2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Automated Annotator: Capturing Expert Knowledge for Free

Sebastian Elmes[1] Tapabrata Chakraborti[1] Mengran Fan[1] Holm Uhlig[3] Jens Rittscher[1,2,3]

***Abstract*— Deep learning enabled medical image analysis is heavily reliant on expert annotations which is costly. We present a simple yet effective automated annotation pipeline that uses autoencoder based heatmaps to exploit high level information that can be extracted from a histology viewer in an unobtrusive fashion. By predicting heatmaps on unseen images the model effectively acts like a robot annotator. The method is demonstrated in the context of coeliac disease histology images in this initial work, but the approach is task agnostic and may be used for other medical image annotation applications.The results are evaluated by a pathologist and also empirically using a deep network for coeliac disease classification. Initial results using this simple but effective approach are encouraging and merit further investigation, specially considering the possibility of scaling this up to a large number of users.**

***Index Terms*— automated annotation, explainable deeplearning, autoencoder, heatmap visualisation, coeliac disease**

I. INTRODUCTION

Deep learning based medical image analysis has reached near human accuracy in a range of segmentation/classification tasks [1], [2]. But supervised deep learning requires large amount of labeled data for robust and generalised performance, which in turn needs significant amount of annotations by human experts [3]. This is particularly costly for medical imaging applications, as it is difficult to have access to the time of clinical experts. The need of the hour is a machine learning tool that can be trained from a limited number of expert annotations available and then learn to bootstrap that for automated annotation of incoming new images.

The present work does exactly that. Only information that can be obtained from image viewers in an unintrusive fashion is being utilised. We present a simple yet effective pipeline using kernel density estimation of autoencoder generated heatmaps. The heatmaps produced are used to annotate further images and then subsequently used for classification of coeliacs disease as an exemplar test case. This was chosen because it involves the assessment of separate and clearly defined regions and as a result attention should be clearly targeted to those regions as well [4]. This means that an attention heatmap would be more informative than in the case where the pathologist is required to look more broadly across the image. Coeliac disease is a condition whereby the ingestion of gluten triggers an immune response which attacks the small intestine. It has a prevalence of approximately 1% of the populations of the US and Europe and symptoms include diarrhea and abdominal pain [5]. In addition to clinical and serological examination, histopathological examination plays a major role in its diagnosis. Biopsies are taken of the small bowel and imaged. Diagnosis is based on the analysis of two main features: the structure of the crypts and villi, and the number of lymphocytes present [6].

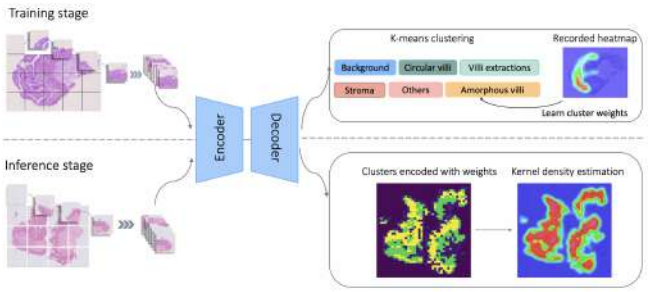


Fig. 1: **Schematic of the proposed automated annotator**

Recently, the idea of computer assisted annotations for medical imaging tasks [7], [8] as been explored. But the reliability of these annotation methods [9] is still an open problem of research and needs pathologist validation. The main drawback of the existing works on integration with a small set of expert annotations with a machine learning based automated annotator is the human experts only focus on the ROI (region of interest) of the image for decision and disregard most of the other regions, as a result of which heatmaps with sharp edges are created which are less specific [10], [11]. Thus a focus of attention of model with uneven probability distribution for attention is needed resulting in a smooth heatmap for ROIs, and the present work provides exactly that.

