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

Machine Learning Method for Functional Assessment of Retinal Models

Nikolas Papadopoulos*, Nikos Melanitis†, Antonio Lozano‡, Cristina Soto-Sanchez‡, Eduardo Fernandez‡ and Konstantina S. Nikita†

***Abstract*— Challenges in the field of retinal prostheses motivate the development of retinal models to accurately simulate Retinal Ganglion Cells (RGCs) responses. The goal of retinal prostheses is to enable blind individuals to solve complex, real-life visual tasks. In this paper, we introduce the functional assessment (FA) of retinal models, which describes the concept of evaluating the performance of retinal models on visual understanding tasks. We present a machine learning method for FA: we feed traditional machine learning classifiers with RGC responses generated by retinal models, to solve object and digit recognition tasks (CIFAR-10, MNIST, Fashion MNIST, Imagenette). We examined critical FA aspects, including how the performance of FA depends on the task, how to optimally feed RGC responses to the classifiers and how the number of output neurons correlates with the model's accuracy. To increase the number of output neurons, we manipulated input images - by splitting and then feeding them to the retinal model - and we found that image splitting does not significantly improve the model's accuracy. We also show that differences in the structure of datasets result in largely divergent performance of the retinal model (MNIST and Fashion MNIST exceeded 80% accuracy, while CIFAR-10 and Imagenette achieved ∼40%). Furthermore, retinal models which perform better in standard evaluation, i.e. more accurately predict RGC response, perform better in FA as well. However, unlike standard evaluation, FA results can be straightforwardly interpreted in the context of comparing the quality of visual perception.**

***Index Terms*— retinal models, functional assessment, machine learning, retinal prosthesis, visual recognition**

I. INTRODUCTION

The increasing knowledge of visual systems along with technological advances give novel results in prevention, limitation or even treatment of visual impairment [1], [2]. However, retinal degenerative diseases, such as age-related macular degeneration (AMD) and retinitis pigmentosa cannot be effectively treated with surgery or medication [3]. Retinal prosthesis devices aim to restore vision in such patients, by translating visual stimuli to electrical stimulations that activate the retina. Then, the retina transmits neural signals to the visual centers of the brain, which are responsible for visual perception [4].

Although current retinal implants have managed to restore certain visual functionalities, there are several technological and biological challenges to be overcome. In particular, implants need to simulate natural retinal processing, by incorporating models that faithfully predict the physiological firing of retina cells to visual stimuli [3], [4]. Towards this direction, the development of retinal models aims to simulate the biological neural processing in the retina, by interpreting images to retinal responses. Nowadays, advanced retinal models incorporate computer vision techniques [5], [6] and, more recently, state-of-the-art models use Convolutional Neural Networks (CNNs) [1], [2].

This paper proposes the functional assessment (FA) of retinal models, which describes the concept of evaluating the performance of retinal models on visual understanding tasks. This constitutes a divergence from the currently common practice of evaluating a retinal model comparing the similarity of model-generated and ground-truth RGC responses. Motivation for FA stems from the observation that visual prostheses aim to restore the capacity of individuals to comprehend their visual environment, thus we should directly evaluate our models on such tasks. The need for FA of prosthetic (i.e., acquired through prostheses) vision has been raised in the literature [7]; FA has been applied to evaluate vision in implantees [8] and in augmented-reality interventions in people with severe vision impairment [9]. FA should be focused on the visual functions that seem most important to the blind: mobility, face recognition and reading [3]. In this context, we develop a machine learning method for FA. Using images from well-established computer vision datasets (Table I), we feed traditional classifiers with RGC responses produced by the retinal model, in order to solve object and digit recognition tasks. Our goal is to explore whether retinal models that faithfully reproduce retina output show improved performance in visual understanding tasks and also, draw conclusions on the quality of prosthetic vision that is attainable by assessing the performance of retinal models directly on such tasks.

II. METHODS

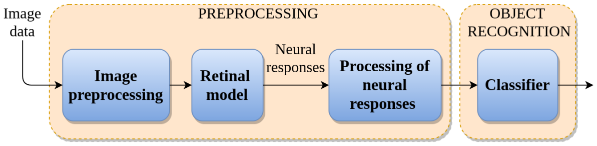

Fig. 1. Functional assessment pipeline.

