PLOS ONE

RESEARCH ARTICLE

Automated summarisation of SDOCT volumes using deep learning: Transfer learning vs *de novo* trained networks

Bhavna Josephine Antony*, Stefan Maetschke, Rahil Garnavi

IBM Research Australia, 22/60 City Road, IBM Center, Southbank, VIC 3006, Australia

* bhavna.antony@au1.ibm.com

Citation: Antony BJ, Maetschke S, Garnavi R (2019) Automated summarisation of SDOCT volumes using deep learning: Transfer learning vs *de novo* trained networks. PLoS ONE 14(5): e0203726. https://doi.org/10.1371/journal.pone.0203726

Editor: Jianjun Hu, University of South Carolina, UNITED STATES

Received: August 16, 2018

Accepted: April 12, 2019

Published: May 13, 2019

Data Availability Statement: The data is part of a dataset that was made available at http://people.duke.edu/~sf59/RPEDC_Ophth_2013_dataset.htm by Dr. Sina Farsiu. Those interested can access the data in the same manner as the authors. The authors had no special access privileges.

Funding: The funder (IBM Research Australia) provided support in the form of salaries for authors (BJA, SM, RG), but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the

Abstract

Spectral-domain optical coherence tomography (SDOCT) is a non-invasive imaging modality that generates high-resolution volumetric images. This modality finds widespread usage in ophthalmology for the diagnosis and management of various ocular conditions. The volumes generated can contain 200 or more B-scans. Manual inspection of such large quantity of scans is time consuming and error prone in most clinical settings. Here, we present a method for the generation of visual summaries of SDOCT volumes, wherein a small set of B-scans that highlight the most clinically relevant features in a volume are extracted. The method was trained and evaluated on data acquired from age-related macular degeneration patients, and "relevance" was defined as the presence of visibly discernible structural abnormalities. The summarisation system consists of a detection module, where relevant B-scans are extracted from the volume, and a set of rules that determines which B-scans are included in the visual summary. Two deep learning approaches are presented and compared for the classification of B-scans—transfer learning and *de novo* learning. Both approaches performed comparably with AUCs of 0.97 and 0.96, respectively, obtained on an independent test set. The *de novo* network, however, was 98% smaller than the transfer learning approach, and had a run-time that was also significantly shorter.

1 Introduction

The detection of key frames is a common approach employed in video analysis, particularly for the summarisation of video sequences. The techniques typically rely on the detection of explicit features of interest such as motion [1, 2] as well as other features such as edge information [3] and self-similarity [4]. Condensing videos via shot boundary detection has also been applied to the summarisation of video sequences [5].

In medical imaging, the detection of keyframes is more commonly found in the analysis of angiogram video sequences, but is less common in other radiographic modalities. Gibson *et al.* [6] described an approach for the compression of angiogram videos by detecting diagnostically relevant frames in videos and ensuring they were preserved. Syeda-Mahmood *et al.* [7] presented an approach for the detection of key frames in angiogram video analysis by detecting

manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: Although IBM is a commercial entity, this does not alter our adherence to PLOS ONE policies on sharing data and materials.

the vessels and selecting frames in which their visibility was best. However, both of these approaches relied on the explicit detection of features of interest in the images in order to identify them as "key".

In ophthalmology, spectral-domain optical coherence tomography (SDOCT) [8] has begun to find widespread use for the diagnosis and management of various ocular conditions. This non-invasive imaging modality relies on laser interferometry to generate high-resolution images of the retina, which allows for the visualisation and quantification of structures in 3-D. These volumetric images are comprised of B-scans, which numbers can range from as few as five to two hundred or more. Summarisation of these volumes in current scanning systems is usually limited to a report that indicates the thicknesses of retinal layers. While such a report shows large pathologies such as choroidal neovascularizations (CNV), smaller abnormal indicators such as drusen, epiretinal membranes and microcytic macular edema would not be visible. Thus, visual summaries could complement the existing approach, by highlighting the pathological conditions that are currently not quantified. Previously, Chakravarthy *et. al* [9] described an approach for the detection of B-scans that show choroidal neovascularization (CNV). The method relies on the detection of the retina, followed by a machine-learning approach for the detection of possible fluid patches in the images. While this approach does in fact extract the specific B-scans, the method is limited to CNVs associated with wet-AMD.

Here, we present a deep learning approach for the automated summarisation of SDOCT volumes. Similar to previous summarisation techniques our proposed system begins with the detection of "key" B-scans. The system was trained and tested on SDOCT volumes acquired from patients that presented with age-related macular degeneration (AMD), and "relevance" was defined on the basis of the presence of visibly discernible structural abnormalities. Using a deep learning approach for this task allows it to be posed as a recognition task, and thus, does not require the explicit extraction of features (such as CNV) in order to characterise the B-scan. We employed and compared two deep learning techniques for keyframe extraction, where one is a transfer learning technique based on a pretrained network, while the second is a *de novo* trained custom convolutional neural network (CNN) that is significantly smaller. Transfer learning is a commonly used technique that allows for the repurposing of pretrained networks in applications where data might be scarce (as is commonly the case in medical imaging). Once the relevant B-scans had been identified, a set of rules were applied to generate the visual summary.

The paper is organised as follows: Section 2 details the data used in this experiment; Section!3 describes the two deep learning networks as well as the summarisation rules. The evaluation and comparison of the two networks is presented in Section!4, and a final discussion of the results can be found in Section 5.

2 Data

The data used in the experiments were SDOCT images acquired as part of the AREDS2 Ancilliary Study. As detailed in [10], the dataset was registered at ClinicalTrials.gov (Identifier: NCT00734487) and approved by the the institutional review boards at Devers Eye Institute, Duke Eye Center, Emory Eye Center, and National Eye Institute. With adherence to the tenets of the Declaration of Helsinki, informed consent was obtained from all subjects.

The study cohort consisted of 115 healthy individuals and 269 patients with age-related macular degeneration (AMD). The images were acquired on a Bioptigen SDOCT scanner (Leica Microsystems Inc., Illinois) from an approximately 6.7×6.7mm area centred on the fovea. Each volumes consisted of 100 B-scans, each containing 1000 A-scans and 512 pixels per A-scan (see Fig 1(a)). Further details of the study are provided in [10].

