Evaluating Transfer Learning for Macular Fluid Detection with Limited Data

Alex Cazañas-Gordón
University of Coimbra
Department of Electrical and Computer Engineering
Instituto de Telecomunicações
Coimbra, Portugal
acazanas@deec.uc.pt

Esther Parra-Mora
University of Coimbra
Department of Electrical and Computer Engineering
Instituto de Telecomunicações
Coimbra, Portugal
eparra@deec.uc.pt

Luís A. da Silva Cruz
University of Coimbra
Department of Electrical and Computer Engineering
Instituto de Telecomunicações
Coimbra, Portugal
lcruz@deec.uc.pt

***Abstract*—The ability to transfer knowledge learned with large datasets to train classifiers where labeled data is scarce, has made transfer learning a prime step on deep learning applications. The availability of models pre-trained on large datasets makes possible to apply deep learning to a wide variety of computer aided diagnosis tasks including retinal image processing. However, with dozens of deep convolutional network architectures, choosing the right model for transfer learning is not a trivial task. While current art implies that the deeper the model the better, here we verified this assertion by fine-tuning a suite of deep convolutional network architectures on a scenario with limited resources. Contrary to what was expected, deeper architectures did not perform as well with limited data as with large datasets.**

***Index Terms*—transfer learning, fine tuning, convolutional neural networks, macular fluid detection, optical coherence tomography**

I. Introduction

Excessive accumulation of fluid within the central portion of the retina – the macula, signals damage of the vasculature of the eye and is often the consequence of retinal diseases such as age-related macular degeneration (AMD), retinal vein occlusion (RVO) or diabetic macular edema (DME). In a healthy eye the macula is the most sensitive area of the retina and is responsible for the central, high-resolution, color vision. When leakage from damaged blood vessels occurs nearby the central retina, fluid pockets form causing edema and deteriorating the macular function. The condition degrades the vision impeding patients to see objects in fine detail. With the development of anti-vascular endothelial growth factor (anti-VEGF) therapy, patients of exudative retinal diseases can retain useful vision. However, early detection and continuous monitoring are critical for effective treatment. Diagnosis of retinal pathology relies on a wide variety of imaging modalities among which spectral domain optical coherence tomography (OCT) has become a standard for guiding macular edema therapy. OCT is a non-invasive modality that produces high resolution volumetric images by acquiring a series of 2D cross-sectional slides (B-scans) of the retina and underlying structures. In the OCT of a healthy eye, the inner retina exhibits a smooth contour with a gradual depression at the center of the macula. By contrast, the OCT of an eye with macular edema shows a thickened retina with hypo-reflective areas indicating fluid-filled regions. Fig. 1 depicts an OCT of a normal eye next to an OCT of a retina with macular edema. Although OCT allows to detect and quantify fluid accumulation within sub-retinal layers, manual examination of B-scans is labor intensive and requires specialized knowledge. Considering the growing prevalence of retinal diseases, increasing demand of eye care might put pressure over the healthcare system in the non-distant future. AMD and DME are now leading causes of blindness worldwide and are projected to increase in the years to come due to rising global pervasiveness of diabetes and aging population [1], [2]. This has motivated a great deal of research on computer aided diagnosis of macular pathology based on the processing of OCT images.

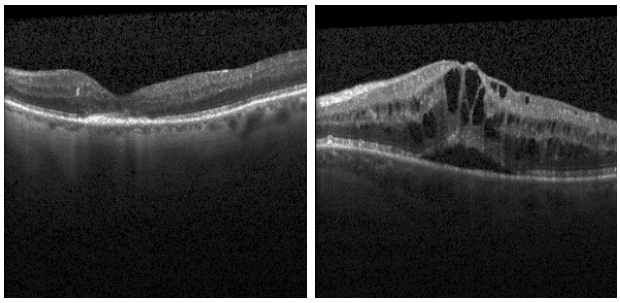

Fig. 1. OCT retinal B-scans centered around the macula. Left: normal eye. Right: macular edema. Images from RETOUCH [3]

Following the success of deep convolutional neural networks in image classification tasks, deep learning methods have become the preferred approach to solve computer vision problems. Deep learning delivers high performance in image classification tasks, but it is costly in that it requires large sets of annotated data. This limits the application of deep learning to scenarios where annotated data is abundant and