A. Functional assessment pipeline

Functional assessment pipeline (Fig. 1) includes image preprocessing (Steps II-A.1, II-A.2, II-A.3, II-A.4), feeding

* Nikolas Papadopoulos is with the School of Electrical and Computer Engineering, National Technical University of Athens, Greece nikpap555@gmail.com

† K.S. Nikita and N. Melanitis are with the School of Electrical and Computer Engineering, National Technical University of Athens, Greece

‡ E.Fernandez, C. Soto-Sanchez and A. Lozano are with Instituto de Bioingenieria, Universidad Miguel Hernandez, Alicante, Spain

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

the retinal model with images (Step II-A.5), extracting and processing RGC responses (Steps II-A.6, II-A.7, II-A.8) and finally, feeding them to the classifiers. All steps, together with the corresponding design decisions, are analyzed below.

1) Input: Transform each image from rgb to grayscale, only for CIFAR-10 and Imagenette. Then, normalize images in range (0, 1) and finally, reshape each image to (50, 50, 1) to match the input size of the retinal model.

2) Data augmentation: Implement data augmentation as a standard process to improve model training. Two of the following transformations are applied to each image: *rotate* 45^o or -45^o, use *gaussian noise* of scale = $0.1*255$ or scale = $0.2*255$, *crop* images by 5 or 7 pixels from each side, *translate* images over x or y axis by a percentage ranging from -10% to $+10\%$ of the image size. Parameter values are selected, so as to produce different perspectives of the same image and, at the same time, preserve the object for detection inside the frame of the image. Data augmentation is implemented using *imaug* library[1].

3) Split: Split image in p^2 parts, where $p = 1, 2, 3, 4$ and reshape the size of each part to (50, 50). For $p = 1$, we assume that the image is not split at all. Motivation behind image splitting is to artificially increase the limited number of output neurons (60 in our case) provided by the retinal model, so as to increase the performance of the classifiers.

4) Adjust: The goal of this step is to create a temporal dataset, as described in Section II-C. It should be decided whether to repeat each image five times and get the full response trend for t = 0 ... 50 ms (*Adjust=yes*) or have one row per image with ten image repetitions (see red frames in Fig. 2) and get a snapshot of the response at t = 100 ms (*Adjust=no*)[2].

5) Feed: The retinal model, described in Section II-B, is fed with images and predicts retinal responses for 60 neurons.

6) Valid: In this step, the receptive fields of the neurons are identified using STA Analysis [10]. If STA manages to compute the center of the receptive field for a neuron, then we consider this neuron as valid. For the retinal model used, only 12 valid neurons are found. Therefore, it should be decided whether to keep all 60 neurons (*Valid=no*) or only the 12 valid ones (*Valid=yes*).

7) Combine: If adjustment was previously implemented, there are five arrays of RGC responses for every image, which are combined in one array, by applying elementwise *min* or *max* transformation.

8) Concatenate: If image splitting is implemented in Step II-A.3, there is one array of 60 RGC responses for every part of the split image. These arrays are then concatenated to one larger array of size $a * (p^2)$, where $a = 60$ (or $a = 12$ if valid neurons are selected) and $p = 2, 3, 4$. The set of final arrays is the training dataset for the classifiers and thus, the size of final arrays represents the number of features for the classifiers.

B. Retinal Model

We use a 3-layer CNN retinal model [2] that was trained to predict response rates for sixty simultaneously recorded RGCs. This model was chosen, as it can effectively predict retinal responses to natural images and, being trained with natural images, it can model a wide range of retina's biological properties. To train the retinal model, we used an image dataset consisting of 4890 grayscale natural images of size 50x50 pixels and the recorded retinal responses (retinal responses were recorded at Prof. E. Fernandez lab) [1]. Each frame, corresponding to 10 ms of visual stimulus, was projected onto the retina of a mouse for a total of 50 ms. Thus, each frame was repeated five times and the whole dataset consists of 24450 natural images of total duration 244.5 s.

C. Temporal dataset

The response of RGCs depends not only on the current stimulus, but also on preceding stimulations. To model this temporal dependency, a temporal dataset is created. For each image being projected onto the retina, we keep track of the history of images projected before. The total number of frames used -the actual image projected onto the retina (at $t = t_n$) plus the additional image frames accounting for the stimulus history- is called *temporal_interval*. Every row in the temporal dataset represents an input sample to the retinal model. The number of image repetitions (n) represents the duration ($n * 10$ ms, given that each frame corresponds to 10 ms of visual stimulus) that the retina is exposed to a specific image (Fig. 2).

