Few-shot Learning Using a Small-Sized Dataset of High-Resolution FUNDUS Images for Glaucoma Diagnosis

Mijung Kim Jasper Zuallaert Wesley De Neve
Ghent University - imec, IDLab, Department of Electronics and Information Systems, Belgium
Ghent University Global Campus, Center for Biotech Data Science, Songdo, Republic of Korea
{mijung.kim, jasper.zuallaert, wesley.deneve}@ugent.be

ABSTRACT

Deep learning has recently attracted a lot of attention, mainly thanks to substantial gains in terms of effectiveness. However, there is still room for significant improvement, especially when dealing with use cases that come with a limited availability of data, as is often the case in the area of medical image analysis. In this paper, we introduce a novel approach for early diagnosis of glaucoma in high-resolution FUNDUS images, only requiring a small number of training samples. In particular, we developed a predictive model based on a matching neural network architecture, integrating a high-resolution deep convolutional network that allows preserving the high-fidelity nature of the medical images. Our experimental results show that our predictive model is able to obtain higher levels of effectiveness than vanilla deep convolutional neural networks.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **Life and medical sciences**;

KEYWORDS

Deep learning; glaucoma diagnosis; matching networks; medical image analysis; one-shot learning

1 INTRODUCTION

Deep learning has recently achieved remarkable levels of effectiveness in the area of visual content understanding. In particular, as illustrated by the ImageNet benchmark, state-of-the-art deep learning models are currently more effective than humans in detecting objects and classifying images [8]. Furthermore, Google recently introduced a deep learning model for highly effective diagnosis of diabetic retinopathy [17], a disease caused by diabetes, leading to serious eye impairment. This predictive model attracted substantial attention from the medical community, given that it demonstrates the usefulness of deep learning for medical image analysis.

In practice, however, the application of deep learning techniques, which are data hungry in nature, is sometimes not possible due to overfitting, and where the latter is typically caused by a lack of data. Therefore, an increasing number of research efforts are dedicated to modifying deep learning techniques so that they can be successfully applied to sets having a limited number of data points.

The occurrence of small-sized image datasets is common in the area of medical image analysis, and where the images in question typically have a high resolution in order to facilitate ease of diagnosis by human experts. As a result, in our research, we focus on applying deep learning techniques to small-sized datasets of high-resolution medical images. In addition, to overcome the limited availability of medical images, we leverage few-shot learning, given that the combined application of deep learning and few-shot learning has recently demonstrated to come with a high potential [12]. Among several problems in the field of medical image analysis, we decided to steer our research towards early diagnosis of an eye disease called glaucoma, receiving expert help from Samsung Medical Center.

Glaucoma is one of the leading causes of human vision loss in the world [11]. The disease finds its origin in an increasing eye pressure, damaging the optical nerve, with patients gradually losing peripheral vision, leading to tunnel vision, and in the end, to a complete loss of vision. Fortunately, glaucoma can be controlled through early diagnosis and proper medicine and treatment. As shown in Figure 1, ophthalmologists, which are specialists in medical and surgical eye problems, diagnose glaucoma by examining the eyes of patients using various types of eye images, including fundus imaging (hereafter referred to as FUNDUS), Retinal Nerve Fiber Layer (RNFL) imaging, Optical Coherence Tomography (OCT) imaging of the optical disc and for macular measurement, and/or perimetry images. Together, these different types of images can help in reaching the correct diagnosis.

For radiologists, however, it is time consuming to capture different types of images and to subsequently examine them in a manual way. Moreover, with the exception of general hospitals, most ophthalmologists do not have access to all of the aforementioned image types. Therefore, the development of an effective computational model for early diagnosis of glaucoma, only making use of one type of image, has the potential to save a significant number of patients from vision loss. Given this observation, we obtained a FUNDUS image dataset from Samsung Medical Center in Korea, and we subsequently developed a predictive model for the early diagnosis of glaucoma, leveraging state-of-the-art techniques for both deep learning and few-shot learning.

Our paper is organized as follows. In Section 2, we review related work. Next, in Section 3, we provide details about our network architecture. We subsequently discuss our experimental setup and results in Section 4. Finally, we conclude our paper in Section 5.

MMHealth'17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-5504-9/17/10…$15.00
https://doi.org/10.1145/3132635.3132650

(a) Normal vision (b) Intermediate glaucoma (c) Advanced glaucoma (d) FUNDUS eye image

Figure 1: When people suffer from glaucoma, they see differently: (a) people with normal vision (peripheral vision); (b) as glaucoma develops, patients gradually lose their vision; and (c) people with advanced glaucoma, only seeing a small portion of an object (i.e., *tunnel vision*). Finally, a FUNDUS eye image (that is, an image of the back of the eye) can be seen in (d), with the arrows indicating where the optic disc and the macula reside in the image (the macula is an oval-shaped pigmented area near the center of the retina of the human eye). The region-of-interest is mainly between the optic disc and the macula.