This work was funded by the Secretariat of Higher Education, Science, Technology and Innovation of the Republic of Ecuador. The authors acknowledge the support of FCT Portugal project UIDB/EEA/50008/2020.

inexpensive, which is not the case of medical imaging. A widespread strategy to overcome scarcity of data is to take advantage of data representations learned with large datasets to learn the representations of a different set of images but with less annotated data. The operating assumption of learning transfers is that networks trained in large datasets should have learned to distinguish a large enough set of low-level features which can be re-purposed as building blocks of high-level features belonging to the representation of a different set of data. In a convolutional neural network low-level features, e.g. edges, are captured by the first layers, while subsequent layers learn to detect increasingly high-level features as we go deeper into the network. Were low-level features to be reused, there would be no need to train earlier layers but only train the later ones to learn the features that are specific to the new data. This makes possible training deep convolutional networks with much less labeled data than training from a blank network. While transfer learning may be implemented in a few different ways, the prevalent variant is that in which a network trained in a large dataset like ImageNet, is fine-tuned to learn the representation of new data, e.g. X-ray, MRI, or OCT.

A. Related work

The outstanding performance of deep convolutional networks on ImageNet Large Scale Visual Recognition Challenge (ILSVRC) has encouraged porting top-performer architectures to domains where labeled data is not abundant. Based on the assumption that knowledge gained by networks trained in ImageNet's massive dataset could be transferred to smaller datasets, transfer learning has become a mainstay in medical image classification. Transfer learning has been successfully applied to automated diagnosis of various diseases including skin cancer [4], [5], diabetic retinopathy [6], age related macular degeneration (AMD) [7], [8], diabetic macular edema (DME) [9]–[11], and Alzheimer's disease [12].

In line with the chronological development of convolutional networks, recent work on detection of macular diseases on OCT exhibit a tendency to favor recent, deeper networks to conduct transfer learning. Kamble et al. [11] reported 99.46% average validation accuracy on binary classification of DME by fine-tuning a pre-trained inception-resnet-v2 model. Treder et al. [8] fine-tuned a pre-trained inception-v3 model and achieved 96% accuracy on the test set for binary classification of exudative AMD. Similarly, Kermany et al. [9] fine-tuned a pre-trained inception-v3 model and reached 93.4% accuracy on binary classification of referrals (urgent vs routine). Karri et al. [10] chose a pre-trained googlenet model for finetuning and reported mean decision pooling of 0.99, 0.89, and 0.86 at classifying cases into: normal, AMD, and DME respectively.

The prevalence of ImageNet top performer architectures suggests that the benchmark's rankings play a principal role on model selection. Comparing performance, the highest score belonged to inception-resnet-v2, which is the deepest architecture. The second deepest architecture, inception-v3, delivered lower performances. However, the least deep architecture, googlenet, achieved a performance that was in between the performances of the deepest and the second deepest architectures. Regarding the size of the train sets range from tens of thousands images [8], [10], [11] to one hundred thousand images [9]. Considering that producing reliable labels require trained staff, such volumes of labeled data imply a high cost associated to data alone.

Despite the increasing application of transfer learning to process retinal images little attention has been drawn towards evaluating deep convolutional networks fine-tuned with limited data. A recent study compared five deep network architectures: alexnet, densenet, inception, resnet, and vgg for multi-class classification of retinal disease on a dataset containing over a hundred thousand OCT B-scans [13]. The study involved a single training per model on a high parallel computing facility. Similarly, Kamble et al. [11] compared performance of three fine-tuned models: inception-resnet-v2, inception-v3, and resnet-50. However, as was the case of works on the literature review, performance rather than data availability was the main concern.

With deep learning, high performance comes at the cost of large amounts of annotated data being needed. Transfer learning is a strategy to reduce deep learning requirements, and so it is increasingly applied to automated detection of macular diseases. However, should we look at dataset sizes in the literature review, it becomes apparent that transfer learning is deployed with large amounts of annotated data in the thousands of images range. This beg the questions: Is the use of large amounts of annotated data necessary to achieve state-of-the-art performance? Would competitive performance be achievable with fewer data, e.g. less than one thousand images? How do deep neural networks behave in such limited data scenarios?