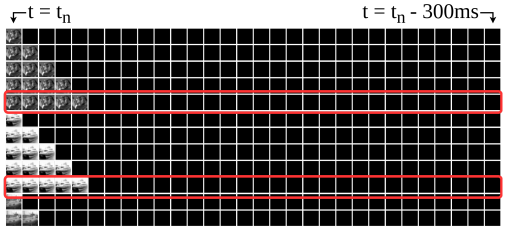

Fig. 2. A clipping of the temporal dataset. Every row in the temporal dataset represents an input sample to the retinal model and it has *temporal_interval* frames (here, *temporal_interval* $= 30$). In each row, the number of image repetitions ($n = 1, 2, 3, 4, 5$) represents the time that the image has been projected onto the retina ($\Delta t = 10, 20, 30, 40, 50$ ms respectively). Black frames represent the intermediate stage between the alternation of images, where no light is projected onto the retina.

D. Design of simulations

We conducted simulations for all different combinations of design decisions during data preprocessing. In particular, we trained four different classifiers: MLP_500_100 (size of input layer = 500, size of hidden layer = 100), MLP_n_n/2 (size of input layer = n, size of hidden layer = n/2, where n = number of features), SVM (kernel='rbf') and Random Forest (max_features=12). Hyperparameters were chosen based on the performance of classifiers in initial simulations. We also chose four datasets (Table I): CIFAR-10, MNIST, Fashion MNIST and Imagenette with ten different classes (ten object

[1] https://github.com/aleju/imgaug

[2] We found that the retina response at t=100 ms has high variance and thus, it could better distinguish different objects (data not shown).

categories) each, so as to compare classification tasks with equal number of classes. Then, 10000 samples from each dataset were augmented, creating in this way 30000 samples, which were randomly divided according to 70%/30% train/test split. For each simulation, we trained each classifier ten times and we calculated the mean and standard deviation of accuracy. We repeated all simulations for two different parametrizations of the CNN retinal model (Section II-B): *RetModel1* with *temporal_interval*=30 and *RetModel2* with *temporal_interval*=40.

TABLE I

DESCRIPTION OF THE DATASETS USED IN SIMULATIONS.

Dataset	Image size	Type of classification	# Classes
CIFAR-10 [11]	32x32	objects	10
MNIST [12]	28x28	digits	10
Fashion MNIST [13]	28x28	clothing items	10
Imagenette [3]	variable	objects	10

III. RESULTS AND DISCUSSION

A. Preprocessing decisions for functional assessment

Initially, we performed simulations with CIFAR-10 and MNIST, splitting images in *None*, $4, 9$ parts ($p = 1, 2, 3$ respectively) for both *Valid=yes/no* (Step II-A.6). Then, we split images of the two more complex datasets, Fashion MNIST and Imagenette, in *None*, $4, 9, 16$ parts for only *Valid=no*. In Fig. 3, we see the percentage differences between 60 neurons (*Split=None*) and max splitting (*Split=9* or *Split=16*). We observe that image splitting (i.e. using more than 60 neurons) increases the performance mostly in MNIST (>10%) and less in Fashion MNIST (<10%), while it does not further improve performance in CIFAR-10 and Imagenette. This can be explained by structural differences mentioned in Section III-B. We further assume that, by splitting images in too many parts, the objects are difficult to be recognized due to oversegmentation and so, performance does not significantly improve with splitting. In addition, if we compare plots with *Adjust=yes* and those with *Adjust=no* (Fig. 3), we see that, keeping a snapshot of the retina response at a critical t (at which retina response has high variance), produces similar results as keeping the full trend of the retina response over time. By keeping only the critical responses, we can also save significant computational resources and time. Finally, we tested a set of different classifiers and we found that the type of classifier is not an important factor -our conclusions remain unchanged, irrespective of the classifier we use- even if the Random Forest had the most efficient and consistent performance across the simulations (Fig. 3).