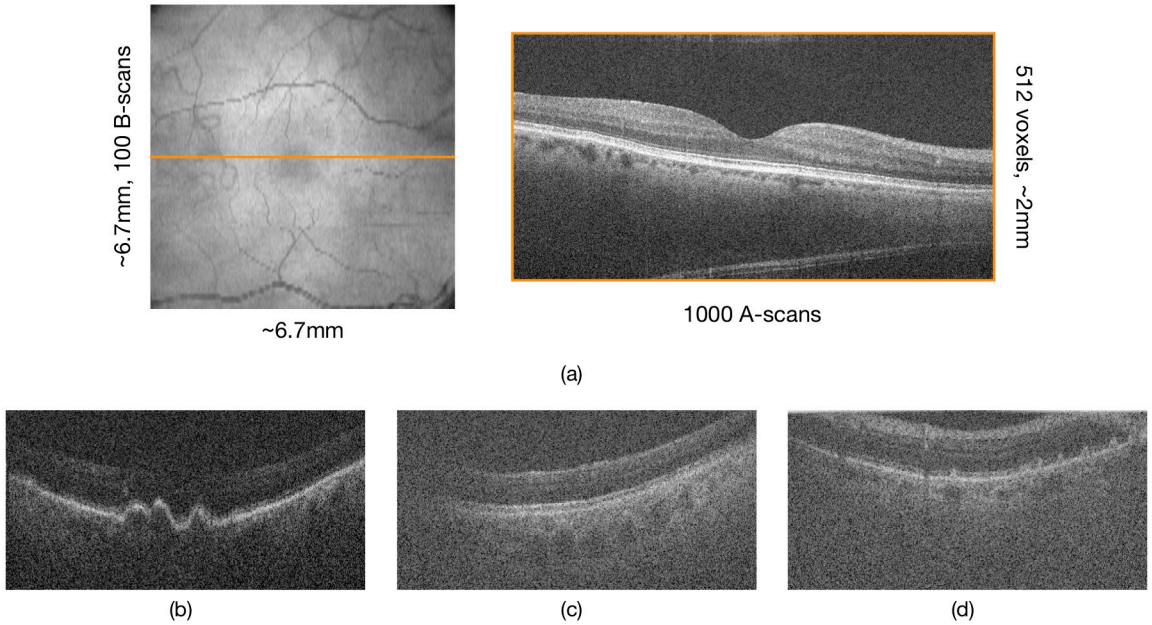

Fig 1. (a) The dimensions of an SDOCT volume depicted on the en-face image (left) and the central B-scan of the volume (right). Examples of poor quality scans with (b) poor contrast, (c) large shadows, and (d) incorrect mirror position.

https://doi.org/10.1371/journal.pone.0203726.g001

2.1 Data annotation

The volumes were manually annotated and labeled as being healthy, relevant (containing visibly discernible structural change) or low-quality. For this, each B-scan was visualised and labeled relevant if any visual structural change was observed. Thus, retinal layer thinning (which is difficult to identify visually) was not considered a key feature. However, B-scans with even minor disruptions like small drusen, reticular pseudodrusen and epiretinal membranes were all labeled as relevant "key" B-scans.

The presence of large shadows, poor contrast and other artefacts (such as vinetting, mirror location errors) were flagged as poor quality B-scans (see Fig 1(b)–1(d)).

3 Methods

Deep learning [11] has been successfully employed for a number of applications in computer vision such as image recognition [12, 13] and semantic segmentation [14, 15]. This technique has also found application in medical imaging [16] for recognition [17], segmentation [18–20] as well as image registration [21–23]. While larger architectures have shown to perform better than shallower networks, their training also requires larger datasets.

Transfer learning is a technique that re-purposes existing, trained models for new tasks by retraining only small parts of the network. As most weights of the network are left unchanged, this reduces the amount of training data required. Transfer learning lends itself to medical imaging quite well, as large datasets are difficult to acquire in the medical domain. Thus, this was the first technique we employed for the detection of relevant B-scans (detailed in Section 3.1). We utilised the 16-layer VGG network [13] that was initially trained for the ImageNet Challenge—a classification problem consisting of 1000 classes of natural scene images [24].

Transfer learning, however, is not without problems. The pre-trained networks were designed for object recognition in natural scene images, and require the input to be a 3-channel RGB image. SDOCT images are however, grayscale. Thus, a B-scan either has to be

replicated three times to meet the required input dimensions, or a section of three slices (and only three) has to be used as network input. Designing and training a network *de novo* allows to circumvent this requirement and potentially could also result in a smaller or more accurate network. For comparison we therefore also designed a significantly smaller convolutional neural network (CNN) that was trained *de novo* (detailed in Section 3.2).

The networks were developed in Keras [25] with TensorFlow [26] as the backend and nutsflow/ml [27] for the data pre-processing.

3.1 Transfer learning approach

The structure of the 16-layer VGG network [13] is as shown in Fig 2(a). It consists of 3x3 convolutional filters with a stride of 1; padded to preserve spatial resolution. All layers utilised rectified linear unit [12] (ReLU) activation. The original network for an input RGB (3-channel) image of 224x224 pixels in size contained 138 million parameters. See [13] regarding training of the original network.

Since the classification task at hand is a 2-class problem, the model was changed to reflect this (see Fig 2(b)). Furthermore, the two fully connected layers prior to the final layer were reduced in size from 4096 to 1024 and 512, respectively. Removal of the two original fully connected layers allowed for the input size of the network to be changed to 300x512x3 pixels. For an input image of this size, the total size of the network is 18.6 million parameters. The five blocks of convolutional filters were used as feature extractors and were not re-trained or fine-tuned, resulting in a network with 3.8 million trainable parameters.

