Robust 3D U-Net Segmentation of Macular Holes

Jonathan Frawley\textsuperscript{1,2} \\
\textsuperscript{1} Department of Computer Science, Durham University, Durham, UK  
\textsuperscript{2} Inegral Limited, Durham, UK \\

Chris G. Willcocks\textsuperscript{1} \\

Maged Habib\textsuperscript{3} \\
\textsuperscript{3} Sunderland Eye Infirmary, Sunderland, UK \\

Caspar Geenen\textsuperscript{3} \\

David H. Steel\textsuperscript{3,4} \\
\textsuperscript{4} Newcastle University, Newcastle Upon Tyne, UK \\

Boguslaw Obara\textsuperscript{1,2,4} \\

Editors: Under Review for MIDL 2021 \\

Keywords: Machine learning, image processing and computer vision, medicine, segmentation, neural nets

Abstract

Macular holes are a common eye condition which result in visual impairment. We look at the application of deep convolutional neural networks to the problem of macular hole segmentation. We use the 3D U-Net architecture as a basis and experiment with a number of design variants. Manually annotating and measuring macular holes is time consuming and error prone. Previous automated approaches to macular hole segmentation take minutes to segment a single 3D scan. Our proposed model generates significantly more accurate segmentations in less than a second. We found that an approach of architectural simplification, by greatly simplifying the network capacity and depth, exceeds both expert performance and state-of-the-art models such as residual 3D U-Nets.

1. Introduction

Idiopathic full thickness macular holes (iFTMH) are a common, and visually disabling condition, being bilateral in 10\% of affected individuals. They occur at a prevalence of approximately 1 in 200 of the over 60-year-old population with an incidence of approximately 4000 per annum in the United Kingdom (UK)\cite{Ali2017,McCannel2009}. If left untreated they result in visual acuity below the definition of blindness and typically less than 6/60.

3D high-resolution images of the retina can be created using optical coherence tomography (OCT) \cite{Hee1995}. It is now the standard tool for diagnosing macular holes \cite{Goldberg2014}. Compared to previous imaging methods, OCT can more easily assist a clinician in differentiating a full-thickness macular hole from mimicking pathology which is important in defining appropriate treatment \cite{Hee1995}.

An OCT scan of a macular hole is a 3D volume. Clinicians typically view OCT images as a series of 2D images, choose the central slice with the maximum dimensions and perform
Figure 1: Small 3D U-Net ($M_1$). The proposed model is a cut-down version of 3D U-Net, as proposed by Çiçek et al. (Çiçek et al., 2016). It has fewer levels and a carefully optimized capacity for our datasets. More advanced architectures, including residual 3D U-Nets (De Fauw et al., 2018) were found to negatively impact performance.

measurements which are predictors of anatomical and visual success such as base diameter, macular hole inner opening and minimum linear diameter (Madi et al., 2016)(Chen et al., 2020)(Murphy et al., 2020)(Steel et al., 2020). This approach is limited as it assumes that the macular hole base is circular, and would give incorrect results when it is elliptical (Nasrulloh et al., 2018), which is typically the case (Chen et al., 2020). With the advent of automated 3D approaches, it is possible to begin to look at measurements in 3D and how they might be predictors of anatomical and visual success.

Neural networks are an interconnected group of artificial neurons, which can be reconfigured to solve a problem based on data. Convolutional neural networks (CNN) are a type of neural network inspired by how the brain processes visual information (Lindsay, 2020). CNNs have been very successful in computer vision problems, such as automating the segmentation of disease in medical images. For a CNN to be able to segment images, it needs to have access to images with associated ground truth (GT) information which highlight the important areas of the image for the task at hand. This has to be done manually and is time consuming and usually requires expert knowledge.

The U-Net CNN architecture (Ronneberger et al., 2015) is a highly-utilized CNN architecture for biomedical image segmentation. It has had success in segmenting other eye diseases such as macular edema, even when dataset sizes are limited (Frawley et al., 2020). The no free lunch theorem states that there is no single model that performs optimally across all task distributions, whereas Occam’s razor indicates we should choose the simplest
or smoothest architecture where available. This motivates us to re-examine simplifications of U-Net without overparameterizations or the latest residual designs (De Fauw et al., 2018). Our proposed model is a smaller version of the model from the original 3D U-Net paper. For comparison, we also implemented and evaluated the proposed model with residual blocks added, similar to those described by He et al. (He et al., 2016). In addition, we implemented DeepMind’s OCT segmentation model (De Fauw et al., 2018) and ran the same tests with it.

