An Inception-like CNN Architecture for GI Disease and Anatomical Landmark Classification

Stefan Petscharnig, Klaus Schöffmann, Mathias Lux
Alpen-Adria-Universität Klagenfurt
first.lastname@itec.aau.at

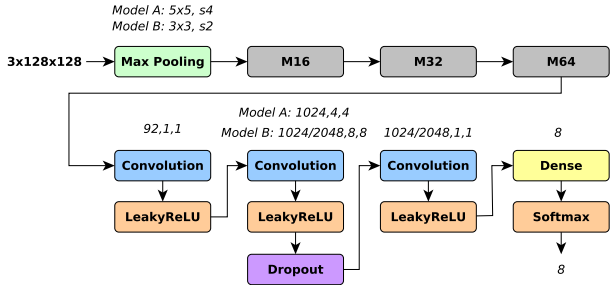
ABSTRACT

In this working note, we describe our approach to gastrointestinal disease and anatomical landmark classification for the Medico task at MediaEval 2017. We propose an inception-like CNN architecture and a fixed-crop data augmentation scheme for training and testing. The architecture is based on GoogLeNet and designed to keep the number of trainable parameters and its computational overhead small. Preliminary experiments show that the architecture is able to learn the classification problem from scratch using a tiny fraction of the provided training data only.

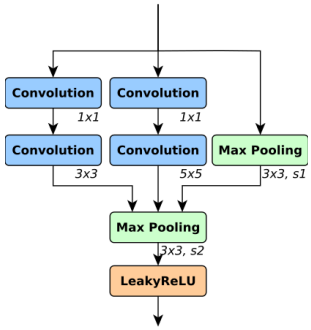
1 INTRODUCTION

With recent developments in computer vision, it seems only natural to transfer the progress to the domain of medical imaging. However, in this domain where neither machines nor domain-experts provide flawless results [8], there is a need for specialized methods in order to make a significant leap forward at tasks such as computer aided diagnosis. The Medico task at MediaEval 2017 [9] aims to improve methods for multimedia-assisted diagnosis in the domain of endoscopic imaging for the special case of the gastrointestinal (GI) tract. More precisely, participants of this task should develop approaches for GI disease and anatomical landmark detection. The goal of the task is (efficient) classification of diseases with as little training data as possible. To cope with this task, a rich training dataset comprising of 4000 images (500 per class) [7] was made available by the task organizers. The individual performance is benchmarked on a similarily sized test dataset. Recently, much effort in the field of deep learning in medical image analysis has been conducted. For the use case of surgical action and anatomical structure recognition in laparoscopic interventions, different off-the-shelf architectures have been investigated by our research group [4, 5]. For the use case of cholecystectomy, Twinanda et al. [12] altered a well-known CNN architecture to suit the domain and use-case of temporal segmentation. Within medical image analysis of the gastrointestinal (GI) tract, Pogorelov et al. [6] provide a system for disease detection. Automated polyp detection in colonoscopy videos was proposed by Tajbakhsh et al. [11]. For more information on automated please refer to Litjens et al. [2]. In this paper, we propose an efficient inception-like CNN architecture, which is capable of learning the classification problem from scratch using only as small amount of training data. To achieve this goal, we propose a crop-based data augmentation scheme, which is tailored to the use case of GI disease detection. We furthermore propose a variant of our architecture increasing predictive performance at the expense of increased computational cost and number of trainable parameters (model size).

MediaEval '17, 13-15 September 2017, Dublin, Ireland



(a) Overview on the two variants of the proposed CNN Architecture with max-pooling input preparation, stacked inception-like modules for feature extraction and classification part.



(b) Inception-like module structure.

Figure 1: Proposed inception-like CNN architecture.

2 APPROACH

We approach this problem by defining a CNN architecture that is capable of learning the distinction of a relatively small amount of classes from a small training set. We base our work on GoogLeNet [10], an already existing CNN architecture, which yields a decent performance in many tasks. Its prominent architectural feature is the inception module. The basic idea behind the inception module is that the network may select at training time whether pooling, small convolution, or wider convolution suits the underlying data best. Therefore, the aforementioned operations are calculated in parallel and their results are merged and the feature dimensionality is reduced. Our inception module is depicted in Figure 1b and consists of three main branches: small convolution, large convolution, and pooling. We also use 1x1 convolutions before the main convolutional layers in order to reduce computational cost. Furthermore, we use padding in order to preserve the size of the feature maps. After the main branches are merged channel-wise, we use max-pooling with a stride of 2 for dimensionality reduction, reducing feature map size by a factor of 4. After pooling, we use a LeakyReLU [3] with a negative slope of 0.01 as non-linear activation function. We refer to such inception-like modules as *MX* where X denotes the number of

