Distilling the knowledge in CNN for WCE screening tool

Thomas Garbay, Orlando Chuquimia, Andrea Pinna, Hichem Sahbi, Xavier Dray, Bertrand Granado

Distilling the knowledge in CNN for WCE screening tool

Thomas GARBAY
LIP6, CNRS UMR 7606
Sorbonne Universite, Paris, France
thomas.garbay@lip6.fr

Orlando CHUQUIMIA
LIP6, CNRS UMR 7606
Sorbonne Universite, Paris, France
orlando.chuquimia@lip6.fr

Andrea PINNA
LIP6, CNRS UMR 7606
Sorbonne Universite, Paris, France
andrea.pinna@lip6.fr

Hichem SAHBI
LIP6, CNRS UMR 7606
Sorbonne Universite, Paris, France
hichem.sahbi@lip6.fr

Xavier DRAY
APHP - Hpital Saint-Antoine
Sorbonne Universite, Paris, France
xavier.dray@aphp.fr

Bertrand GRANADO
LIP6, CNRS UMR 7606
Sorbonne Universite, Paris, France
bertrand.granado@lip6.fr

Abstract—**A way to improve the early detection of colorectal cancer is screening. Polyps are a marker of colorectal cancer and the best modality to detect them is the image. In 2003 Wireless Capsule Endoscopy was introduced and opened a way to integrate automatic image processing to realize a screening tool. Moreover, the capacity to detect polyp with Convolutional Neural Network was shown in many scientific studies, but one issue is the integration of these networks. In this article, we present our works to integrate CNN or image processing based on a CNN inside a WCE to realize a powerful screening tool. We apply the knowledge distillation method. We prove that knowledge distillation is efficient from VGG16 to Squeezenet in polyp detection context**

I. Introduction

In 2018 Colorectal cancer (CRC) was the second highest cause of death by cancer worldwide. The mortality rate was 47.62% and corresponding to 880,792 deaths. [1], [20]. It is a public health problem. In about 95% of the cases, the beginning of CRC is a growth on the inner lining of the colon or rectum called polyp [6]. The European Code Against Cancer recommends an early detection to greatly improve outcomes through screening of gastrointestinal (GI) tract [7]. Indeed, in about 90% of the cases, CRC is treatable if it is detected before polyp become adenocarcinomas [18].

Today, to find polyps, image is the modality for analyzing the colon. Screening, diagnosis, and therapy in the gastrointestinal tract are done with the same tool: the colonoscopy. However, it is a painful examination and poorly tolerated by patients. The colonoscopy is invasive and needs anesthesia, a specialist and a controlled environment. Furthermore, the colonoscopy doesn't allow the visibility of all the regions near the colon. Another method, the colorectal tomography (CTC), is non-invasive, but it can not detect polyps less than 1cm and patients are exposed to radiation [14].

In many countries, the screening process starts with a test, a Faecal Occult Blood Test (FOBT) or a Fecal Immunochemical Test (FIT). The need for a colonoscopy is determined by these tests. However, FOBT has a poor sensitivity of only 38% [1]. For the FIT test sensibility depends on the calibration of $\mu g/g$ of blood. A study shows a variation from 89% for a FIT calibrated to detect less than 20 $\mu g/g$ of blood to 70% if it is calibrated to detect 20 to 50 $\mu g/g$ of blood [17]. Decreasing the sensibility of FIT test increases its specificity and reduce the number of useless colonoscopies. Thus, it is preferred to increase the specificity and decrease the sensibility of the FIT test and realize a periodic screening.

As exposed, there is a need for a screening tool with high sensibility and high specificity. In 2003 Paul Swain and al [20] have introduced Wireless Capsule Endoscopy (WCE), a simple pill that patient swallows and that transmits images of the gastrointestinal tract via a Radio Frequency communication through the body. More than 1.6 million patients worldwide have used this technology for the small bowel, esophagus, and colon (more than 125,000 procedures a year). The available WCEs for the colon [14], example are visible in table I, has a length of 31 mm and diameter of 11 mm, battery life of 10 hours, a resolution of 256x256 pixels and an image sampling rate around 2 to 4 frames per second.