The results presented in this report support the claim that the proposed technology can be used in two different settings:

- **Monitoring pathologists.** As this methodology only requires user input that can be extracted remotely it will be possible to run this at a large scale involving several experts without imposing any restrictions on their regular workflow. This makes the generation of large data sets possible at low cost.
- **Enhancing machine learning algorithms.** The ability

[1]Institute of Biomedical Engineering (IBME) and the Big Data Institute (BDI), Dept. of Engineering Science, University of Oxford, Oxford, UK
[2]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, Oxfordshire, UK
[3]Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK
*MF received financial support from the Arthritis Therapy Programme (A-TAP) funded by the Kennedy Trust. JR and TC are supported by the Innovate UK PathLAKE Consortium and UKRI Innovate UK DART Porgramme. TC is also supported by the Oxford CRUK Cancer Centre.

to make accurate predictions for heatmaps using a small number of training samples could be used to produce large numbers of weakly annotated data points. The validation results also give a strong indication that the predicted heatmaps are accurate enough to be useful as a prior for the focus of attention of an image classification algorithm.

II. Methodology and Experimental Setup

Here we present the data preprocessing, model architecture and training protocol. The method is summaried in Figure 1.

A. Data preparation and preprocessing

30 images of small bowel biopsies, collected by through the authors' affiliated hospital, were used to demonstrate the automated annotation tool. Of these, 24 were used to train the algorithm to generate heatmaps, which were then used to test the remaining six. 5-fold cross validation was performed whereby training was performed and expert annotations were obtained by a pathologist for groundtruth. The main discriminating regions were the villi, stroma, crypts and lymphocytes, in that order. The images were segmented into 128×128 pixel tiles. This size was chosen as it was found to be at a similar scale to the width of a villus extrusion. This meant that villi could be identified as a feature and be separated from the rest of the image as a region of interest. This also meant that the effective number of patch level training samples generated by the process was large enough to train the model. The tiling procedure generated a total of 40,000 patches, which were clustered into 8 categories. Though the number of samples belonging to the background cluster consituted around 50% of the samples. The other categories (villi, crypt, stroma, etc.) were more evenly distributed, with at least 1000 patches per cluster, which turned out to be enough for effective training of the autoencoder model. The 40,000 image segments were divided into eight clusters using the k-means algorithm. The five largest clusters could broadly be defined as follows: background - 21,344 segments, villi extrusions - 4,615 segments, clear circular villi cross sections - 2,830 segments, amorphous villi cross sections - 5,110 segments, stroma - 2,782 segments. The content of the segments in the remaining three clusters (of sizes 1,175, 1,073 and 1,071 segments) was less clearly defined upon visual inspection. These were most commonly found around the edges of sections of tissue and may have contained some combination of the features from the larger clusters.

B. Autoencoder architecture and training

A feature vector was generated for each patch using a convolutional autoencoder. The autoencoder architecture consisted of alternating layers of convolution (with 5×5 filters and ReLu activation) and max pooling layers (downsample by factor of 2). Five stages of this using varying numbers of filters at each stage resulted in a 4 by 4 by 1 encoded version of each segment. This final size was chosen as a balance between being large enough to generate a complex representation of each segment without being so large that the computation would suffer due to its high dimensionality. Five layers of transposed convolution with a stride of 2 were used to form the decoder. Some of the training settings used are listed below:

- 80% of the patches were used to train the model and the remaining 20% to validate the model on unseen patches and check for overfitting and parameter tuning.
- He initialization [12] was used for the convolutional layers to keep the initial variance of each of the layers equal.
- The autoencoder was optimised by minimising the pixel wise mean squared error between the input segment and its decoded reconstruction. It measures pixel by pixel how different the reconstruction is from the input image. This, along with the l1 regularization term constituted the loss function to be minimised.
- The Adam optimiser was used [13] with an adaptive learning rate. Mini-batches of 128 samples of the training data were used on each iteration. Optimization was terminated after 20 epochs (runs through the data).

III. Heatmap generation and evaluation

A. Heatmap Generation

The training patches were run through the encoder stage of the autoencoder. The euclidean distance from the encoded version (or feature vector) of each segment to each of the cluster centres was calculated and the segment was assigned to the closest one. The recorded heatmaps were integrated over the segments in each cluster to give the cluster's weightings. This was repeated for each image and the results summed to find a total weighting for each cluster. The weighting of the cluster which represented the image backgrounds was set to zero.