In this study we conducted a comparative evaluation of deep convolutional neural network architectures in a scenario with limited data. The evaluation encompasses popular architectures in the state of the art with focus on retinal fluid detection. A thorough analysis of performance resulting from fine-tuning pre-trained models aims to provide a clear picture on how different model architectures behave with limited data. Moreover, this study provides empirical data to support decision making on network architecture selection for transfer learning when labeled data is scarce.

II. Data and methods

Training and test data were sourced from the Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge RETOUCH [3]. The source dataset contains OCT volumes with various retinal fluid conditions and reference annotations. The OCT volumes were acquired with three different OCT devices, and the reference standard was obtained from manual voxel-annotations of fluid lesions by human graders from two medical centers. All OCT volumes are stored in ITK MetaImage format containing an ASCII readable header and a separate raw image data file.

For this study we processed a set of 24 OCT volumes acquired with an OCT Spectralis device (Heidelberg Engineer-

ing, Heidelberg, Germany) and extracted 1176 individual B-scans size 512x496 pixels. Samples of the dataset are shown in Fig. 1. In addition, we labeled each B-scan with one of two labels: "fluid" or "normal". The dataset composition after labeling was 711 B-scans (60%) with label "fluid", and 465 B-scans with label "normal" (40%).

Ten convolutional neural network models trained on ImageNet were fine-tuned to classify the selected B-scans. The model selection included a representative sample of architectures with varying depth, size, and recency. Model sources trained on ImageNet were obtained from MATLAB [14]. Table I lists the main characteristics of the evaluated architectures as well as the batch size training parameter which will be described on the protocol description below.

Training and data pre-processing were conducted with MATLAB 2019b (The MathWorks, Inc., Natick, Massachusetts, United States) and CUDA library 9.0 (NVIDIA Corporation, Santa Clara, California, United States). The models were trained on a computer with Windows 10 operating system, processor Intel i7 8700K CPU @ 3.7 GHz - 6 cores and 32 GB RAM, using a GPU NVIDIA GeForce GTX 1080 Ti with 11 GB RAM.

TABLE I
EVALUATED NETWORK ARCHITECTURES

Network	Params	Depth	Input	BS
vgg16	138	16	224x224	32
alexnet	61	8	227x227	64
inceptionresnetv2	55.9	164	299x299	16
inceptionv3	23.9	48	299x299	64
xception	22.9	71	299x299	16
densenet201	20	201	224x224	16
resnet18	11.7	18	224x224	64
googlenet	7	22	224x224	64
mobilenetv2	3.5	53	224x224	64
squeezenet	1.24	18	227x227	64

Params: Number of parameters in millions.
Input: Image input size. BS: Largest batch size.

A. Protocol Details

Each pre-trained network was modified for binary classification by removing the last classification layer, softmax layer and output layer; all remaining layers were left as they were. A final classification layer was added with the number of classes set to two. The weights of the new classification layer were initialized with the Glorot initializer [15].

To match models' input size we center-cropped the OCT B-scans to the size of each architecture input (Table I). Inputs were subsequently rescaled to the range [-1 1]. Apart from rescaling and cropping, no additional enhancement was performed. We did not use data augmentation either.

Every model was trained with learning rate 1e-4 and momentum 0.9. We set an early stop condition based on the validation loss value where we checked the loss value twice per epoch and halted training whenever the loss failed to improve for more than five consecutive checkpoints. If the early condition was not met, models were trained for a maximum of 10 epochs. Training data were shuffled before each epoch to avoid overfitting due to accommodation of data ordering. Model loss was computed with the crossentropy function.

We evaluated two experimental designs to investigate how deep convolutional networks perform with limited data. For the first design, we performed a stratified split of the data such that the training set contained 90% of the available data while the remaining 10% was used for validation. For the second scenario, we reversed the splitting percentages, so the training set used 10% of the available data and the validation set used the remaining 90%.

In addition, in both designs and for different batch sizes, we recorded the number of iterations elapsed up until the best performance was achieved. This number was used as a measure of model efficiency. Given that each architecture has different memory requirements, we estimated the largest batch size for each architecture and varied the batch size from size 8 up to the largest size that fitted in memory. The last column of Table I shows the estimated largest batch size for each network.