TABLE I
Commercially available WCE for Colon

Manufacturer	PillCam COLON	PillCam COLON2
Size[mm](Length x diameter)	31x11	31x11
Battery [h]	10	10
Image Resolution[pixels]	256x256	256x256
Image Sampling rate [fps]	4	4-35

Our idea is to integrate inside a WCE an intelligent image processing that can detect a polyp. Question is what is the good image processing?

As in many domains, Convolutional Neural Network (CNN) demonstrates its capacity to detect polyp lesions.

A. Polyps and CNN state of the art

In [12], AlexNet model [15] pre-trained on the ILSVRC 2012 dataset was used to detect polyps. The input was modified to an image size of 96x96. Furthermore, the kernel size of the two first pooling layers is decreased from 3 to 2 and the last pooling layer is removed to modify the output layer for two outputs: polyp or non-polyp. They increased the number of examples applying random mirroring, rotation, up- and down-scaling, cropping, and brightness adjustment in the original database. They use a sliding-window strategy to determine the polyp presence or absence in a video sequence. They evaluate their CNN performance in a dataset of 120 frames (60 with a polyp), they use 80 images (40 with a polyp) to train and the rest to testing, their experimentation has shown an accuracy of 60%.

In [21], authors use three CNN trained at different image scales (x1, x0.5, x0.25). They remove the last fully connected layer, the outputs of the convolutional layer of each CNN are fed as input to a single Multi-Layer Perceptron (MLP) network that is trained separately. The training was performed exclusively on the CVC-CLINIC(Computer Vision Center/Universitat Autonoma de Barcelona and Hospital Clinic from Barcelona, Spain) database composed by 1200 WCE images. Such a CNN achieves an accuracy classification of 90%.

Despite CNN achieve very good performance on polyp detection, all these methods are running on an external computer and contribute to helping the physician in his diagnosis. Unfortunately, they are not useable to be integrated inside a WCE. Indeed, they do not consider WCE constraints such as real-time execution, the form factor of the pill, energy consumption. Especially, CNN methods use a great number of parameters, that are not implantable in an 8x8 mm^2 chip inside a pill.

How it is possible to use CNN, such a powerful tool, to embed its processing inside an iWCE ?

B. Reducing a CNN

One potential solution is to reduce the number of parameters in CNN. In [16], Lecun and his team were among the first using pruning to reduce the network complexity and over-fitting. The idea of pruning is to reduce the size of a network by deleting unnecessary weights. Recently, authors in [10], applied this method on state of the art CNN like AlexNet or VGG16, with no loss of accuracy on PASCAL, COCO, KITTI, and ImageNet datasets. Moreover, in [9] quantization method and weight sharing compress the network through the reduction of the required number of bits to represent each weight. More than quantization, they also applied pruning and Huffman coding. They reduced the number of parameters of Alexnet by 9x, from 61 million to 6.7 million and the number of parameters of VGG16 by 13x, from 138 million to 10.3 million.

These parameter reduction examples of a CNN are focused on the software part, but improvement can also be done on the hardware part. In [8] with their energy efficient inference engine, they accelerated the resulting sparse matrix-vector multiplication with the weight sharing method. They considerably improve the speed and energy efficiency compared to CPU and GPU.

Finally, a method named knowledge distillation, whose feasibility was first shown by the team of Bucilla [4], has been used on the resolution of eight problems with excellent results and negligible losses due to size reduction of the CNN.

The obtained compressed network, on average, was a thousand times smaller and a thousand times faster than the original network. It was trained by mimicking the output of the original network. The team of Hinton [11] tested this methodology on the MNIST database and for the processing of natural language issues and achieved good results in applying the prediction of the teacher network as a soft label. Then, Chen [5] used distillation for object detection on datasets as PASCAL, KITTI, ILSVRC2014 and MS-COCO. They reduced the required storage for AlexNet by a factor of 35 and the required storage for VGG16 by a factor of 49.