B. Functional assessment performance on different datasets

Fig. 4 compares the maximum performance of classifiers between different visual understanding tasks, i.e. different datasets (Table I). CIFAR-10 and Imagenette achieve $\sim$40% accuracy, while MNIST and Fashion MNIST exceed 80%. We see that image resolution is not a crucial factor in the model's performance. Although Imagenette was compared to CIFAR-10 to test images with higher resolution, it performs only slightly better -and in some cases- worse than CIFAR-10. Furthermore, MNIST and Fashion MNIST have twice the accuracy of CIFAR-10 and Imagenette, even if they consist of images with lower resolution. Significant differences in performance between MNIST/Fashion MNIST and CIFAR-10/Imagenette can be explained, if we take a closer look at the structural differences between these datasets. Images in both MNIST and Fashion MNIST have a dark background with the object for classification situated in the middle. On the other hand, CIFAR-10 and Imagenette consist of real-life images, with unclear background-foreground segregation and more complex structures. This plays a key role, if we additionally consider the retina only reacts in spatiotemporal changes among different image frames [14]. This retina's property favors both MNIST and Fashion MNIST, where only the object under classification changes over different dataset images.

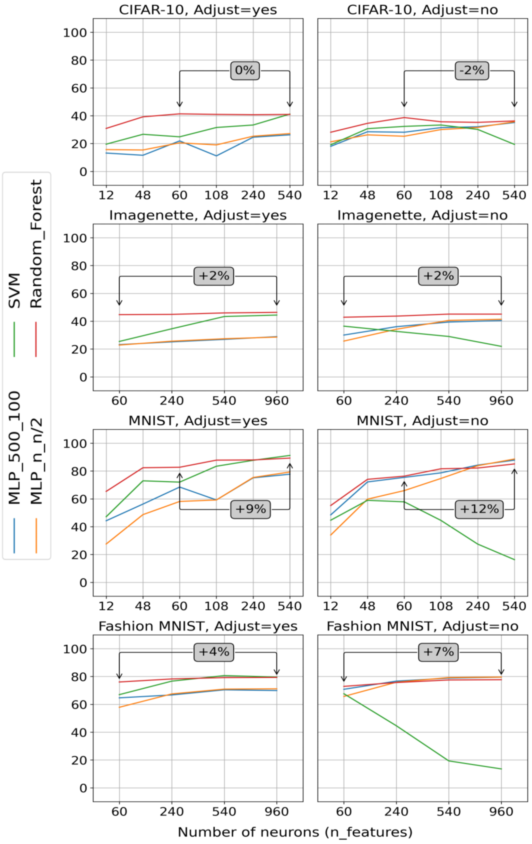

Fig. 3. Sensitivity plots correlating the accuracy of classifiers (CIFAR-10, MNIST, Fashion MNIST, Imagenette) with the number of neurons used as input (*n_features*), for both *Adjust=yes/no* (Step II-A.4).

C. Functional assessment of different retinal models

Finally, functional assessment is applied in order to compare the performance of two different retinal models,

[3] Available at https://github.com/fastai/imagenette/

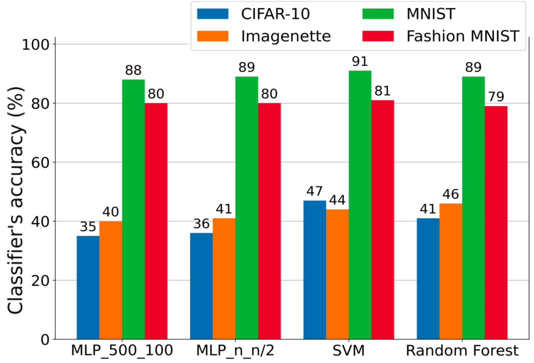

Fig. 4. Comparison of the maximum performance of classifiers between datasets.

RetModel1 and *RetModel2* (Section II-D). Results show that the retinal model with the lowest mean squared error (MSE) (i.e. the one being closer to the biological retina) performs better across the whole range of functional simulations. Therefore, it seems that functional assessment is in accordance with standard evaluation techniques. Furthermore, by functionally assessing a model, we get a direct and more easily interpretable estimate of how well an implantee may perform in a visual task of interest.

D. Limitations and future directions of functional assessment

During the development of this work, several limitations were encountered. We used a retinal model that has been trained to predict only a limited number of RGC responses (Section II-B). However, a higher number of RGC responses could better represent complex images and provide classifiers with a richer set of features to solve more efficiently real-life visual tasks. Therefore, there is a need to increase the number of RGC responses given by retinal models, either by using improved experimental methods to collect the data or using a different architecture for the retinal model.