Network weights were opimized by Adam [28], with parameters set to recommended values (learning rate set to 1^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1^{-8}$). The loss function was the balanced cross-entropy loss function:

$$\mathcal{L} = \sum_i - C_2 y_i \log \hat{y}_i - C_1 (1 - y_i) \log (1 - \hat{y}_i) \quad (1)$$

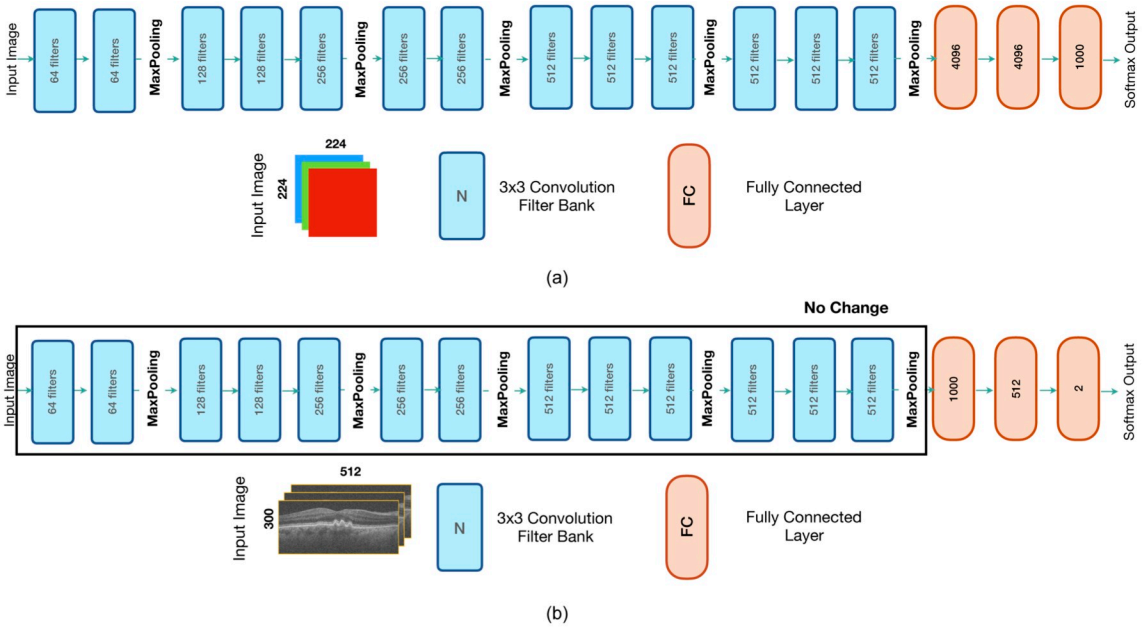

Fig 2. The structure of the (a) original VGG-16 network [13], and (b) the modified transfer network for keyframe detection. Note that the layers in the first five banks of CNNs have not been changed from the original network. First, the last three fully-connected layers were re-trained, followed by a fine-tuning step where all layers were retrained.

https://doi.org/10.1371/journal.pone.0203726.g002

where, y_i is the true label and $\hat{y}_i$ is the predicated label of the i-th sample, and C_1 and C_2 are the number of samples of the first and second class in the batch, respectively. This loss function, being normalised by the number of samples in each class helps with class imbalances. Training was stopped when the validation loss did not decrease by more than 0.1 or after 150 epochs. A second stage of fine-tuning was conducted where the all layers of the network were retrained for an additional 50 epochs.

Data preprocessing. The individual B-scans in each volume are 512x1024 pixels in size. The retina however, does not encompass the entire B-scan, with a large portion of the image showing the vitreous, choroid and scleral tissue. Thus, detecting and extracting the image region that contains the retina reduces the image size. Therefore, the image was filtered with a gaussian derivative filter (first order, $\sigma = 6.0$). This generated a high response at the internal limiting membrane (ILM), and the ellipsoid zone of the photoreceptors as shown in Fig 3(a) and 3(b). This response was then thresholded (using a threshold obtained by Otsu's method [29]), and the largest connected components were detected. Since the largest two components belong to the retinal surfaces, their locations defined the bounding boxes (at least 300 pixels in height) around the retina, and B-scan were cropped to this size. The cropped B-scan was finally resized to 300x512, and replicated three times to match the requirements of the VGG-16 network, which expects the input to be a 3-channel image. The training data was augmented through random rotations (±5˚), translations (±10 pixels), contrast scaling (0.3, 1.7) and flipping along the horizontal axis.

3.2 *De novo network*

The *de novo* CNN consisted of 7 convolutional layers, using 3x3 filters in size (see Fig 4). Skip connections [30] were introduced between the layers, with the outputs from the previous layers being concatenated prior to pooling. ReLU activation, and pooling (maximum in a 2x2 window) followed each CNN layer. A global average pooling (GAP) layer was added to enable

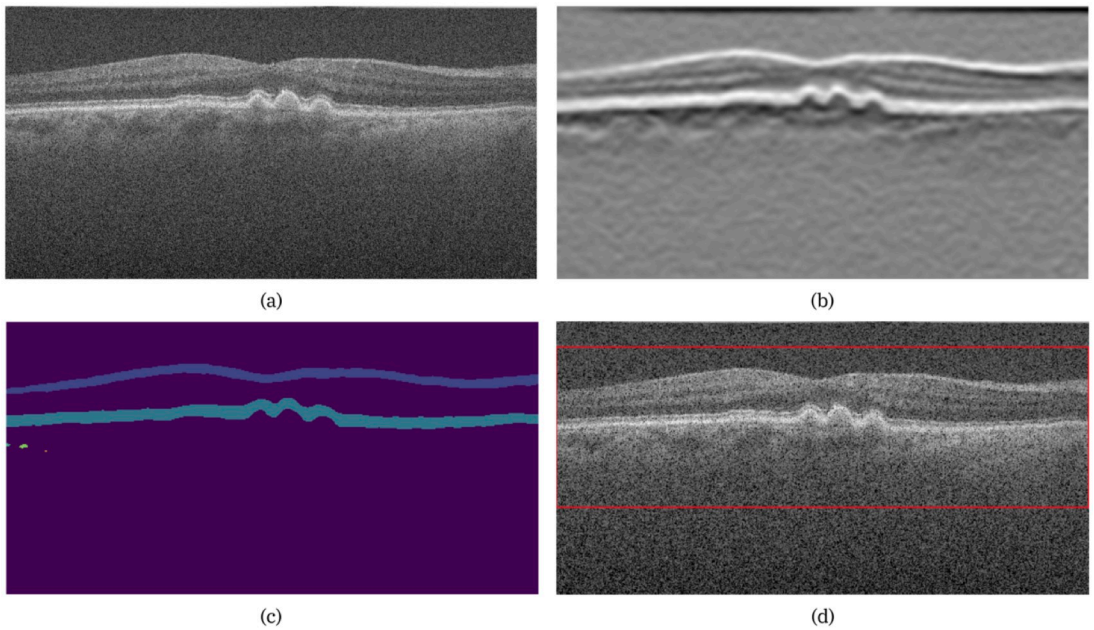

Fig 3. Outline of the data preprocessing steps showing (a) the original slice, (b) the filtered image, (c) connected components of the thresholded image, and (d) the final cropped image indicated by the red box.