The purpose of this article is to describe our first work to study the possibility to integrate a CNN inside a WCE to detect polyps. We present our study based on a Deep Learning algorithm, a Convolutional Neural Network (CNN) and we apply the knowledge distillation method to a smaller CNN.

II. A CNN TO DETECT POLYPS

We first train a CNN, the VGGnet [19], invented by Visual Geometry Group from the University of Oxford, to classify images in two classes: with polyps and without polyp. The architecture of this CNN is shown in Figure 1. This CNN was the first runner-up of the 2014 ILSVRC Contest, which is the first year that deep learning models obtained the error rate under 10%, with a Top-5 error-rate of 8.8% for VGG16. It has 138 million parameters. This state of the art network is one of the biggest and regards to the distillation method, it could be better to start with a big network, which should have a good performance on polyp classification.

We modify the last layer of VGG16, the fully-connected layer, with a structure with two outputs dedicated determining the presence of a polyp.

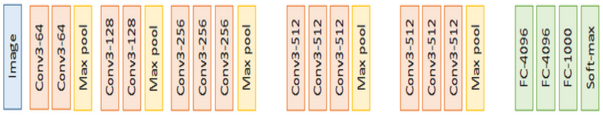

Fig. 1. The teacher network : VGG16

To train this network we use a 11952 images database divided into 10023 images with polyp and 1929 images without polyp. These images are issued from 18 videos of endoscopic examinations of Hospital Clinic, Barcelona, Spain. This dataset was used in EndoVisSub2017-GIANA contest [3]. Each image is associated with a ground-truth: a binary-image that indicates the position of the polyp in the image.

We use 70% of the dataset to train VGG16, and the remaining 30% to test its performance. The result shows an accuracy of more than 98%. Although we have achieved very good accuracy, it is impossible to embed this convolutional neural network with one hundred and thirty height million parameters into a WCE.

We then propose to use the distillation method to do so for polyp detection.

III. Proposed method

Previous works on the knowledge distillation method, [4], [5], [11], show that it is possible for a faster-compressed network to be able to approximate the function learned previously by a larger and slower network, but intrinsically more efficient. We apply the distillation method N times to observe the limit of accuracy we can obtain with state of the art CNN. The CNN generated by the distillation is named the student network and it will replace the teacher network, here VGG16, in each further iteration.

The efficiency of CNN was measured with two important metrics: sensibility (1), it is the image rate with polyp correctly predict and specificity (2), it is the image rate without polyp correctly predict.

$$Sensibility = \frac{VP}{VP+FP} \tag{1}$$

$$Specificity = \frac{VN}{VN+FN} \tag{2}$$

Where :

- TP = True Positive, number of images with polyp which is correctly classified.
- FP = False Positive, number of images with polyp which is not correctly classified.
- TN = True Negative, number of images without polyp which is correctly classified.
- FN = False Negative, number of images without polyp which is not correctly classified.

We can compute the Accuracy, that is the global performance of the CNN, as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

A. The teacher network: VGG16

We have trained VGG16, the teacher network, which will allow the training of another small network. The training and validation dataset was composed as follow:

- training set: 5857 images including 4910 with polyp(s) and 947 without polyp.
- validation set: 2509 images including 2106 with polyp(s) and 403 without polyp.

It was tested on five test datasets. Each of these was randomly created from our polyp images database precedently described. Thanks to these five datasets, we did cross-validation of the accuracy of our teacher network. These test datasets are composed of 3586 images, including 3008 with polyp(s) and 578 without polyp. It represents 30% of the entire database. Every test, even those on small networks will be done on these five test datasets to keep the same test reference. Test results are visible on table II for the teacher network VGG16:

TABLE II
VGG16 polyp detection

VGG16	Dataset test 1	Dataset test 2	Dataset test 3	Dataset test 4	Dataset test 5	Mean
Accuracy	0.9908	0.9889	0.9925	0.9877	0.9869	0.9894
Sensibility	0.9914	0.9897	0.9917	0.99	0.9897	0.9905
Specificity	0.9879	0.9845	0.9965	0.9758	0.9724	0.9834

As shown on table II, VGG16 is excellent to detect polyp. The sensibility and the specificity are respectively 99% and 98.3%. As we wanted, this network could be a good teacher network to apply distillation.