learned filers per convolutional layer. Whereas GoogLeNet features a 1x1 convolution branch, our preliminary experiments showed that there is no performance gain for this specific domain, thus we do not use a fourth branch. Furthermore, we only use activation functions at the end of each module and not after each convolutional layer and skip the 1x1 convolution after the pooling branch. A graphical overview on our proposed architecture is given in Figure 1a. The first layer of the proposed architecture is overlapping max-pooling. Please note that we propose two different variants of the architecture: Model A and Model B. The difference between the variants lies in the first (max pooling) layer: Model A uses a stride of 4 and max-pooling window size 5, whereas Model B uses stride 2 window size 3. We then use three stacked inception-like modules as feature extraction stage. The number of convolutions per convolutional layer is doubled the deeper we delve into the network to compensate for the spatial reduction. The feature extraction stage is followed by a 1x1 convolution with 92 learned filters. This convolutional step is a further dimensionality reduction in the channels axis (whereas the inception-like modules reduce the dimensionality in the spatial axis). The second convolution reduces the feature map to a spatial size of 1x1. We also experiment with the number of channels in the late stages and therefore, the output of this layer has - depending on the actual choice - 1024 or 2048 channels. The convolution size for this layer is dependent of the model variant (4x4 and 8x8 for Model A and Model B respectively). The only regularization technique we use is a dropout layer (with a dropout chance of 0.2). This is followed by a further 1x1 convolution layer and a dense layer with softmax activation. Henceforth, we refer to our architecture variant with a combination of model identifier (A or B) number of neurons in deep layers (1024 or 2048) and percent of training data used, e.g., $B_{1024-10}$ refers to model B (3x3 pooling at low network size and 8x8 convolution before dropout) with 1024 neurons in the deep layers and was trained on 10 percent of the training data. Input to our proposed network is a 128x128 image patch. We augment the training set by extracting seven different image patches according to Figure 2 and resizing them to the input shape. The patch selection was motivated from the consideration that the most important image details are in the center of the image. We furthermore extract the patches from three different scales. Furthermore, we randomly mirror the patches at training time. We standardize the training set by subtracting the mean image pixel. For testing, we extract the a set image patches from each testing image the same way as we augment our test data (see Figure 2). We classify each of these patches and aggregate the results by using a simple average.

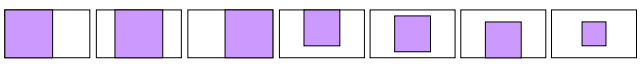


Figure 2: Crops used for training augmentation and testing.

3 RESULTS AND ANALYSIS

We used three different variants for the tasks: Model A as well as Model B with 1024 and 2048 neurons in the deep layers. An overview on the individual results is given in Table 1. All in all, we observe the trend that generalization performance increases with more training data available. Generally, we discover that all the models are confusing dyed resection margins with dyed-lifted-polyps as well as polyps with ulcerative-colitis. We argue that this weakness originates in the choice of training data augmentation: polyps and resection margins are not always visible on center-like crops. Our models also show minor weaknesses at distinguishing normal-z-line from esophagitis. In preliminary experiments, we also tried to distinguish these selected classes with a binary CNN classifier and the fusion of global features at a deep level, but this did not improve results. Model A is used in the speed runs. Its strength is the small number of parameters (2.8M against 7.3M for Model B_{1024}, 16.5M for Model B_{2048}) and its small computational cost. We measure forward passes over 1000 iterations using a GeForce GTX Titan X (maxwell) graphics card. The model takes 2.25ms per forward pass, in contrast to 2.91ms and 3.42ms for Model B_{1024} and B_{2048} respectively. Caffenet [1], an AlexNet variant takes 3.27ms per forward pass, the computation time GoogLeNet 14.16ms. Variants from Model B are used in the detection runs and As baseline in the speed run. Interestingly, we observe that the lower capacity model B_{1024} is superior to B_{2048}. These results were surprising on first view, as our preliminary evaluations indicated the opposite. We conclude that the larger model tends to better adapt to the training data. Thus, the model suffers from over-fitting and yields smaller generalization performance.

Table 1: Medico '17 benchmark results on the speed and detection subtasks. Bold values indicate best predictive performance per subtask.

