frontiers in Computational Neuroscience

BRIEF RESEARCH REPORT
published: 24 August 2021
doi: 10.3389/fncom.2021.609721

Quantifying the Brain Predictivity of Artificial Neural Networks With Nonlinear Response Mapping

Aditi Anand[1,2], Sanchari Sen[2,3] and Kaushik Roy[2]*

[1] West Lafayette Junior/Senior High School, West Lafayette, IN, United States, [2] Center for Brain-Inspired Computing, School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, United States, [3] IBM Thomas J. Watson Research Center, Yorktown Heights, NY, United States

Quantifying the similarity between artificial neural networks (ANNs) and their biological counterparts is an important step toward building more brain-like artificial intelligence systems. Recent efforts in this direction use *neural predictivity*, or the ability to predict the responses of a biological brain given the information in an ANN (such as its internal activations), when both are presented with the same stimulus. We propose a new approach to quantifying neural predictivity by explicitly mapping the activations of an ANN to brain responses with a non-linear function, and measuring the error between the predicted and actual brain responses. Further, we propose to use a neural network to approximate this mapping function by training it on a set of neural recordings. The proposed method was implemented within the TensorFlow framework and evaluated on a suite of 8 state-of-the-art image recognition ANNs. Our experiments suggest that the use of a non-linear mapping function leads to higher neural predictivity. Our findings also reaffirm the observation that the latest advances in classification performance of image recognition ANNs are not matched by improvements in their neural predictivity. Finally, we examine the impact of pruning, a widely used ANN optimization, on neural predictivity, and demonstrate that network sparsity leads to higher neural predictivity.

Keywords: artificial neural networks (ANN), brain-inspired computing, neuromorphic systems, brain similarity, neural recordings, neural predictivity

OPEN ACCESS

Edited by:
Si Wu,
Peking University, China

Reviewed by:
Bao Ge,
Shaanxi Normal University, China
Jonathan Mapelli,
University of Modena and Reggio Emilia, Italy

****Correspondence:***
Aditi Anand
anand86@purdue.edu

Received: *24 September 2020*
Accepted: *26 July 2021*
Published: *24 August 2021*

Citation:
Anand A, Sen S and Roy K (2021) Quantifying the Brain Predictivity of Artificial Neural Networks With Nonlinear Response Mapping. Front. Comput. Neurosci. 15:609721. doi: 10.3389/fncom.2021.609721

INTRODUCTION

The fields of machine learning and neuroscience have a long and deeply intertwined history (Hassabis et al., 2017). In the quest for developing intelligent systems capable of learning and thinking by themselves, researchers have repeatedly looked for inspirations in the biological brain. The first generation of Artificial Neural Networks (ANNs) developed in the 1950s utilized perceptrons, which are abstract mathematical models of biological neurons (Rosenblatt, 1958). In subsequent generations of ANNs, engineering efforts to successfully train these networks eventually led to the design of artificial neuron models that differ from their biological counterparts. Simultaneously, researchers continued to seek and implement biological inspirations for improving

ANNs, including their structure and function. For instance, multi-layer convolutional neural networks developed in the 1990s (Fukushima, 1980; Lecun et al., 1998) were heavily inspired by the functioning of simple and complex cells in the human visual cortex (Hubel and Weisel, 1962). More recently, the development of attention networks (Vaswani et al., 2017) was motivated by the observation that human brains "attend to" certain parts of inputs when processing large amounts of information.

While the desire to emulate more advanced functions of biological brains serves as one driver of brain-inspiration in the field of ANNs, a second, equally important motivation arises from the need for efficiency. While ANNs have matched or surpassed human performance in many machine learning tasks, including image recognition, machine translation and speech recognition, the computational cost required to do so is quite high and increasing rapidly. Amidst the justified excitement about the success of artificial intelligence in man vs. machine contests such as IBM's Watson (IBM) and Google's AlphaGo (Deepmind AlphaGo), the gap in energy efficiency between artificial and natural intelligence continues to grow. Improved energy efficiency is crucial in the face of exploding computational requirements for training state-of-the-art ANNs on the one hand (AI and Compute, 1998), and the need to deploy them in highly energy-constrained energy devices on the other hand (Venkataramani et al., 2016). Recent efforts also suggest that biologically inspired mechanisms have the potential to improve the robustness of ANNs to adversarial attacks (Sharmin et al., 2019; Dapello et al., 2020).