We use that the above techniques to develop an automated approach to macular hole segmentation which yields significantly improved results compared to the current state of the art. We present a thorough comparative analysis of the above-mentioned models against the current state-of-the-art automated approach (Nasrulloh et al., 2018) along with a comparison against expert performance.

2. Method

Image segmentation involves the labelling of objects of interest in an image. For a 3D image, this is done by assigning voxels with shared characteristics to corresponding class labels. We wish to assign areas of the macular hole volume in an OCT image to white voxels and all other regions to black voxels.

We use binary cross-entropy as our loss function, which tells us the probability of a voxel being part of a macular hole region:

$$L_{BCE} = -\sum_{i=1}^{n} p_i \log q_i$$

$$= -(p_i \log q_i + (1-p_i) \log(1-q_i)),$$

where $p_i$ is the ground truth and $q_i$ is the output of our model. For images with multiple annotations in our training set, we trust them with equal integrity and the target probabilities are averaged. The validation set has no samples with multiple annotations and for the unseen test set we use a voting approach as described in Section 3.

U-Net takes as input a 2D image and outputs a set of probabilities. Each entry in the output is the probability of each part of the image being a part of the segmented region. It is a U-shaped CNN architecture, consisting of a contracting path and an expansive path. The contracting path consists of 2D convolutions, ReLU activations and 2D max pooling at each level. The expansive path’s levels use skip connections to their contracting path equivalent, along with up convolutions and ReLU activations. Skip connections allow for high-resolution information to be captured by the model while the contracting/expansive paths capture the abstract shape of the segmentation. The 3D U-Net architecture (Çiçek et al., 2016) is a version of U-Net designed for use with 3D images which uses 3D convolutions, up convolutions and max pooling layers. This allows for improved segmentation of 3D images as the context from multiple slices are used to decide whether an individual voxel is an object or not.

A number of models based on the 3D U-Net architecture were compared:

$M_1$: Small 3D U-Net (Proposal)
$M_2$: Small residual 3D U-Net (Residual)

$M_3$: Residual 3D U-Net for 2D slices (DeepMind) (De Fauw et al., 2018)

A diagram of model $M_1$ is shown in Figure 1. Our experiments showed that using three levels for this model resulted in the best performance. A scaled-down input image of $160 \times 188 \times 49$ yielded the best results for models $M_1$ and $M_2$. The output is of a similar dimension to the input. $M_2$ is similar to $M_1$ except that residual blocks have been added to each level. $M_3$ is a residual 3D U-Net architecture which takes nine slices of the OCT image as input and outputs a 2D probability map as output, representing the segmentation of a single slice of the OCT image. For $M_3$, the slice which we want to segment, along with 4 slices on either side is input to the model, which is a $321 \times 376 \times 9$ image. For slices near the boundaries, we use mirroring of slices. It outputs a set of $321 \times 376$ probabilities, corresponding to one slice of the 3D OCT. $M_3$, therefore, requires 49 iterations to segment a whole 3D OCT image in our dataset. Model $M_3$ has the highest number of parameters of the models, with $M_1$ having the fewest parameters.

Jaccard index was used as the primary metric for measuring the performance of each method. This is a standard measure of the performance of image segmentation methods, especially in medical image segmentation (Taha and Hanbury, 2015). It measures the overlap of two segmentations: the intersection of the segmentations divided by their union. Values close to 0 indicate little overlap and values close to 1 indicate significant overlap.

3. Implementation

Our experiments were all conducted using the PyTorch (Paszke et al., 2019) deep learning framework on NVIDIA Turing GPUs with 24GB of memory. We trained each model for 500 epochs where each epoch ran over 10 3D images. This means that the models which output a 3D segmentation ($M_1$ and $M_2$) had 10 iterations per epoch, and the slice-based model ($M_3$) had 490 iterations per epoch. As source code was not released for DeepMind’s model, $M_3$ is implemented as closely as possible from the description provided in the original paper and slightly adapted to fit the binary classification problem.