2 RELATED WORK

In this section, we review several machine learning techniques, paying particular attention to few-shot learning, matching networks (MNs), and high-resolution convolutional neural networks (CNNs).

2.1 Few-shot Learning

Despite the current success of deep neural networks in various application domains, it is still a challenge to apply these networks to small-sized datasets. To overcome this challenge, Google DeepMind introduced a few-shot learning approach in 2016 [12]. Based on meta-learning [14, 15] and Memory-Augmented Neural Networks (MANNs) like Neural Turing Machines (NTMs; [4]), the newly introduced approach only needs a few samples per class for training purposes (that is, one, five, or ten), outperforming Long Short-Term Memory (LSTM) [5] and humans for the task of Omniglot [1] classification.

2.2 Matching Networks

MNs were developed by the same team that developed the few-shot learning approach described in Section 2.1. In particular, MNs were introduced in [16], emphasizing two design aspects: (1) the use of an attention mechanism that leverages cosine similarity and softmax, and (2) the adoption of the machine learning principle that test and train conditions must match. Furthermore, the embedding function of the MNs consists of a CNN stack and succeeding LSTM layers. However, unlike for the ImageNet task [2] and the Penn Treebank (PTB) task [10], the additional LSTM layers did not bring significant improvement for the Omniglot task. The MNs did outperform a MANN and a convolutional Siamese network [7].

2.3 High-Resolution CNNs

Most deep learning approaches that target visual classification tasks make use of various techniques to reduce the dimensionality of the input images, including downscaling the original image resolution [3]. Since downscaling may cause a loss of key features in medical images, an observation that explains why radiologists prefer the use of high-resolution imagery, the effectiveness of deep learning approaches may be hampered. For that reason, the authors of [3] decided not to downscale the original images, but to aggressively reduce their dimensionality by relying on CNN layers. Specifically, by leveraging high-resolution CNNs, they were able to improve the accuracy of their model for breast cancer screening.

3 ARCHITECTURE

Our approach mainly focuses on tackling two challenges: (1) applying deep learning techniques to small-sized datasets and (2) preserving the quality of the original images as much as possible when feeding them into the embedding function, so to be able to minimize the loss of helpful features. For dealing with small-sized datasets, we started from the MNs introduced in [16], and for facilitating the usage of high-resolution medical images, we started from the high-resolution CNNs introduced in [3]. We discuss the architecture of our neural network, which can be seen in Figure 2, in more detail in the next sections.

3.1 Attention Mechanism

The attention mechanism used by our approach is based on the attention mechanism used in [16]. Specifically, we first calculate the cosine distance c_T between each example image x_i of the support set $T = \{x_1, \ldots, x_n\}$ and a target image x_t:

$$c_T = \frac{\theta_T(x_t) \cdot \theta_T(x_i)}{\|\theta_T(x_t)\|_2 \|\theta_T(x_i)\|_2}, \tag{1}$$

where θ_T is the embedding function. This embedding function takes the form of a deep CNN that is able to deal with high-resolution imagery and that is followed by a bidirectional LSTM. Note that we make use of the same embedding function for both x_i and x_t. Next, we obtain a predicted label y_t_pred by using the softmax over the cosine distance:

$$y_t\_pred = \sum_{i=1}^{n} softmax(c_T) y_i, \tag{2}$$

where y_i is the corresponding label of x_i.

Note that, different from the MNs introduced in [16], our model is parametric in nature, given that there is no need for classifying unseen classes during validation and testing.

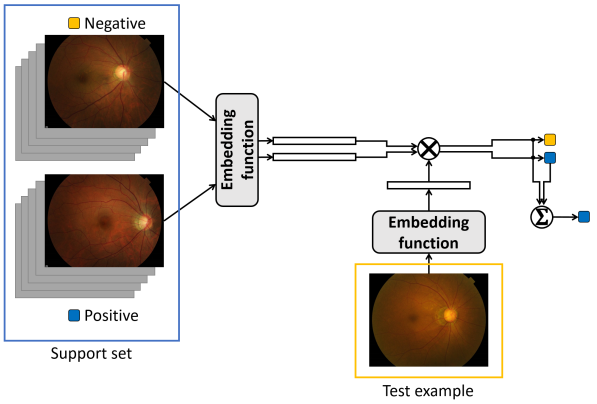

Figure 2: Overall architecture used by our model, showing five-shot learning. Five positive and five negative example images are fed into the embedding function, leading to a positive and a negative embedding, respectively. Contrary to the number of example images, only one target image (test example) is consistently used for any-shot learning.

