest fire smoke detection based on faster R-CNN using synthetic smoke images. In: Procedia engineering. Publisher: Elsevier; 2018, vol 211. p. 441–6.
6. Alshirbaji TA, Jalal NA, Mündermann L, Möller K. Classifying smoke in laparoscopic videos using SVM. Curr Dir Biomed Emg Jan. 2017; 3. https://doi.org/10.1515/cdbme-2017-0040.
7. Loukas C, Georgiou E. Smoke detection in endoscopic surgery videos: a first step towards retrieval of semantic events: smoke detection in endoscopic surgery videos. Int J Med Robot Comput Assist Surg Mar. 2015;11:80–94.

8. Leibetseder A, Primus MJ, Petscharnig S, Schoeffmann K. Image-based smoke detection in laparoscopic videos. In: Computer assisted and robotic endoscopy and clinical image-based procedures. Springer; 2017:70–87 pp.
9. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imag Jan. 2017;36:86–97.
10. Funke I, Jenke A, Mees ST, Weitz J, Speidel S, Bodenstedt S. Temporal coherence-based self-supervised learning for laparoscopic workflow analysis. In: OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis. Cham: Springer International Publishing; 2018, vol 11041: 85–93 pp.
11. Ross T, Zimmerer D, Vemuri A, Isensee F, Wiesenfarth M, Bodenstedt S, et al. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. en, Int J CARS Jun. 2018;13:925–33.
12. Laine S, Aila T. Temporal ensembling for semi-supervised learning. Mar. 2017. arXiv:1610.02242 [cs], arXiv: 1610.02242.