In order to evaluate models $M_1$ and $M_2$, we scale up the output probability map to its original size using trilinear interpolation and threshold it at 0.5 to generate a binary segmentation. For model $M_3$, we individually run over all 49 slices of an image and recombine the 49 2D probability maps into a single 3D probability map. We then threshold this combined map at 0.5 to generate a 3D binary segmentation. The Adam optimization algorithm was used to optimize parameters of the models (Kingma and Ba, 2015), with hyperparameters being found by experimentation. For model $M_1$, a learning rate of $1 \times 10^{-4}$ and weight decay of $1 \times 10^{-6}$ was used. For model $M_2$, a learning rate of $1 \times 10^{-4}$ and weight decay of $1 \times 10^{-5}$ was used. For model $M_3$, a learning rate of $7.5 \times 10^{-5}$ was used and weight decay was disabled. Images were normalized to the $[0, 1]$ range before being input to the models.

Each model was trained and evaluated separately three times to assess the consistency of our results. Due to the fact that we only had a small number of images with multiple authors, we decided to keep the training, validation and unseen test set static for all tests rather than using k-fold cross validation. We reserved all images which had three annotations for the unseen test set, in order for us to compare our results with expert disagreement: expert 1 with expert 2, expert 2 with expert 3 and expert 1 with expert 3.
Robust 3D U-Net Segmentation of Macular Holes

Figure 2: Qualitative sample output of our trained macular hole model ($M_1$) compared with the ground truth, the state-of-the-art automated approach (Nasrulloh) (Nasrulloh et al., 2018), the residual model ($M_2$) and DeepMind’s model ($M_3$). For clarity, we zoomed in on the predicted regions.

4. Materials

OCT images were exported from a Heidelberg SPECTRALIS (Heidelberg, Germany) machine. A high density central horizontal scanning protocol with 29-30 micron line spacing was used in the central 15 by 5 degrees. The individual OCT line scans had spacing which varied slightly between images but it was typically 5.47 microns per pixel on the $x$-axis and 3.87 microns per pixel on the $y$-axis. All scans used a 16 automatic real-time setting enabling multisampling and noise reduction over 16 images.

All images were cropped to the same size and unnecessary information such as the fundus image were removed. Annotations were created by a mixture of clinicians and image experts using a 3D image annotation tool. Pixels on each slice of the OCT scan which represented macular hole were highlighted. There were 85 (image, annotation) pairs in the training dataset, 56 after combining annotations from multiple authors. There were 22 pairs in the validation dataset and 9 in the unseen test set.

Originally we had three annotations for each OCT image in the unseen test set. However, due to inconsistencies between authors, we combined all ground truths into a single ground truth per image. To do this, we used a voting system, where if $\frac{2}{3}$ of the authors had annotated a voxel, that voxel was annotated in the resultant ground truth. All images and ground truths at full size had dimensions $321 \times 376 \times 49$.

5. Results

In this section, we evaluate our results both qualitatively and quantitatively.
Table 1: 3D and 2D segmentation output of model $M_1$ (Proposal) on unseen test images along with ground truth.

| 3D segmentation | 2D segmentation | 2D ground truth |
|-----------------|----------------|----------------|
| ![3D segmentation](image1) | ![2D segmentation](image2) | ![2D ground truth](image3) |
| ![3D segmentation](image4) | ![2D segmentation](image5) | ![2D ground truth](image6) |
| ![3D segmentation](image7) | ![2D segmentation](image8) | ![2D ground truth](image9) |
| ![3D segmentation](image10) | ![2D segmentation](image11) | ![2D ground truth](image12) |

5.1. Qualitative Results

The qualitative results of running the trained macular hole model are generally quite close to the ground truth, as seen in Figure 2. In general, predictions from model $M_1$ are qualitatively better than the other models and the state-of-the-art automated approach. Table 1 shows the output of our proposed model in 3D.