3.2 Embedding Function

Our embedding function consists of two components: CNN layers and a bidirectional LSTM layer. Each component can be optionally implemented.

3.2.1 CNN Component. As shown in Table 1, the CNN component is a stack of deep convolutional neural layers that is able to deal with high-resolution images. First, we begin with coarse strides, two or three, at the lower convolutional layers and the max-pooling layers, so to be able to aggressively reduce input image dimensionality. After going through all the layers, the CNN component returns a flattened 6,400-D feature vector.

3.2.2 LSTM Component. Depending on the input, different LSTM functions are applied to the embedding generated by the CNN component. For the example input images, the bidirectional LSTM component g is trained over the flattened output of the CNN component. For a target image, the forward LSTM component f is used. Both outputs are then used for calculating the cosine distance and the subsequent softmax. The overall approach is as follows, with θ denoting the embedding function:

$$g(x_i) = biLSTM(\theta(x_i)) + \theta(x_i) \tag{3}$$

$$f(x_t) = LSTM(\theta(x_t)) + \theta(x_t) \tag{4}$$

$$a(f(x_t), g(x_i)) = softmax(f(x_t)^T \cdot g(x_i)) \tag{5}$$

Table 1: The CNN embedding function used by our predictive model.

Layer	Kernel size	Stride	Depth	Repetition
input 512×512×3				
convolution	3×3	2	32	-
max pooling	3×3	3	32	-
convolution	3×3	2	64	-
convolution	3×3	1	64	2
max pooling	2×2	2	64	-
convolution	3×3	1	128	3
max pooling	2×2	2	128	-
convolution	3×3	1	128	3
max pooling	2×2	2	128	-
convolution	3×3	1	256	3
global average pooling			256	-

4 EXPERIMENTS

In this section, we discuss the outcome of several experiments, comparing our approach with other state-of-the-art neural network architectures. Our main focus is on studying the effectiveness of binary classification for our small-sized dataset of high-resolution medical images. To evaluate our model, we make use of accuracy: (#True Positives + #True Negatives) / #Predictions. All approaches have 680 images in the training set, 200 images in the validation set, and 200 images in the test set.

4.1 Experimental Setup

4.1.1 Dataset. Our dataset consists of 1,080 high-resolution FUNDUS RGB images, made available by Samsung Medical Center in Korea. This dataset comes with two perfectly balanced classes: negative, which denotes absence of glaucoma, and positive, which denotes presence of glaucoma. Specifically, each class comes with 540 high-resolution FUNDUS images, with the resolution ranging from 1172×1500 to 2500×3200. Since the size of the images varies, we center-cropped each image to a region-of-interest with a size of 1024×1024, keeping the three RGB channels. In other words, we did not make use of downscaling, nor did we make use of grayscale conversion, so to be able to preserve the high-fidelity nature of the original images. Note that we decided to make use of center-cropping because all of the important features for diagnosing glaucoma are located between the optic disc and the macula, as shown in Figure 1.

4.1.2 Training. For comparison purposes, we made use of several state-of-the-art deep neural networks, as also included in Table 2. In general, we have used the default settings for the different neural networks [3, 13], using various input sizes.

First, we have run LeNet [9] with an image size of 256×256. We have then run VGG16 and Inception ResNet V2 with their default settings, as described in the respective papers. Furthermore, to alleviate the problem of overfitting, we applied data augmentation.

We implemented our predictive model by means of the TensorFlow framework developed by Google, training this model using NVIDIA Titan X GPUs. We center-cropped and eventually resized the input images to three different resolutions (that is, 256×256, 512×512, and 1024×1024), maintaining the three original color channels, thus making it possible study the impact of image quality on the effectiveness of classification. We set up our model once with and once without data augmentation.

Table 2: Results obtained by the different predictive models

Model	Input size	Data Aug.	Acc.
VGG-16	$224 \times 224 \times 3$	Yes	65.2%
Inception ResNet V2	$239 \times 239 \times 3$	Yes	89.5%
Our model (low)[a]	$256 \times 256 \times 3$	No	79.0%
Our model (low)	$256 \times 256 \times 3$	Yes	77.2%
Our model (mid)[b]	$512 \times 512 \times 3$	No	81.2%
Our model (mid)	$512 \times 512 \times 3$	Yes	83.4%
Our model (high)[c]	$1024 \times 1024 \times 3$	No	88.1%
Our model (high)	$1024 \times 1024 \times 3$	Yes	87.9%

[a] low denotes a center-cropped image down-sized to 256×256.
[b] mid denotes a center-cropped image down-sized to 512×512.
[c] high denotes the use of the original center-cropped image.