https://doi.org/10.1371/journal.pone.0203726.g003

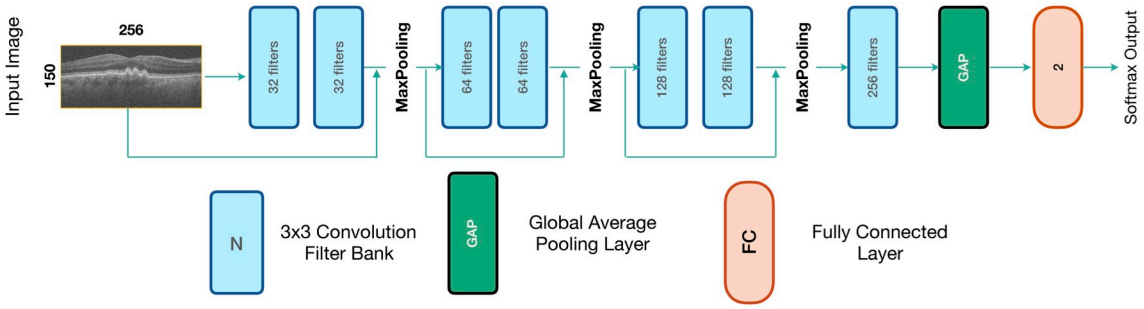

Fig 4. Architecture of the *de novo* network for keyframe detection.

https://doi.org/10.1371/journal.pone.0203726.g004

the generation of class activation maps (CAM) [31]. Finally, a fully connected layer (size = 2) with a softmax output provided the class probability for the input B-scan. The resulting network contained only 1.4 million trainable parameters, and is 8% the size of the transfer learning network. As before, training was performed by Adam with the balanced cross-entropy loss function (see Eq 1).

Data preprocessing. The individual B-scans were processed as described for the transfer-learning approach, beginning with the detection of the retina followed by image cropping at the bounding box. The resulting images were down-sampled by a factor of 2 (final size 150x256 pixels) and directly inputted into the network as 1-channel gray-scale images. Training data was augmented as before, employing random rotations (± 5˚), translations (± 10 pixels), contrast scaling (0.3, 1.7) and flipping along the horizontal axis.

3.3 Experimental setup

Of the total 38,382 available B-scans, 32,492 were found to be of sufficient quality for analysis. A 10-fold cross validation experiment was used to assess the performance of both networks. The finale set of annotated B-scans divided into training, validation and testing sets containing 75%, 10% and 15% of the SDOCT volumes, respectively. The allocation of an entire volume to a set ensured that B-scans from a single volume were not distributed across the sets. The division of the data into 3 sets ensures that the test set in each cross validation split consists of volumes that have not been previously encountered by the network during training or validation. The training termination criteria is set using the validation set; when no change is detected in the area under the curve (AUC) in the validation set, the training is terminated.

The performance of the two networks was evaluated using AUC, which was statistically evaluated using a Wilcoxon rank sum test. The false positive and false negative rates were also computed and compared.

3.4 Summarisation rules

Once the relevant B-scans have been extracted, a set of rules is imposed to select the keyframes. If no relevant B-scans were identified by the deep learning framework, then three slices (two peripheral and one central slice) are returned by the system. Otherwise, the set of relevant B-scans F_i, $i = 1, 2, \ldots N$ are grouped into regions. If two of the identified B-scans are only separated by a small distance (preset threshold T), they are considered to be part of the same region. This not only compensates for mislabeled B-scans (not correctly identified as relevant for the summary), but also aggregates small regions that show disease-induced change. For instance, drusen may be present in a small number of B-scans near the fovea, but the individual slices may be separated by a few slices that show no pathology. In such a situation, it is

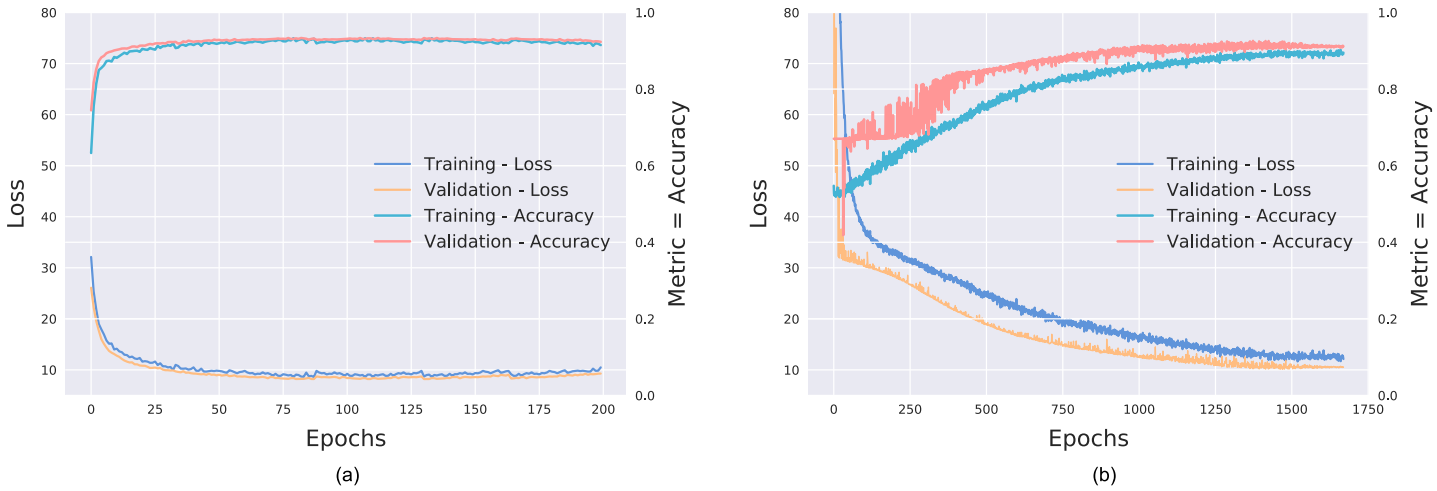

Fig 5. Loss and accuracy (AUC) monitored during (a) transfer learning and (b) *de novo* training.

https://doi.org/10.1371/journal.pone.0203726.g005

reasonable to aggregate them into a larger region. A flexible threshold T, controls the aggregation during run time.