For all experiments we performed 10-fold cross-validation with stratified splits of the training set. To evaluate performance, we chose accuracy as primary metric due to its prevalence in literature. Moreover, we computed the area under receiver operating curve (AUC) as complementary metric for further insight into the model prediction performance. All performance metrics were evaluated at every iteration during training, and we kept the model that maximized accuracy.

III. RESULTS AND ANALYSIS

Table II presents mean and standard deviation values of accuracy, and AUC of top-performing architectures in the baseline design (train set = 90%) . Both metrics were computed across ten folds on the validation set. Overall, we observed that performance decreased as batch size increased, which suggest that setting batch size to the largest value that fits in memory would be sub-optimal performance wise. This is consistent with Keskaret al. [16] who observed a decrease in performance with largest batch sizes.

TABLE II
MEAN (STANDARD DEVIATION) ACCURACY, AUC OF TOP PERFORMERS ACROSS TEN FOLDS OF THE BASELINE DESIGN (TRAIN SET = 90%).

Batch size	Architecture	Accuracy	AUC
8	vgg16	0.92 (0.06)	0.97 (0.04)
	alexnet	0.88 (0.08)	0.93 (0.08)
	googlenet	0.88 (0.10)	0.91 (0.12)
16	vgg16	0.90 (0.07)	0.96 (0.05)
	alexnet	0.86 (0.07)	0.93 (0.08)
	googlenet	0.87 (0.09)	0.92 (0.16)
32	vgg16	0.90 (0.07)	0.96 (0.04)
	googlenet	0.86 (0.09)	0.90 (0.16)
	alexnet	0.85 (0.09)	0.92 (0.11)
64	alexnet	0.82 (0.09)	0.91 (0.11)
	inceptionv3	0.79 (0.07)	0.83 (0.16)
	googlenet	0.71 (0.05)	0.87 (0.13)

Regarding the duration of the training, we observed that bigger batch sizes led to shorter training but at cost of diminishing accuracy. Fig. 2 shows mean accuracy versus average iterations to best accuracy. Batch sizes are shown next to corresponding markers and the averages were computed across the ten folds. If we look at the data points of one network, e.g. vgg16, we can observe that batch size 32 shorten the duration of training by almost half compared to batch size 8; however the accuracy of the larger batch size is lower (0.9) than that of the shorter (0.92). The same observation holds for corresponding batch sizes with networks: alexnet, and inceptionv3. Furthermore, we observed that accuracy and iterations approximate diminishing-return curves where performance increase at a rate that is inversely proportional to the increase of iterations (Fig. 3). For instance, in the curve describing vgg16 (top-most trend line), accuracy values belonging to batch sizes 32, 16, and 8 are separated less than 2 percent however the number of iterations to reach the highest value (0.92) almost double the iterations required for the second highest (0.9). The same observation stood for the other top-performing architectures. This observation coincides with those of [17] which noted the possibility of finding "sweet spots" where performance balances out computing efficiency.

Given that the early-stop condition resulted in variable training duration, we conducted an additional experiment where models were trained for a fixed number of iterations with every batch size. Here, we observed that the bigger the batch size the earlier the validation loss deteriorated, most likely due to model overfitting. In addition, we observed that beyond the early-stopping point the accuracy saturated or decreased. This confirmed the early stopping criterion used effectiveness in preventing overfitting and achieving high training efficiency by avoiding unnecessary training iterations.

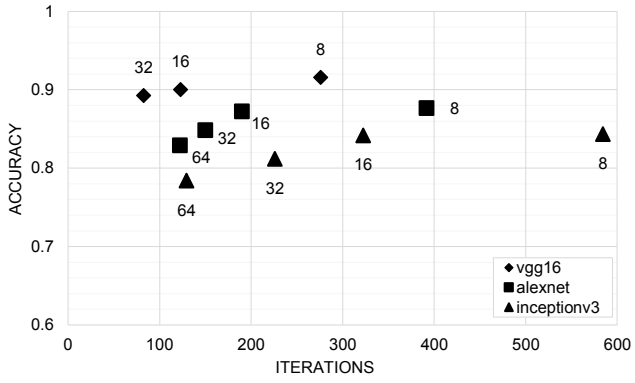

Fig. 2. Mean accuracy versus iterations of top performers. Data labels denote batch size. Iterations to reach best accuracy averaged across ten folds of the baseline design (train set = 90%)