B. The student network: Squeezenet

To be sure of the efficiency of distillation method on polyp detection, we first train and test on one of the state-of-the-art convolutional neural networks, Squeezenet [13], by DeepScale, UC Berkeley and Stanford University. The architecture of Squeezenet is based on fire module, composed by a squeeze convolution layer with only 1x1 filters and expand layer that has a mix of 1x1 and 3x3 filters. This architecture is represented in figure 2. Moreover, this network has around one million parameters without compression, far less than VGG16.

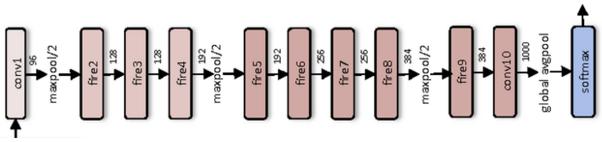

Fig. 2. The first student network : Squeezenet

We first compare the validation results of Squeezenet between two datasets:

- our original dataset, precedently described
- our original dataset, labeled by trained VGG16

We trained Squeezenet on both sets during 500 epochs to observe the behavior of the network on a large scale. Results show a better accuracy for Squeezenet trained on the dataset labeled by VGG16, namely 0.9829 compare to accuracy for Squeezenet trained on the original dataset, namely 0.9785. But the difference is only 0.0044 which is very small and shows that distillation allows training efficiently a student network. Figure 3 underline the influence of distillation on validation results.

The blue curve represents Squeezenet accuracy on validation set, previously trained on dataset labelled by VGG16. The orange curve represents Squeezenet accuracy on validation, previously trained on the original dataset. Both reach the beginning of convergence after twenty epochs but the blue curve needs around fifty epochs to reach stabilization compared to forty epochs for the other. Figure 3 allows us to determine the right number of epochs to obtain the best performance

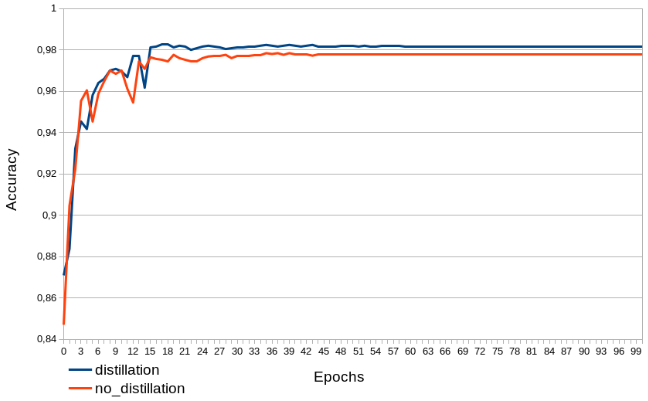

Fig. 3. Distillation influence on Squeezenet validation results

during the test. In our distillation case, theoretically, best performance should be raised after training Squeezenet around thirty epochs, just before stabilization.

Then we test Squeezenet with distillation on the five test datasets. Results are shown in table III.

TABLE III
SQUEEZENET WITH DISTILLATION

Squeezenet	Dataset test 1	Dataset test 2	Dataset test 3	Dataset test 4	Dataset test 5	Mean
Accuracy	0.9824	0.9780	0.9816	0.978	0.9769	0.9794
Sensibility	0.9904	0.9870	0.9877	0.9870	0.9844	0.9873
Specificity	0.9413	0.9309	0.9499	0.9309	0.9378	0.9382

As shown in table III, Squeezenet with distillation has a mean specificity of 93.82% and a mean sensibility of 98.73%. Performances demonstrate a good result for polyp detection with a complexity reduction of 2 order of magnitude. Even if there is still too much parameter to be embedded inside a pill, this result shows a clear way to define a smaller CNN that can be integrable inside a WCE.