Next, each region is represented by the first, median and last scan of each region. Thus, if M regions are detected in the volume, a total of $3M$ key-frames will be returned by the algorithm. Note, that selecting the median and not the midpoint between the first and last B-scans, ensures that the B-scan included in the summary will be one that was identified by the deep learning classifier as being relevant for the visual summary.

4 Results

The training of the transfer learning network (including the fine-tuning) required approximately 200 epochs, while *de novo* training needed nearly 2000 epochs. This is to be expected as the transfer learning network is pre-trained. The fine-turning however, did not improve the performance of the network. The initial loss was also substantially larger for the *de novo* network. Fig 5 shows the training process of the transfer learning and *de novo* network for a single split.

The mean AUC (standard deviation expressed in parentheses) computed on the test set was 0.967 (0.004) and 0.965 (0.011) ($p = 0.51$) for the transfer learning and the *de novo* networks, respectively. The standard deviations seen in the 10-fold cross-validation experiments is quite small, indicating the stability of both networks.

The confusion matrices and AUC curves for a single split are as shown Fig 6. The sensitivity (lower right quadrant of the confusion matrix) was found to be 0.91 for the transfer learning and 0.90 for the *de novo* network (see Fig 6(a) and 6(b)). Similarly, the specificity was found to be 0.87 and 0.89 for the transfer learning and *de novo* networks, respectively. Fig 7 shows instances of B-scans with mild ((a)-(c)) and severe ((d)-(f)) AMD-related pathologies that were correctly identified by the networks as being relevant to the visual summary. The CAM visualisation for mild and severe B-scans are shown in Fig 7.

In the control datasets, 3.5% and 5.3% ($p < 0.01$), of the B-scans were erroneously detected as abnormal by the TL and *de novo* networks, respectively. There was no statistical difference between the false positive rates detected by the two networks. Visualisations of these B-scans showed that poor quality scans were sometimes misclassified. Errors in the retinal localisation as shown in Fig 8(b) also led to misclassifications.

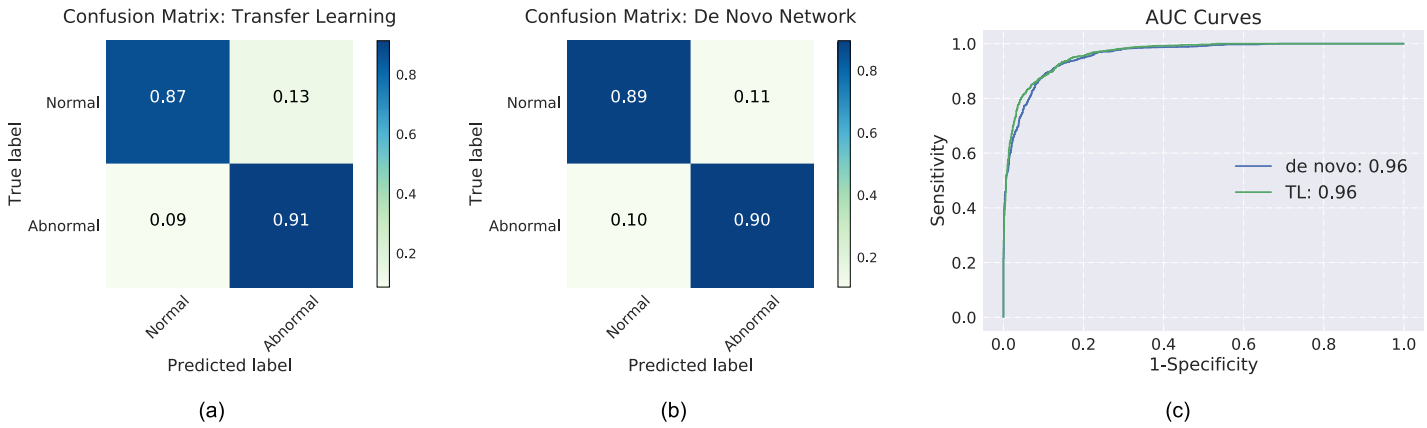

Fig 6. Confusion matrices for the (a) transfer learning (TL) and (b) the *de novo* network. (c) The AUC plot for the two networks.

https://doi.org/10.1371/journal.pone.0203726.g006

The mean FNR noted in the TL and *de novo* trained networks were 3.2% and 2.9% ($p < 0.01$), respectively. Visually inspection of the results obtained from a single split showed that approximately 90% of the false negatives obtained using either network contained small pathological conditions such as isolated drusen (see Fig 8(c)). However, there were also instances where geographic atrophy was not correctly identified as a pathology (see Fig 8(d)). The dataset consists of horizontal as well as vertical scans [10], where normal B-scans close to the optic nerve head are visually very similar to geographic atrophy. Since the model did not incorporate this additional piece of information (horizontal or vertical scan), it is not surprising to see misclassifications of this particular pathology.