Looking at the top-performing architectures, vgg16 delivered the highest accuracy followed by alexnet and googlenet. These architectures also ranked top-3 with other batch sizes, except with batch size 64 where vgg16 was not tested due to memory limitations. Contrary to what was expected, the number of parameters of the architecture was a better performance

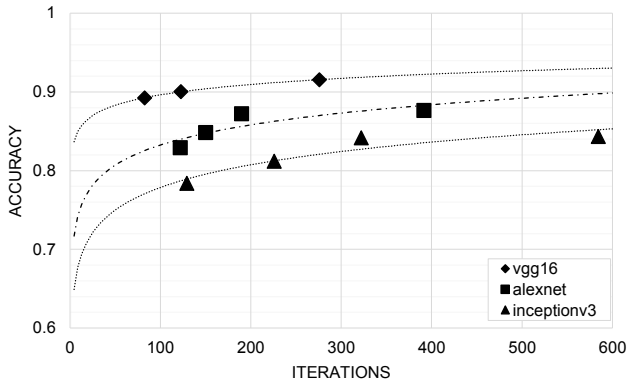

Fig. 3. Mean accuracy versus iterations of top performers architectures grouped by architecture for batch size: 8,16, and 32. Iterations to reach best accuracy averaged across ten folds of the baseline design (train set = 90%)

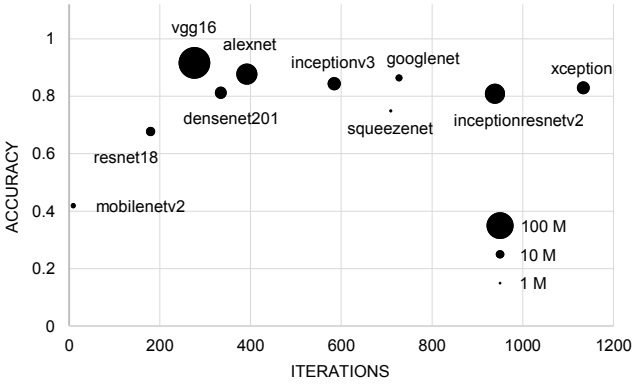

Fig. 4. Mean accuracy versus iterations of top performers architectures across ten folds for batch size: 8. 1M = 1 million. Iterations to reach best accuracy averaged across ten folds of the baseline design (train set = 90%)

predictor than depth. Fig. 4 shows mean accuracy against mean iterations, along with the number of parameters (represented by bubble size) of each architecture. As it can be observed, top performers: alexnet and vgg16, are the architectures with the largest number of parameters. It is also observed that these two architectures reach on average their best performance faster than other architectures scoring over 80% accuracy. This suggests a higher algorithm efficiency from more complex architectures than from streamlined models.

Previous work noted an inverse relationship between the number parameters and performance on ImageNet top-1error ranking, placing early, more complex networks like alexnet and vgg16 below more recent, less complex networks such as inception-resnetv2 and densenet201 [18]. However, we observed the opposite behavior in that networks with more parameters performed better than more streamlined models when trained with limited data.

Although the main objective of this study was not optimizing performance, the results here presented are realistic in that they are comparable with current art. For instance, RETOUCH reported mean (standard deviation) AUC of 0.93 (0.05) across teams for fluid detection on OCT scans acquired with Spectralis device [3]. Nevertheless, we believe that there is room for

improvement in the form of hyperparameter optimization, data augmentation, and input image preprocessing, e.g. denoising.

As per the second experimental design (train set = 10%), we observed that training with very limited data resulted in significant reduction of accuracy (Table III) almost matching the lowest performances of the baseline design. Consistent with the baseline experiment, larger batches resulted in faster training, and lower accuracy; while networks with larger number of parameters achieved higher accuracy within fewer iterations.

TABLE III
MEAN (STANDARD DEVIATION) ACCURACY, AUC ACROSS TEN FOLDS OF THE LIMITED-DATA DESIGN (TRAIN SET = 10%).