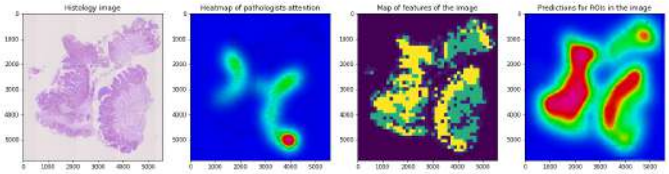


Fig. 2: **An example input image and the results of the three main objectives of the project:** a heatmap showing where a pathologist looked in the image, a colour coded map of the image indicating the physiological feature at each location and a heatmap showing predictions for the locations of ROIs in the image.

Attention heatmaps were generated as the pathologist analysed the images by recording which part of the image was in the centre of the image viewer and how long it was held there for. Kernel density estimation was used to convert this to a probability density function for attention at a given location. This was to reflect the chance that the pathologist was not looking directly at the centre of the screen. A probability at each position was calculated by summing the values of surrounding data points (or samples).

Each sample was weighted according to how close it is to the position being evaluated and by how long the viewer was held there for. The kernel, in this case, is the function which calculates the weighting for each sample [14]. Several different functions are commonly used and it has been shown that the choice of function only has a small effect on the accuracy of the algorithm when trained to learn known distributions [15]. As the kernel density estimate is a convolution of the data with the kernel, Fourier transforms can be used for more efficient computation [15]. Using a Gaussian kernel has computational benefits therefore as its Fourier transform can be found explicitly and so was chosen for use in this work.

Heatmaps were displayed by converting the image to HSV format and replacing the hue value with that of the heatmap. This meant that the heatmaps could be viewed overlayed with the image. Red delineated a region which received high amounts of attention and blue a region which received minimal amounts of attention.

B. Heatmap Prediction

The feature clusters were then used to predict attention heatmaps for an unseen image using the following method:

1) The image is segmented and a feature vector generated for each segment using the same filtering and dimensionality reduction process as the one described previously.
2) The segments are assigned to one of the seven clusters from the 'seen' image. This is done according to the least squared Euclidean distance between the segment's feature vector and the centres of the clusters.
3) The value of each cluster is assigned according to the cluster weightings from the first image. Cluster weightings are adjusted by raising them all to the power of τ where τ represents the specificity of the clusters chosen. As τ increases, the variance between cluster weightings will increase. Thus, high τ will result in only the most salient clusters being used and a more specific heatmap.
4) Kernel density estimation is used to generate the heatmap from this, taking the sample at each location in the image as the weighting of the cluster that the pixel in that location's segment had been assigned to.
5) Two parameters needed to be tuned at this stage: The specificity and the bandwidth of the kernel density estimation. Changing the specificity would vary how much contextual information is highlighted (i.e. tissue surrounding the villi extrusions) by controlling the weighting of the villi extrusions cluster relative to the others. Choice of this would depend on how diagnostically salient the contextual information was. Changing the bandwidth would vary how detailed the heatmap was. A small bandwidth would generate a heatmap which focused solely on the segments in the highest weighted clusters whereas a larger bandwidth would result in some overlap into the surrounding regions and thus a more conservative prior.

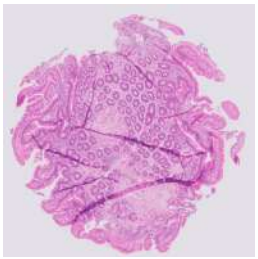
C. Heatmap Evaluation

We evaluate the heatmaps generated by both human expert and also through a classification tasks using a deep network. An illustrative example of results is provided in figure 2.

1) Human evaluation by expert: When validating the heatmaps, two types of error had to be tested for: false positive (when regions were highlighted which weren't in fact relevant to diagnosis) and false negative (when regions which were in fact relevant were ignored).