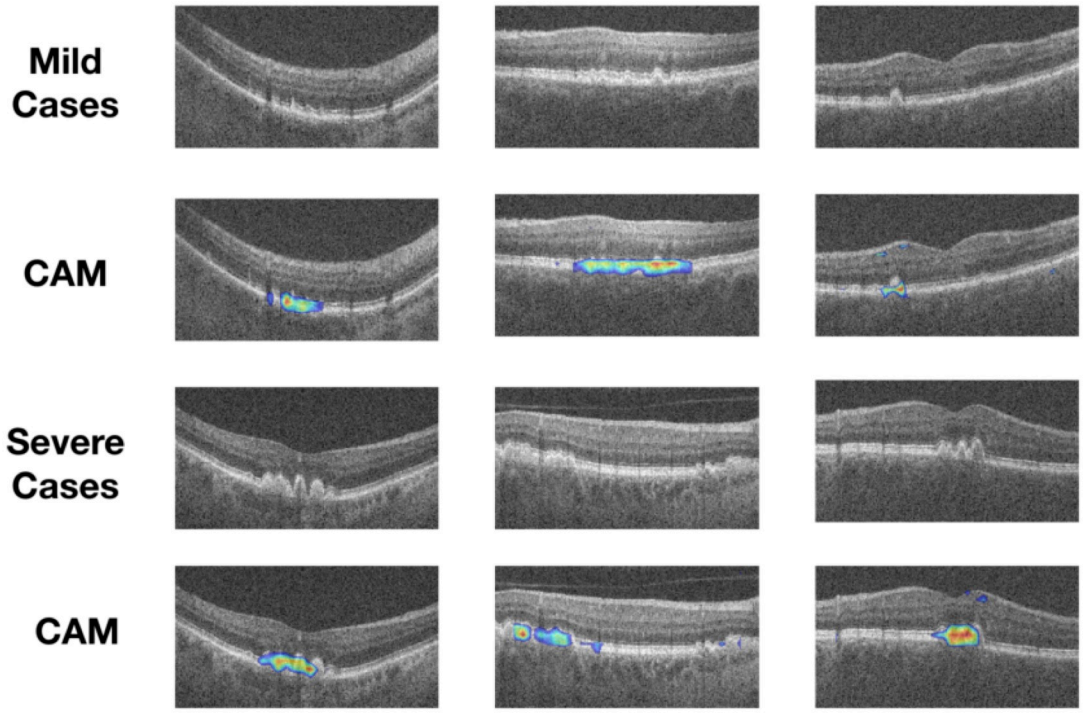

Fig 7. Examples of AMD B-scans with small/mild (top row) and significant pathologies (third row). CAMs from the *de novo* network for the same images are shown in the second and fourth rows, respectively.

https://doi.org/10.1371/journal.pone.0203726.g007

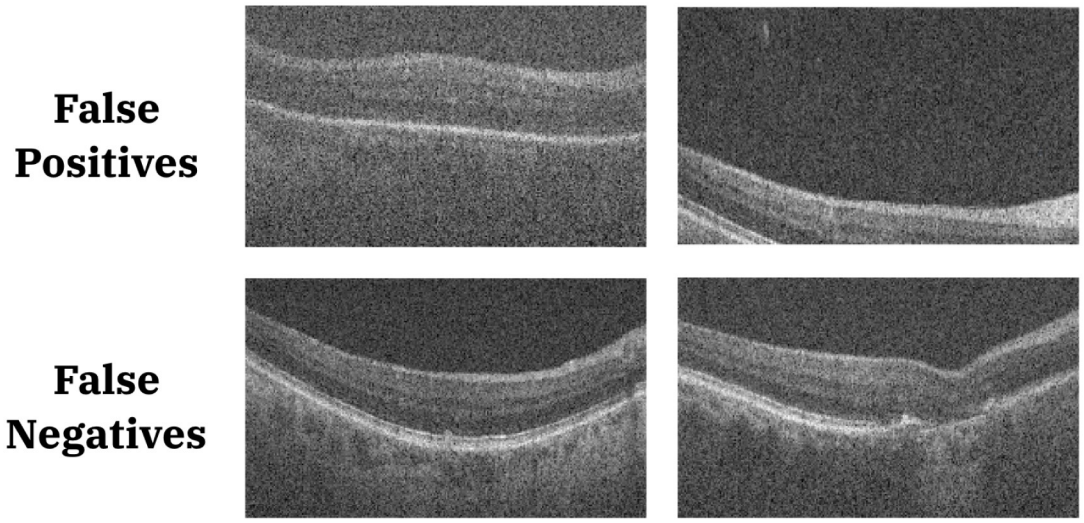

Fig 8. Examples of false positives from control subject scans (top row), and false positives in AMD scans (bottom row).

https://doi.org/10.1371/journal.pone.0203726.g008

4.1 Summarisation result

An example of a visual summary generated by the system is displayed in Fig 9. This scan showed pigment epithelial detachment at the fovea, and only a single region was identified by the summarisation rules.

A second example of an SDOCT volume with drusen and geographic atrophy is presented in Fig 10. Here three separate regions were detected by the summarisation algorithm as indicated by the blue, red and green regions in Fig 10(b). The final visual summary consisting of the B-scans that represent the three regions depicted in Fig 10(c). The individual B-scans from the three regions are shown in the three rows Fig 10(d)–10(l), with the colours of the bounding boxes corresponding to the location indicated in Fig 10(c).

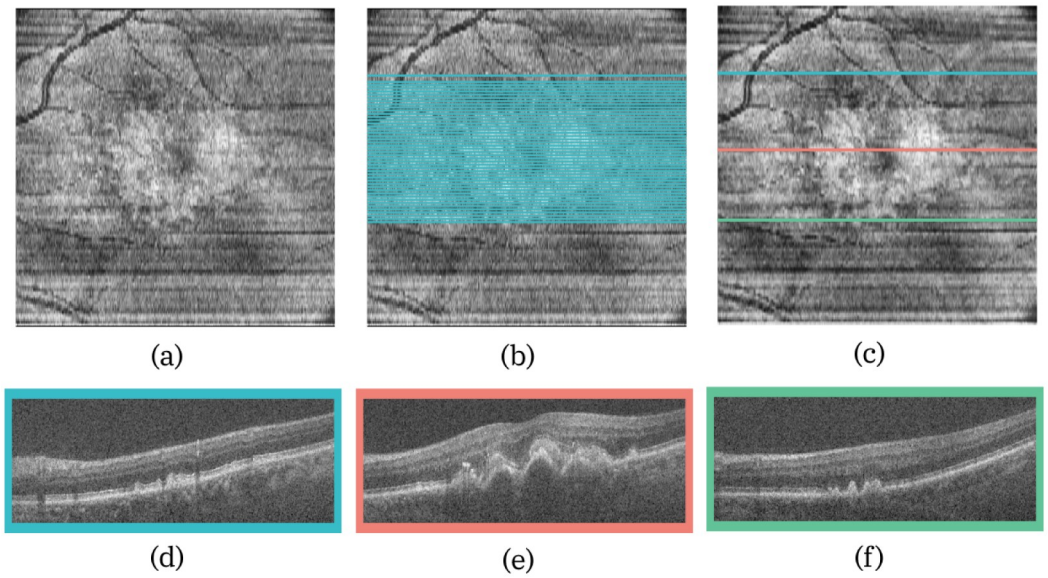

Fig 9. (a) The en-face projection image of the volume, with (b) all the key B-scans and (c) the final visual summary. The B-scans that correspond to the three locations are shown below, and colour coded to indicate their location in the volume.