Speed	REC	PREC	SPEC	ACC	F1	MCC	RK
$A_{1024-10}$	0.687	0.687	0.955	0.922	0.687	0.642	0.643
$A_{1024-50}$	0.706	0.706	0.958	0.927	0.706	0.664	0.672
$A_{1024-90}$	0.727	0.727	0.961	0.932	0.727	0.688	0.695
$B_{1024-90}$	**0.755**	**0.755**	**0.965**	**0.939**	**0.755**	**0.720**	**0.724**
Detection	REC	PREC	SPEC	ACC	F1	MCC	RK
$B_{1024-10}$	0.649	0.649	0.950	0.912	0.649	0.599	0.607
$B_{1024-50}$	0.750	0.750	0.964	0.938	0.750	0.715	0.717
$B_{1024-90}$	**0.755**	**0.755**	**0.965**	**0.939**	**0.755**	**0.720**	**0.724**
$B_{2048-50}$	0.740	0.740	0.963	0.935	0.740	0.703	0.705
$B_{2048-90}$	0.747	0.747	0.964	0.937	0.747	0.710	0.715

4 DISCUSSION AND OUTLOOK

We provide a CNN architecture capable of learning classification with little training data. The proposed architecture is able to provide acceptable results with even as little as 50 training examples per class. The presented data augmentation and testing method using a fixed set of selected crops per image is beneficial for the overall performance. In preliminary experiments, we also investigated late fusion of global feature to the CNN architecture which did not lead to significant performance improvement. For future work, we want to investigate two main aspects: (1) whether the architecture is performing well in the domain of laparoscopic surgery, and (2) whether the model is capable of efficiently dealing with more input channels for early fusion of temporal information which can be used in action recognition in laparoscopic surgery.

ACKNOWLEDGMENTS

This work was supported by Universität Klagenfurt and Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF 20214 u. 3520/ 26336/38165.

REFERENCES

[1] Jeff Donahue. 2014. BVLC Caffenet. (2014). https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet Online, Accessed: 2017-09-54.

[2] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42 (2017), 60 – 88. https://doi.org/10.1016/j.media.2017.07.005

[3] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, Vol. 30.

[4] Stefan Petscharnig and Klaus Schoeffmann. 2017. Deep Learning of Shot Classification in Gynecologic Surgery Videos. In *International Conference on Multimedia Modeling*, Laurent Amsaleg, Gylfi Þór Guð mundsson, Cathal Gurrin, Björn Þór Jónsson, and Shin'ichi Satoh (Eds.). Springer, Cham, 702–713.

[5] Stefan Petscharnig and Klaus Schoeffmann. 2017. Learning laparoscopic video shot classification for gynecological surgery. *Multimedia Tools and Applications* (apr 2017), 1–19. https://doi.org/10.1007/s11042-017-4699-5

[6] Konstantin Pogorelov, Sigrun Losada Eskeland, Thomas de Lange, Carsten Griwodz, Kristin Ranheim Randel, Håkon Kvale Stensland, Duc-Tien Dang-Nguyen, Concetto Spampinato, Dag Johansen, Michael Riegler, and Pål Halvorsen. 2017. A Holistic Multimedia System for Gastrointestinal Tract Disease Detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*. ACM, New York, NY, USA, 112–123. https://doi.org/10.1145/3083187.3083189

[7] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*. ACM, New York, NY, USA, 164–169. http://dx.doi.org/10.1145/3083187.3083212

[8] Michael Riegler, Mathias Lux, Carsten Griwodz, Concetto Spampinato, Thomas de Lange, Sigrun L. Eskeland, Konstantin Pogorelov, Wallapak Tavanapong, Peter T. Schmidt, Cathal Gurrin, Dag Johansen, Håvard Johansen, and Pål Halvorsen. 2016. Multimedia and Medicine: Teammates for Better Disease Detection and Survival. In *Proceedings of the 2016 ACM on Multimedia Conference (MM '16)*. ACM, New York, NY, USA, 968–977. https://doi.org/10.1145/2964284.2976760

[9] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Carsten Griwodz, Thomas de Lange, Kristin Ranheim Randel, Sigrun Losada Eskeland, Duc-Tien Dang-Nguyen, Mathias Lux, and Cibcetti Spampinato. 2017. Multimedia for Medicine: The Medico Task at MediaEval 2017. In *Proc of the MediaEval 2017 Workshop*. Dublin, Ireland, Sept. 13-15, 2017.

[10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1–9.

[11] N. Tajbakhsh, S. R. Gurudu, and J. Liang. 2016. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *IEEE Transactions on Medical Imaging* 35, 2 (Feb 2016), 630–644. https://doi.org/10.1109/TMI.2015.2487997

[12] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy. 2017. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Transactions on Medical Imaging* 36, 1 (Jan 2017), 86–97. https://doi.org/10.1109/TMI.2016.2593957