5.2. Quantitative Results

Figure 3 shows how the Jaccard index improves as $M_1$ is trained and we can see that after 200 epochs it surpasses the performance of the state of the art and expert disagreement. The trained macular hole models perform very well compared to the state-of-the-art...
Table 2: Jaccard index comparison between the state of the art (Nasrulloh et al., 2018) and tested models. $M_1$ and $M_2$ correspond to our cut-down 3D U-Net models, without and with residual blocks respectively. $M_3$ is DeepMind’s model. Each entry shows the mean and standard deviation over three runs.

| Image | Nasrulloh et al. | $M_1$ (Proposal) | $M_2$ (Residual) | $M_3$ (DeepMind) |
|-------|-----------------|-----------------|-----------------|-----------------|
| Image 1 | 0.714 | 0.865 ± 0.009 | 0.868 ± 0.002 | 0.832 ± 0.006 |
| Image 2 | 0.743 | 0.891 ± 0.02 | 0.887 ± 0.014 | 0.893 ± 0.012 |
| Image 3 | 0.772 | 0.887 ± 0.004 | 0.885 ± 0.002 | 0.872 ± 0.006 |
| Image 4 | 0.811 | 0.895 ± 0.012 | 0.884 ± 0.001 | 0.894 ± 0.006 |
| Image 5 | 0.787 | 0.894 ± 0.005 | 0.901 ± 0.003 | 0.875 ± 0.014 |
| Image 6 | 0.678 | 0.804 ± 0.008 | 0.815 ± 0.007 | 0.765 ± 0.006 |
| Image 7 | 0.845 | 0.907 ± 0.002 | 0.905 ± 0.004 | 0.893 ± 0.009 |
| Image 8 | 0.874 | 0.874 ± 0.012 | 0.862 ± 0.002 | 0.893 ± 0.006 |
| Image 9 | 0.787 | 0.869 ± 0.019 | 0.853 ± 0.008 | 0.835 ± 0.007 |
| Mean | 0.779 | **0.876 ± 0.012** | 0.874 ± 0.006 | 0.861 ± 0.008 |

Table 3: Detailed statistics of model $M_1$ (Proposal). Each entry shows the mean and standard deviation over three runs.

| Image | Precision | Recall | DSC | AVD | AP |
|-------|-----------|--------|-----|-----|----|
| Image 1 | 0.93 ± 0.009 | 0.926 ± 0.012 | 0.928 ± 0.005 | 1352 ± 908.357 | 0.862 ± 0.01 |
| Image 2 | 0.954 ± 0.008 | 0.931 ± 0.014 | 0.942 ± 0.011 | 1379 ± 369.396 | 0.889 ± 0.021 |
| Image 3 | 0.949 ± 0.003 | 0.931 ± 0.003 | 0.94 ± 0.002 | 2308 ± 517.533 | 0.885 ± 0.004 |
| Image 4 | 0.974 ± 0.005 | 0.917 ± 0.016 | 0.945 ± 0.007 | 5293 ± 1817.849 | 0.895 ± 0.011 |
| Image 5 | 0.915 ± 0.012 | 0.974 ± 0.007 | 0.944 ± 0.003 | 2320 ± 763.08 | 0.892 ± 0.005 |
| Image 6 | 0.848 ± 0.003 | 0.94 ± 0.007 | 0.891 ± 0.005 | 1911 ± 80.168 | 0.797 ± 0.009 |
| Image 7 | 0.965 ± 0.002 | 0.938 ± 0.001 | 0.951 ± 0.001 | 1564 ± 157.11 | 0.906 ± 0.002 |
| Image 8 | 0.898 ± 0.006 | 0.971 ± 0.008 | 0.933 ± 0.007 | 4544 ± 155.656 | 0.872 ± 0.013 |
| Image 9 | 0.917 ± 0.016 | 0.943 ± 0.012 | 0.93 ± 0.011 | 1186 ± 906.832 | 0.865 ± 0.02 |
Figure 3: Average Jaccard index of our proposed model ($M_1$) over 3 runs on the unseen test set as the model was trained (higher is better). We see that the model gives significantly better results than the state-of-the-art automated approach (Nasrulloh et al., 2018) and expert agreement.

approach (Nasrulloh et al., 2018) as we see in Table 2. Despite $M_1$ having the fewest parameters, it achieves the best performance of all of the models. We can see that all of the models achieved better results than the state-of-the-art method on our unseen test set. The state-of-the-art method is a level set approach which does not use deep learning. Further results in Table 3 show that $M_1$ performs consistently well under other standard segmentation quality measures.