https://doi.org/10.1371/journal.pone.0203726.g009

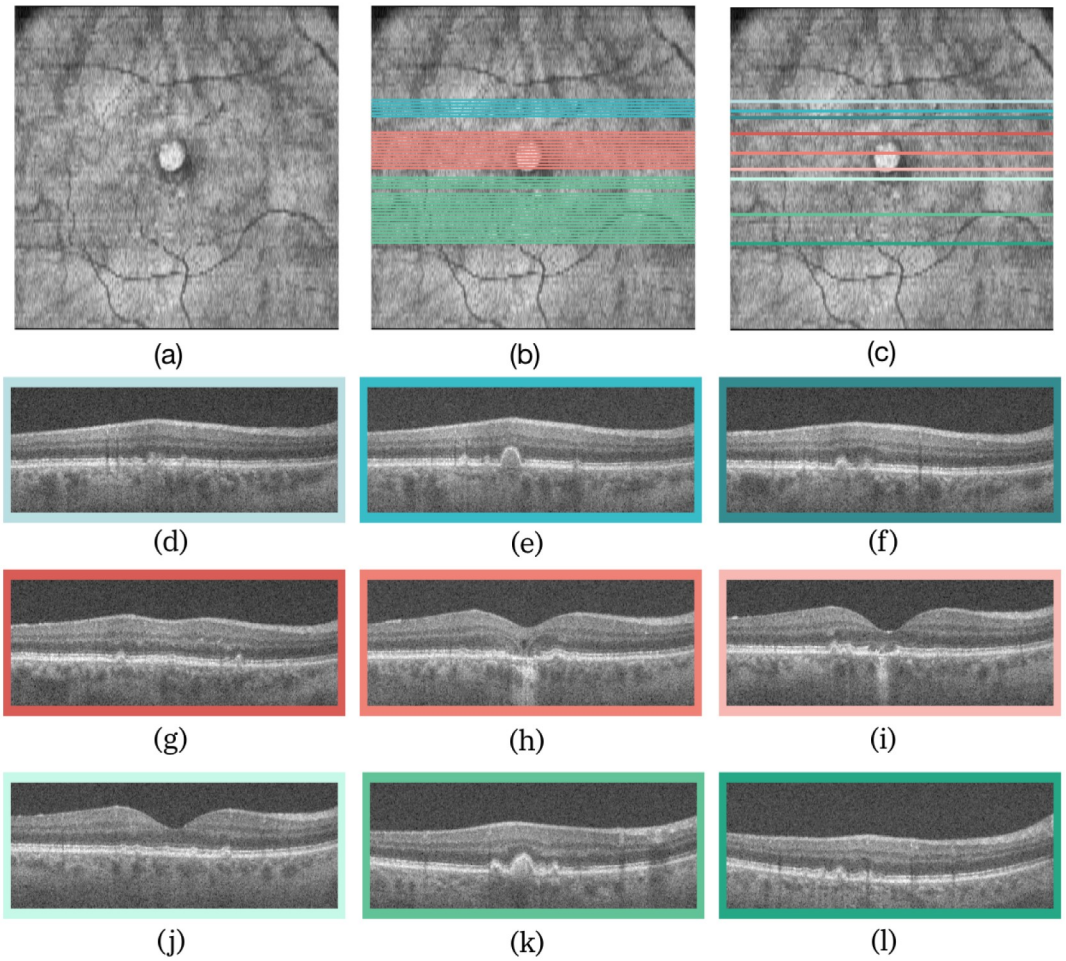

Fig 10. (a) The en-face projection image of the volume, with (b) all the key B-scans and (c) the final visual summary. The three B-scans corresponding (d)—(f) to the first region (blues), (g)—(i) the second region (reds), and (j)—(l) to the third region (greens).

https://doi.org/10.1371/journal.pone.0203726.g010

5 Discussion & conclusions

SDOCT finds extensive use in ophthalmology for the visualisation and quantification of structures in the retina. This high-resolution modality generates vast quantities of data ($\sim$ 50MB per volume), making the visual inspection of these images time-consuming, tiring and therefore error prone. Summarisation of the SDOCT volumes has been limited to the extraction of structural measurements, such as retinal layer thicknesses or optic nerve head parameters such as cup-to-disc ratio. However, other conditions such as epiretinal membranes or intra-layer cysts (that do not affect retinal layer thickness) require the manual inspection of the B-scans in the OCT volume. Visual summaries that retrieve key B-scans and identify relevant regions of the scan can be a valuable addition to the existing diagnostic framework.

Previously proposed summarisation methods for videos or volumetric medical images rely on the detection of relevant features, similar to our approach. However, our method does not explicitly segment AMD-related pathologies (see [9]), but uses a deep learning network for the detection of B-scans that show structural abnormalities. Our method can be extended to any

structural abnormality or even use a different definition of "relevance". For instance, a similar system could be designed to extract B-scans where maximal temporal change is identified. This would allow to monitor a variety of conditions such as AMD (dry or wet) or even glaucoma, where changes at the optic cup are recorded over time.

In this work, we also compared a commonly used technique—transfer learning—with a *de novo* trained network. Medical imaging applications typically do not have large datasets to work with and thus, transfer learning is commonly used. Here, we reduced the size of the network as well as the size of the input image, and gauged the ability of a lightweight *de novo* trained network to accomplish the same task. The *de novo* networks was 92% smaller than the TL network, however, the AUC obtained using these two networks are equivalent, with no statistical difference between the two networks. It is also important to note here that the standard deviation of the AUCs was very small for both networks, indicating the robustness of both frameworks. The training and run-time tells a different story, with the TL network only requiring 3 days for training (on a K80 NVIDIA-GPU), while the *de novo* network required nearly 9 days. The run-time was, as expected, significantly smaller for the *de novo* network, which was able to classify an entire SDOCT volume in four seconds, while the larger transfer learning network took 190 seconds (computed on a 2.5GHz Intel Core i7, 16GB RAM system).

The pre-defined expectations of inputs was also found to be a hindrance rather than a benefit in this particular application.

A separate experiment was conducted with the transfer learning network, where the input consisted of three adjacent B-scans instead of a replication of a single B-scan. Intuitively, one would expect that the use of adjacent B-scans would bolster the network's ability to detect the key B-scans, but this network performed worse. The conclusion to be drawn is not that the adjacent B-scans have no additional useful information, but that the transfer learning network is ill-equipped to leverage this. The VGG-16 network was originally designed for three-channel colour images where each channel presented different colour characteristics of the same image. Here, the adjacent B-scans might show differing structures (healthy B-scans adjacent to one with small drusen), and the network was not able to efficiently leverage this additional information. A *de novo* network, designed and trained for this, might do better, but we did not pursue this in the current work.