False negatives were measured by comparing the heatmaps predicted with those recorded from the pathologist. Success was defined by the percentage of the ROIs identified by the pathologist also highlighted as salient in the prediction. In the recorded heatmaps most of the image (70-80%) received no attention at all. This is expected since the pathologist could move straight to distinct regions of interest and ignore the rest of the image. An ROI was therefore defined as anywhere in the heatmap which had received attention. Across the 25 images tested (5 from each of the 5 cross validations) the average coverage with an 8 pixel bandwidth was 79% with a minimum coverage of 46% and a maximum of 100%.

False positives could not be identified simply by comparing the two heatmaps for an image. The pathologist was shown the full image and then cropped sections of the image which were potential false positives. These were then ranked 1-3 according to how informative they were for forming a diagnosis: 1 for completely irrelevant, 2 for somewhat informative and 3 for highly informative. 24 sections were analysed and it was found that of these: 11 (46%) were highly informative, 7 (29%) were somewhat informative and only 6 (25%) were completely irrelevant.



(a) **Tissue folding**: folds in the tissue lead to darker regions in the image.

(b) **Square outlines**: due to structure of cartridge holding tissue sample.

Fig. 3: **Examples of artefacts of the imaging process**.

2) Automated evaluation by deep network: The effectiveness of generated heatmaps was also assessed by a standard classification network to measure whether annotated regions could contribute to patch-based classification performance. As mentioned previously, clinical diagnosis is based primarily on the analysis of villi extrusion structures, and our previous results have demonstrated that villi regions can be identified and separated from the rest of the image patches. Therefore, we will use the generated heatmaps as a region of interest (RoI) detector to localise the most discriminative

patches within each whole slide image (WSI).

The WSIs were labeled by a pathologist based on modified Marsh score [18], which was developed to measure the severity of coeliac disease. Our goal was to classify the images into five grades: normal, Marsh II, March IIIa, March IIIb and March IIIc. The severity scores of WSIs were assigned to all their patches. We constructed two patch-based dataset. In the first dataset we randomly cropped 1000 patches of size 512 $\times$ 512 from each WSI. In the second dataset we extracted the centre of each villi segment and generated weakly annotated data points. As a result, a large number of 512 $\times$ 512 patches around the annotation points were cut. We trained our classifier with a MXNet implementation of ResNet-18 on two datasets. We used the SGD optimizer with a learning rate of 0.0001 and momentum of 0.9 for 200 epochs. We find that on the dataset with random cropping, the ResNet yields a classification accuracy of 51.2%, whereas when the dataset with heatmap based villi detection is used, the accuracy with the same ResNet backbone improves to 76.6%. We attribute the improvements to the localisation of villi regions, which successfully discarded irrelevant patches and enhanced the discrimination of patch appearance.

3) Analysis of failure cases: These results show that the majority of the salient regions of the images were identified, however there were still some errors. This was often either due to an artefact of the scanning process making it into the image (see figure 3) or simply because it was an unusual feature which was present in a large enough number of segments to form its own cluster. Two of the three images with the lowest coverage showed tissue folding, highlighting the impact of the failure to identify these rare occurrences. These problems could be fixed to some extent by using more data. With enough images there would be enough examples of even very rare features to form a clear representation of them. There are distinct types of artefact that are present in histology images and with enough examples the algorithm could learn to filter them out. Alternatively, if this fails, a classifier for specific artefacts could be learned so that images containing them could be flagged and these samples diagnosed by a pathologist.

IV. CONCLUSION

We present a machine learning enabled "automated annotator" for medical image analysis, that 1) learns from a limited set of human expert annotations obtained "free" in an unobtrusive manner through pathologist focus of attention tracking, and 2) utilises that knowledge to label incoming new samples which may then be used for subsequent classification/segmentation tasks. This helps to mitigate the age old prohibitive costs of expert annotation in health science for enough data to effectively train deep networks. The proposed method was tested on coeliac disease image data, but the results are encouraging enough for the method to be employed for other similar medical imaging tasks in the near future.

DECLARATION

JR is a co-founder of Ground Truth Labs, there are no other author conflicts. All relevant ethical approval and privacy consent were ensured for human patient data.