Another limitation arises from choosing to process retinal responses, which are one-dimensional firing-rate arrays, with traditional machine learning classifiers. From a biological perspective, the visual pathway includes neural processing both in the retina and mainly, in the brain's visual centers, which are responsible for higher visual functionalities, like object recognition. Literature indicates that the brain's visual centers can be effectively modeled by deep neural architectures [15]. Deep learning networks have been also used, in high-impact research on biological vision, to model the ventral visual stream in order to elucidate retinal mechanisms [16]. Taking those insights into consideration, we suggest that future efforts for FA should be focused on combining end-to-end deep learning architectures to model both the retina and the rest of the visual pathway. Moreover, we may train on tasks with unsupervised methods, which produce biologically plausible ventral visual system models and follow bio-plausible sensory learning procedures [17]. Given also the interpretable nature of FA, future development involves explainable models, providing insights into the relationship between image properties and retina output [18].

IV. Conclusion

In this work, we introduced the concept of functional assessment and designed a machine learning framework for it. We investigated how retinal models, trained to faithfully reproduce retina output, perform in visual understanding tasks. We show that FA is comparable with the established evaluation method; yet, FA provides a direct and easily interpretable way of assessment, based on the performance on visual tasks. We also found that performance in FA is closely dependent on the given task. Finally, image splitting, as a way to increase the number of output neurons, does not significantly improve accuracy; still, restoring functional vision requires that retinal models interpret images to a larger number of RGC responses.

References

[1] A. Lozano, C. Soto, J. Garrigós, J. Martínez, J. Ferrández, and E. Fernandez, "A 3d convolutional neural network to model retinal ganglion cell's responses to light patterns in mice," *International Journal of Neural Systems*, vol. 28, 2018.

[2] L. McIntosh, N. Maheswaranathan, A. Nayebi, S. Ganguli, and S. Baccus, "Deep learning models of the retinal response to natural scenes," in *Advances in neural information processing systems*, 2016, pp. 1369–1377.

[3] E. Fernandez, "Development of visual neuroprostheses: trends and challenges," *Bioelectronic medicine*, vol. 4, no. 1, pp. 1–8, 2018.

[4] L. Yue, J. D. Weiland, B. Roska, and M. S. Humayun, "Retinal stimulation strategies to restore vision: Fundamentals and systems," *Progress in Retinal and Eye Research*, vol. 53, pp. 21–47, 2016.

[5] N. Melanitis and K. S. Nikita, "Biologically-inspired image processing in computational retina models," *Computers in Biology and Medicine*, vol. 113, p. 103399, 2019.

[6] A. Alevizaki, N. Melanitis, and K. Nikita, "Predicting eye fixations using computer vision techniques," in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2019, conference, pp. 309–315.

[7] B. Lepri, "Is acuity enough? other considerations in clinical investigations of visual prostheses," *Journal of neural engineering*, vol. 6, p. 035003, 2009.

[8] A. Demchinsky *et al.*, "The first deaf-blind patient in russia with argus ii retinal prosthesis system. what he sees and why," *Journal of Neural Engineering*, vol. 16, 2019.

[9] A. N. Angelopoulos, H. Ameri, D. Mitra, and M. Humayun, "Enhanced depth navigation through augmented reality depth mapping in patients with low vision," *Scientific Reports*, vol. 9, p. 11230, 2019.

[10] E. Chichilnisky, "A simple white noise analysis of neuronal light responses," *Network (Bristol, England)*, vol. 12, pp. 199–213, 2001.

[11] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[12] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *ATT Labs*, vol. 2, 2010.

[13] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *CoRR*, vol. abs/1708.07747, 2017.

[14] J. S. Werner and L. M. Chalupa, Eds., *The New Visual Neurosciences*. MIT Press, 2014.

[15] M. Schrimpf *et al.*, "Brain-score: Which artificial neural network for object recognition is most brain-like?" *bioRxiv*, 2020.

[16] J. Lindsey, S. A. Ocko, S. Ganguli, and S. Deny, "A unified theory of early visual representations from retina to cortex through anatomically constrained deep cnns," 2019.

[17] C. Zhuang *et al.*, "Unsupervised neural network models of the ventral visual stream," *Proceedings of the National Academy of Sciences*, vol. 118, no. 3, 2021.

[18] M. Athanasiou, K. Sfrintzeri, K. Zarkogianni, A. C. Thanopoulou, and K. S. Nikita, "An explainable xgboost–based approach towards assessing the risk of cardiovascular disease in patients with type 2 diabetes mellitus," in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2020, pp. 859–864.