The key aspect of the *de novo* network remains its ability to be designed and trained for specific applications. In this instance, examples of this adaptability include the easy incorporation of a the global average pooling layer (to generate the CAMs), as well as skip-connections (known to assist in training). The inclusion of the CAM brings a degree of "explainability" to the system, where this visualisation indicates the source of the final class label. This output could also be used as an input to the rules that generate the visual summaries, where the size of would CAMs impact the inclusion in the visual summary.

In conclusion, the presented *de novo* network allows for the rapid and reliable detection of key B-scans in SDOCT volumes, which are used to generate visual summaries of volumes. In the future, we intend to extend the summarisation techniques to other disease models, as well as explore the use of small networks, designed specifically for the task at hand.

Author Contributions

Conceptualization: Bhavna Josephine Antony, Rahil Garnavi.

Investigation: Bhavna Josephine Antony, Stefan Maetschke.

Methodology: Bhavna Josephine Antony, Stefan Maetschke.

Supervision: Rahil Garnavi.

Validation: Bhavna Josephine Antony.

Writing – original draft: Bhavna Josephine Antony.

Writing – review & editing: Stefan Maetschke, Rahil Garnavi.

References

1. Liu T, Zhang HJ, Qi F. A novel video key-frame-extraction algorithm based on perceived motion energy model. IEEE Transactions on Circuits and Systems for Video Technology. 2003; 13(10):1006–1013. https://doi.org/10.1109/TCSVT.2003.816521
2. Fauvet B, Bouthemy P, Gros P, Spindler F. A geometrical key-frame selection method exploiting dominant motion estimation in video. International Conference on Image and Video Retrieval. 2004; p. 419–427.
3. Nam J, Tewfik AH. Detection of gradual transitions in video sequences using B-spline interpolation. IEEE Transactions on Multimedia. 2005; 7(4):667–679. https://doi.org/10.1109/TMM.2005.843362
4. Huang P, Hilton A, Starck J. Automatic 3D video summarization: key frame extraction from self-similarity. The Fourth International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT'08). 2008; p. 71–78.
5. Cotsaces C, Pitas I, Nikolaidis N. Video shot detection and condensed representation: a review. IEEE Signal Processing Magazine. 2006; 23(2):28–37. https://doi.org/10.1109/MSP.2006.1621446
6. Gibson D, Spann M, Woolley SI. A wavelet-based region of interest encoder for the compression of angiogram video sequences. IEEE Transactions on Information Technology in Biomedicine. 2004; 8 (2):103–113. https://doi.org/10.1109/TITB.2004.826722 PMID: 15217255
7. Syeda-Mahmood T, Beymer D, Wang F, Mahmood A, Lundstrom RJ, Shafee N, et al. Automatic selection of keyframes from angiogram videos. Proceedings—International Conference on Pattern Recognition. 2010; p. 4008–4011.
8. Huang D, Swanson EA, Lin CP, Schuman JS, Stinson WG, Chang W, et al. Optical coherence tomography. Science. 1991; 254(5035):1178–1181. https://doi.org/10.1126/science.1957169 PMID: 1957169
9. Chakravarthy U, Goldenberg D, Young G, Havilio M, Rafaeli O, Benyamini G, et al. Automated identification of lesion activity in neovascular age-related macular degeneration. Ophthalmology. 2016; 123 (8):1731–1736. https://doi.org/10.1016/j.ophtha.2016.04.005 PMID: 27206840
10. Farsiu S, Chiu SJ, Connell RVO, Folgar FA, Yuan E, Izatt JA, et al. Quantitative classification of eyes with and without Intermediate age-related macular degeneration using optical coherence tomography. Ophthalmology. 2013; p. 1–11.
11. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553):436–444. https://doi.org/10.1038/nature14539 PMID: 26017442
12. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances In Neural Information Processing Systems; 2012. p. 1097–1105.
13. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICRL); 2015. p. 1–14. Available from: http://arxiv.org/abs/1409.1556.
14. Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: fully convolutional DenseNets for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2016. p. 1175–1183. Available from: http://arxiv.org/abs/1611.09326.
15. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv 160600915v2. 2016; p. 1–14.
16. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017; 42:60–88. https://doi.org/10.1016/j.media.2017.07.005 PMID: 28778026
17. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018; 172(5):1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010 PMID: 29474911
18. Devalla SK, Chin KS, Mari JM, Tun TA, Strouthidis NG, Aung T, et al. A deep learning approach to digitally stain optical coherence tomography images of the optic nerve head. Investigative Opthalmology & Visual Science. 2018; 59(1):63–74.
19. Shah A, Abramoff MD, Wu X. Simultaneous multiple surface segmentation using deep learning. arXiv preprint 170507142v1. 2017.

20. He Y, Carass A, Jedynak BM, Solomon SD, Saidha S, Calabresi PA, et al. Topology guaranteed segmentation of the human retina from OCT using convolutional neural networks. arXiv preprint 180305120v1. 2018.
21. de Vos BD, Berendsen FF, Viergever MA, Staring M, Isgum I. End-to-end unsupervised deformable image network. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. 2017; p. 204–212.
22. Yang X, Kwitt R, Styner M, Niethammer M. Quicksilver: Fast predictive image registration—a deep learning approach. NeuroImage. 2017; 158:378–396. https://doi.org/10.1016/j.neuroimage.2017.07.008 PMID: 28705497
23. Dwarikanath M, Antony BJ, Sedai S, Garnavi R. Deformable medical image registration using generative adversarial networks. In: IEEE International Symposium on Biomedical Imaging; 2018. p. 1449–1453.
24. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision. 2015; 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y
25. Chollet F. Keras; 2015. Available from: https://keras.io.
26. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous systems; 2015. Available from: http://www.tensorflow.org.
27. Maetschke S, Tennakoon R, Vecchiola C, Garnavi R. nuts-flow/ml: data pre-processing for deep learning. arXiv preprint 170806046v2. 2018.
28. Kingma DP, Ba JL. Adam: a method for stochastic optimization. International Conference on Learning Representations 2015. 2015; p. 1–15.
29. Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics. 1979; 9:62–66. https://doi.org/10.1109/TSMC.1979.4310076
30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–778. Available from: http://arxiv.org/pdf/1512.03385v1.pdf.
31. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 2921–2929. Available from: http://arxiv.org/abs/1512.04150.