6. Conclusion

All of the models tested exceeded the performance of the state-of-the-art automated approach which is a level set method. It is clear that deep learning methods allow for the generation of segmentations which are closer to what humans provide. Of the models tested, model $M_1$ gives the best qualitative and quantitative results and surpasses the performance of experts. It is also a quick model to run, requiring only one pass through the whole 3D image, whereas $M_3$ requires one pass per slice. For these reasons, $M_1$ is the best candidate to form the basis of future studies in a clinical setting. These findings show an interesting case of the no free lunch theorem and Occam’s razor; careful tuning and in some cases architectural simplification can, for some simple task distributions, be more effective than very deep residual designs.
Acknowledgments

In accordance with his ethical obligation as a researcher, Jonathan Frawley reports that he received funding for his PhD from Intogral Ltd. Some of the work described was developed as part of his work as an employee at Intogral Ltd. Intogral Ltd also provided annotations created by non-clinicians. Data and annotations by the clinician for this project were kindly provided by Maged Habib, Caspar Geenen and David H. Steel of the Sunderland Eye Infirmary, South Tyneside and Sunderland NHS Foundation Trust, UK. All images were collected as part of routine care and anonymised.

References

Ferhina S Ali, Joshua D Stein, Taylor S Blachley, Sarah Ackley, and Jay M Stewart. Incidence of and risk factors for developing idiopathic macular hole among a diverse group of patients throughout the United States. *JAMA Ophthalmology*, 135(4):299–305, 2017.

Yunzi Chen, Amar V Nasrulloh, Ian Wilson, Caspar Geenen, Maged Habib, Boguslaw Obara, and David HW Steel. Macular hole morphology and measurement using an automated three-dimensional image segmentation algorithm. *BMJ Open Ophthalmology*, 5(1):e000404, 2020.

Özgür Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432, 2016.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342, 2018.

J. Frawley, C. G. Willcocks, M. Habib, C. Geenen, D. H. Steel, and B. Obara. Segmentation of macular edema datasets with small residual 3d u-net architectures. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 582–587, 2020. doi: 10.1109/BIBE50027.2020.00100.

Roger A Goldberg, Nadia K Waheed, and Jay S Duker. Optical coherence tomography in the preoperative and postoperative management of macular hole and epiretinal membrane. *British Journal of Ophthalmology*, 98(Suppl 2):ii20–ii23, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference On Computer Vision*, pages 630–645, 2016.

Michael R Hee, Carmen A Puliafito, Carlton Wong, Jay S Duker, Elias Reichel, Joel S Schuman, Eric A Swanson, and James G Fujimoto. Optical coherence tomography of macular holes. *Ophthalmology*, 102(5):748–756, 1995.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations*, 2015.
Grace W. Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, page 1–15, Feb 2020. ISSN 1530-8898. doi: 10.1162/jocn_a_01544. URL http://dx.doi.org/10.1162/jocn_a_01544.

Haifa A Madi, Ibrahim Masri, and David H Steel. Optimal management of idiopathic macular holes. *Clinical Ophthalmology (Auckland, NZ)*, 10:97, 2016.

Colin A McCannel, Jennifer L Ensminger, Nancy N Diehl, and David N Hodge. Population-based incidence of macular holes. *Ophthalmology*, 116(7):1366–1369, 2009.

Declan C. Murphy, Amar V. Nasrulloh, Clare Lendrem, Sara Graziado, Mark Alberti, Morten la Cour, Boguslaw Obara, and David H.W. Steel. Predicting postoperative vision for macular hole with automated image analysis. *Ophthalmology Retina*, 2020. ISSN 2468-6530. doi: https://doi.org/10.1016/j.oret.2020.06.005. URL http://www.sciencedirect.com/science/article/pii/S2468653020302311.

Amar Nasrulloh, Chris Willcocks, Philip T.G. Jackson, Caspar Geenen, Maged S. Habib, David H.W. Steel, and Boguslaw Obara. Multi-scale segmentation and surface fitting for measuring 3D macular holes. *IEEE Transactions on Medical Imaging*, 37(2):580–589, 2018.

Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pages 234–241, 2015.

DH Steel, PHJ Donachie, GW Aylward, DA Laidlaw, TH Williamson, and D Yorston. Factors affecting anatomical and visual outcome after macular hole surgery: findings from a large prospective UK cohort. *Eye*, pages 1–10, 2020.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015.