IV. Conclusion

CNNs have demonstrated their powerful generalization capacity for the particular task of colorectal cancer detection. Integrated inside a WCE, CNNs are successfully used for efficient early diagnosis of this disease. However, creating embedded CNNs on limited hardware resources is not trivial and requires a careful design in order to reduce the complexity of the original CNNs. Our solution presented in this paper relies on distillation and makes it possible to reduce the complexity of CNNs by two orders of magnitude with only 1% loss of sensibility and 5% loss of specificity. This encouraging result opens the way to compress CNNs with at least one additional order of magnitude and to make them embeddable [2].

References

[1] Allison, J. E., Tekawa, I. S., Ransom, L. J., and Adrain, A. L. A comparison of fecal occult-blood tests for colorectal-cancer screening. *New England Journal of Medicine 334*, 3 (1996), 155–160. PMID: 8531970.

[2] Andrea Pinna, Annick Alexandre, S. V. E. B. P. G. B. G. Connectionist retina: a neural networks system integrated into an electronic retina. vol. 4948.

[3] Bernal, J. J., Histace, A., Masana, M., Angermann, Q., Sánchez-Montes, C., Rodriguez, C., Hammami, M., Garcia-Rodriguez, A., Córdova, H., Romain, O., Fernández-Esparrach, G., Dray, X., and Sanchez, J. Polyp Detection Benchmark in Colonoscopy Videos using GTCreator: A Novel Fully Configurable Tool for Easy and Fast Annotation of Image Databases. In *Proceedings of 32nd CARS conference* (Berlin, Germany, June 2018).

[4] Bucilu, C., Caruana, R., and Niculescu-Mizil, A. Model Compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2006), KDD '06, ACM, pp. 535–541.

[5] Chen, G., Choi, W., Yu, X., Han, T., and Chandraker, M. Learning Efficient Object Detection Models with Knowledge Distillation. 10.

[6] Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer 136*, 5 (2015), E359–E386.

[7] Halloran, S., Launoy, G., and Zappa, M. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First Edition Faecal occult blood testing. *Endoscopy 44*, S 03 (Sept. 2012), SE65–SE87.

[8] Han, S., Liu, X., Mao, H., Pu, J., Pedram, A., Horowitz, M. A., and Dally, W. J. EIE: Efficient Inference Engine on Compressed Deep Neural Network. In *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)* (June 2016), pp. 243–254.

[9] Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, 2015.

[10] Han, S., Pool, J., Tran, J., and Dally, W. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems* (2015), pp. 1135–1143.

[11] Hinton, G., Vinyals, O., and Dean, J. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531 [cs, stat]* (Mar. 2015). arXiv: 1503.02531.

[12] Hwang, S. Bag-of-visual-words approach based on surf features to polyp detection in wireless capsule endoscopy videos. In *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)* (2011), The Steering Committee of The World Congress in Computer Science, Computer , p. 1.

[13] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size, 2016.

[14] Karargyris, A., and Bourbakis, N. Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos. *IEEE transactions on BioMedical Engineering 58*, 10 (2011), 2777–2786.

[15] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105.

[16] LeCun, Y., Denker, J. S., and Solla, S. A. Optimal brain damage. In *Advances in neural information processing systems* (1990), pp. 598–605.

[17] Lee, J. K., Liles, E. G., Bent, S., Levin, T. R., and Corley, D. A. Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. *Annals of internal medicine 160*, 3 (2014), 171–181.

[18] QLQC30, E. Aaronson nk, ahmedzai s, bergman b, bullinger m, cull a, duez nj, et al. the european organization for research and treatment of cancer qlq-c30: a qualityof-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute 85*, 5 (1993), 365–76.

[19] Simonyan, K., and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (Sept. 2014). arXiv: 1409.1556.

[20] Swain, P. Wireless capsule endoscopy. *Gut 52*, 90004 (June 2003), 48iv–50.

[21] Zhao, Q., Dassopoulos, T., Mullin, G. E., Meng, M., and Kumar, R. A decision fusion strategy for polyp detection in capsule endoscopy. *Studies in health technology and informatics 173* (2012), 559–565.