Batch size	Architecture	Accuracy	AUC
8	vgg16	0.80 (0.06)	0.87 (0.06)
	alexnet	0.77 (0.07)	0.82 (0.07)
	googlenet	0.73 (0.06)	0.77 (0.12)
16	vgg16	0.78 (0.05)	0.86 (0.05)
	alexnet	0.72 (0.06)	0.80 (0.07)
	densenet201	0.70 (0.03)	0.79 (0.04)

IV. CONCLUSIONS

Our results coincide with previous work on deep convolutional networks regarding the usefulness of transfer learning to overcome data scarcity. Ten deep convolutional network architectures were fine-tuned to conduct binary classification in a scenario where the supply of annotated data was limited.

Contrary to the ideal scenario where provision of annotated data is abundant, our results showed that recent, deeper, streamlined architectures did not deliver higher performance than older architectures, which are relatively more complex and shallower.

We also observed that more complex, shallower architectures reached their best performance faster than deeper, less complex models. This suggests that the earlier provide higher efficiency than the latter in a limited-data scenario.

In line with previous work on network hyperparamater, we observed a trade-off between computing efficiency and performance. We verified that large batch sizes shorten the training duration but at cost of diminishing accuracy. On the other hand, smaller batch sizes lead to higher accuracy but at cost of longer training.

Our results challenge the widespread reliance in very deep models and large datasets to deliver high performance in macular fluid detection. We achieved competitive performance with limited data and shallower architectures than those in the state of the art. Nevertheless, we believe that further research is needed. e.g. by extending this work to other imaging modalities and datasets.

REFERENCES

[1] K. L. Pennington and M. M. DeAngelis, "Epidemiology of age-related macular degeneration (AMD): associations with cardiovascular disease phenotypes and lipid factors," *Eye and vision (London, England)*, vol. 3, pp. 34–34, 2016.

[2] W. L. Wong, X. Su, X. Li, C. M. G. Cheung, R. Klein, C.-Y. Cheng, and T. Y. Wong, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis," *The Lancet Global Health*, vol. 2, no. 2, pp. e106–e116, 2014.

[3] H. Bogunović, F. Venhuizen, S. Klimscha, S. Apostolopoulos, A. Bab-Hadiashar, U. Bagci, M. F. Beg, L. Bekalo, Q. Chen, and C. Ciller, "Retouch: The retinal oct fluid detection and segmentation benchmark and challenge," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1858–1874, 2019.

[4] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[5] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle, "Knowledge transfer for melanoma screening with deep learning," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, Conference Proceedings, pp. 297–300.

[6] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative ophthalmology & visual science*, vol. 57, no. 13, pp. 5200–5206, 2016.

[7] P. M. Burlina, N. Joshi, M. Pekala, K. D. Pacheco, D. E. Freund, and N. M. Bressler, "Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks," *JAMA Ophthalmology*, vol. 135, no. 11, pp. 1170–1176, 2017.

[8] M. Treder, J. L. Lauermann, and N. Eter, "Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 256, no. 2, pp. 259–265, 2018.

[9] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. L. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. N. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, and K. Zhang, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.

[10] S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," *Biomedical Optics Express*, vol. 8, no. 2, pp. 579–592, 2017.

[11] R. M. Kamble, G. C. Y. Chan, O. Perdomo, M. Kokare, F. A. González, H. Müller, and F. Mériaudeau, "Automated diabetic macular edema (DME) analysis using fine tuning with inception-resnet-v2 on oct images," in *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2018, Conference Proceedings, pp. 442–446.

[12] Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, D. Lituiev, T. P. Copeland, M. S. Aboian, and C. Mari Aparici, "A deep learning model to predict a diagnosis of alzheimer disease by using 18F-FDG PET of the brain," *Radiology*, vol. 290, no. 2, pp. 456–464, 2019.

[13] R. Gelman, "Evaluation of transfer learning for classification of:(1) diabetic retinopathy by digital fundus photography and (2) diabetic macular edema, choroidal neovascularization and drusen by optical coherence tomography," *arXiv preprint arXiv:1902.04151*, 2019.

[14] MATLAB, "Pretrained deep neural networks," 2019, Accessed on: 2020-02-28. [Online]. Available: https://www.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html

[15] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, Conference Proceedings, pp. 249–256.

[16] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.

[17] Y. Bengio, *Practical recommendations for gradient-based training of deep architectures*. Springer, 2012, pp. 437–478.

[18] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64 270–64 277, 2018.