Similar to the ImageNet setup discussed in [16], we make use of a 1-shot, 5-shot, 10-shot, and 20-shot approach per class, thus feeding 1, 5, 10, or 20 positive example images per class and 1, 5, 10, or 20 negative example images per class to the embedding function, and with this function subsequently returning a flattened feature vector for both the positive and the negative images. Similar to [16], these feature vectors are then used to predict a label for an unseen target image using the attention mechanism (i.e., using softmax over cosine distance). The loss was calculated as described in [16] as well, using the ADAM optimizer [6] with a learning rate of 0.01.

In summary, all inputs go through the CNN layers and optionally the LSTM layers, and then through the attention mechanism. Next, the loss, which is calculated based on the last output, is optimized using ADAM.

4.2 Experimental Results

As can be seen in Table 2, VGG-16 did not perform well for the given dataset. Not shown in Table 2, LeNet returned an accuracy of 48.4% for the same task. Considering the balanced nature of the dataset, VGG-16 and LeNet were close to random guessing. Inception ResNet V2 obtained the highest accuracy among all experiments, but this architecture typically requires massive data augmentation, as also pointed out in [13].

Our approach made use of 1-shot, 5-shot, 10-shot, and 20-shot learning. Similar to [16], we could observe the following: the more samples we had per class, the more accurate the results obtained. The same observation could also be made regarding the input image size: the higher the spatial resolution, the better.

Given the above, Table 2 only shows the accuracy results obtained for 20-shot learning, given that 20-shot learning was consistenly outperforming 1-, 5-, and 10-shot learning with a significant margin. As an example, 1-shot learning making use of the highest image resolution was only able to achieve an accuracy of 54.51%.

Finally, we could observe that the models using LSTM layers after the CNN component did not perform well, an observation that is in line with the conclusions of [16] for Omniglot. Therefore, Table 2 also omits the results obtained for an embedding function that consists of both a CNN and LSTM component.

Given that human experts have a diagnosis accuracy of about 80%, the proposed approach demonstrates a higher effectiveness.

5 CONCLUSIONS AND FUTURE RESEARCH

In this paper, we introduced a novel approach towards early diagnosis of glaucoma in medical images, using a few-shot learning technique that leverages a high-resolution CNN. Our experimental results indicate that the effectiveness of our approach is promising, even when training is done by making use of a small-sized dataset.

In future research, we plan to evaluate our approach for different types of diseases and different types of images. In addition, we plan to apply our approach to text-based medical datasets. Finally, we will investigate whether further improvements can be realized through the use of additional data augmentation techniques.

ACKNOWLEDGMENTS

The research effort described in this paper was funded by Ghent University, the Ghent University Global Campus, imec, Flanders Innovation & Entrepreneurship (VLAIO), the Fund for Scientific Research-Flanders (FWOFlanders), and the European Union.

REFERENCES

[1] Simon Ager. 2008. Omniglot writing systems and languages of the world. *Retrieved January* 27 (2008), 2008.
[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.
[3] Krzysztof J Geras, Stacey Wolfson, S Kim, Linda Moy, and Kyunghyun Cho. 2017. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. *arXiv preprint arXiv:1703.07047* (2017).
[4] Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401* (2014).
[5] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[6] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[7] Gregory Koch. 2015. *Siamese neural networks for one-shot image recognition.* Ph.D. Dissertation. University of Toronto.
[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[10] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics* 19, 2 (1993), 313–330.
[11] H A Quigley and A T Broman. 2006. The number of people with glaucoma worldwide in 2010 and 2020. *British Journal of Ophthalmology* 90, 3 (2006), 262–267. https://doi.org/10.1136/bjo.2005.081224 arXiv:http://bjo.bmj.com/content/90/3/262.full.pdf
[12] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065* (2016).
[13] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261* (2016).
[14] Sebastian Thrun. 1998. Lifelong learning algorithms. In *Learning to learn*. Springer, 181–209.
[15] Ricardo Vilalta and Youssef Drissi. 2002. A perspective view and survey of meta-learning. *Artificial Intelligence Review* 18, 2 (2002), 77–95.
[16] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*. 3630–3638.
[17] Tien Yin Wong and Neil M Bressler. 2016. Artificial Intelligence With Deep Learning Technology Looks Into Diabetic Retinopathy Screening. *JAMA* 316, 22 (2016), 2366–2367.