REFERENCES

[1] Dinggang Shen, Guorong Wu, and Heung-Il Suk, "Deeplearning in medical image analysis,"Annual Review of Biomedical Engineering, vol. 19, pp. 221–248, 2017.

[2] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi,Arnaud Arindra Adiyoso Setio, Francesco Ciompi,Mohsen Ghafoorian, Jeroen A.W.M.van der Laak,Bramvan Ginneken, and Clara I. Sanchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.

[3] Daisuke Komura and Shumpei Ishikawa,"Machinelearning methods for histopathological image analysis,"Computational and Structural Biotechnology Journal,vol. 16, pp. 34–42, 2018.

[4] Daniel Adelman, Joseph Murray, Tsung-Teh Wu, Markku Maki, Peter Green, and Ciaran Kelly, "Measur-ing change in small intestinal histology in patients withceliac disease,"The American Journal of Gastroenterol-ogy, vol. 113, pp. 339–347, 2018.

[5] Carlo Catassi, Simona Gatti, and Alessio Fasano, "Thenew epidemiology of celiac disease,"Journal of pedi-atric gastroenterology and nutrition, vol. 59 Suppl 1, pp.Suppl. S7–S9, 2014.

[6] Fei Bao, Peter Green, and Govind Bhagat, "An updateon celiac disease histopathology and the road ahead,"Archives of pathology and laboratory medicine, vol.136, pp. 735–745, 2012.

[7] Hyeonsoo Lee and Won-Ki Jeong,"Scribble2label:Scribble supervised cell segmentation via self-generating pseudo-labels with consistency,"Interna-tional Conference on Medical Image Computing andComputer-Assisted Intervention (MICCAI), pp. 14–23,2020.

[8] Haohan Li and Zhaozheng Yin, "Attention, suggestionand annotation: A deep active learning framework forbiomedical image segmentation,"International Con-ference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 3–13, 2020.

[9] Christian Marzahl et al., "Are fast labeling methods re-liable? a case study of computer-aided expert annota-tions on microscopy slides,"International Conferenceon Medical Image Computing and Computer-AssistedIntervention (MICCAI), pp. 24–32, 2020.

[10] Ezgi Mercan, Selim Aksoy, Linda G Shapiro, Donald LWeaver, Tad T Brunye, and Joann G Elmore, "Local-ization of diagnostically relevant regions of interest inwhole slide images: A comparative study,"Journal ofdigital imaging, vol. 29, no. 4, pp. 496–506, 2016.

[11] David Romo, Eduardo Romero, and Fabio Gonzalez,"Learning regions of interest from low level maps in vir-tual microscopy,"Diagnostic pathology, vol. 6 Suppl 1,pp. S22, 03 2011.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and JianSun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," inPro-ceedings of the IEEE International Conference on Com-puter Vision (ICCV), USA, 2015, p. 1026–1034.

[13] Diederik Kingma and Jimmy Ba, "Adam: A method forstochastic optimization,"International Conference onLearning Representations (ICLR), 2014.

[14] Stuart Russell and Peter Norvig,Artificial Intelligence:A Modern Approach, Book series in Artificial Intelli-gence. Prentice Hall, Upper Saddle River, NJ, third edi-tion, 2010.

[15] Bernard Silverman,Density estimation for statistics anddata analysis, Monographs on statistics and appliedprobability (Series) ; no. 26. London, 1986.

[16] Fogel and Dov Sagi, "Gabor filters as texture discriminator,"Biological Cybernetics, vol. 61, pp. 103–113,1989.

[17] Weitao Li, KeZhi Mao, Hong Zhang, and Tianyou Chai,"Designing compact gabor filter banks for efficient tex-ture feature extraction,"International Conference onControl Automation Robotics Vision, pp. 1193–1197,2010.

[18] Alessio Fasano and Carlo Catassi, "Current approachesto diagnosis and treatment of celiac disease: an evolvingspectrum,"Gastroenterology, vol. 120, no. 3, pp. 636–651, 2001.