Several efforts have explored the use of biologically inspired concepts for improving the energy efficiency and robustness of ANNs, or allowing them to learn from less data. Among these efforts, one group attempts to increase *representational similarity* at the individual neuronal and synaptic level. For instance, spiking neural networks comprise of neurons mimicking the firing behavior of biological neurons while employing different neural coding schemes (Maass, 1997). A second group of efforts explore biologically inspired learning rules like Spike-Timing-Dependent Plasticity (STDP) (Bi and Poo, 1998). Finally, other efforts attempt to create ANNs with topologies that are derived from neuroanatomy (Riesenhuber and Poggio, 1999). In summary, prior efforts have taken various approaches in the attempt to identify desirable features of biological brains and embody them in ANNs.

In this work, we focus on quantifying the information similarity between ANNs and biological networks by comparing their internal responses to a given input stimulus (Schrimpf et al., 2018, 2020). This approach was pioneered by Brain-score (Schrimpf et al., 2018), which quantifies information similarity through a combination of a behavioral sub-score and a neural predictivity sub-score. We specifically focus on *neural predictivity*, which refers to the ability to predict the responses of a biological brain given the information from an ANN (such as its internal activations), when both are presented with the same stimulus. Brain-Score utilizes the Pearson correlation coefficient to capture the correlation between ANN activations and neural recordings from the macaque visual cortex (Schrimpf et al., 2018). The use of Pearson's correlation coefficient implicitly assumes a linear relationship between the ANN activations and neural responses. Alternative metrics such as Mutual Information can quantify correlation under non-linear relationships (Cover and Thomas, 2006). However, methods to compute Mutual Information are only known when the tensors being compared have the same rank and dimensions, which is not true when comparing ANN activations with neural recordings.

In this work, we advocate the use of an explicit, non-linear mapping function to predict neural responses from ANN activations. The rationale behind this approach is that ANN activations are themselves a product of non-linear transformations. In addition, there does not exist a one-to-one correspondence between ANN and brain layers, decreasing the likelihood that the relationship between ANN activations and neural recordings can be modeled by a linear function. A second key idea that we propose is the use of a neural network to approximate the mapping function itself. We note that this is a regression problem, where the form of the function mapping from ANN activations to neural recordings is unknown. Neural networks, which are known to be universal function approximators, have been successfully applied to many regression problems. Hence, we explore their use in our work.

Embodying the approach outlined above, we propose a new method for neural response prediction in order to quantify the informational similarity between an ANN and a set of brain recordings. The method utilizes a neural network, called the neural response predictor (NRP) network, to model the non-linear relationship between ANN activations and brain recordings. Input stimuli (in our case, images) are fed to the ANN, and the activations of its layers are extracted. These activations, along with the corresponding neural recordings (captured after presentation of the same stimuli to a primate) (Schrimpf et al., 2018), are then used to train the NRP network. The prediction error of the NRP network, termed NRP-error, is a quantitative measure of the ANN's neural predictivity.

We implement the proposed method within the TensorFlow (Tensorflow, 2015) machine learning framework and apply it to calculate the NRP-errors of 8 state-of-the-art image classification ANNs. We utilize neural recordings from the IT (168 recording sites) and V4 (88 recording sites) regions of primate brains for 3,200 images (Schrimpf et al., 2018) to evaluate the proposed method. Our results demonstrate considerable improvement in neural predictivity over linear models which are used in previous approaches (Schrimpf et al., 2018). Our results reaffirm the finding that recent advances in image classification ANNs (from AlexNet to Xception) are not accompanied by an improvement in neural predictivity. Finally, we also evaluate the impact of commonly used network optimizations such as pruning on neural predictivity.

MATERIALS AND METHODS

In this section, we first describe the general concept of quantifying brain similarity through neural predictivity. We next present the proposed method to quantify neural predictivity and

finally discuss the experimental setup and methodology used to evaluate our proposal.

Quantifying Brain Similarity Through Neural Predictivity

Neural predictivity refers to the ability to predict biological neural behavior using the information inside an ANN. As illustrated in Eq. 1, one way to quantify neural predictivity is to explicitly formulate a function *f()* that maps ANN activations into predicted neural recordings. In this equation, Act_i refers to the activations of layer i in the ANN and NR_{pred} refers to the predicted neural responses.

$$\boldsymbol{NR}_{pred} = f\left(\mathrm{U}_{i=1}^{L} Act_i\right) \tag{1}$$

$$\boldsymbol{NRP\text{-}error} = \delta\left(NR_{pred}, NR_{measured}\right) \tag{2}$$

The inputs to the function *f()* are the collection of activations from all or a subset of the layers of the ANN. Next, the predicted neural responses are compared to the measured neural recordings using a distance metric δ such as mean absolute error, to quantify neural response prediction error (*NRP-error*), as illustrated in Eq. 2. The NRP-error may be calculated separately for different brain sub-regions (e.g., V1, V4, and IT of the visual cortex) and then averaged to compute the overall NRP-error for the ANN. While there exists a wide range of possibilities for function *f()*, based on the fact that neural networks are universal function approximators, we propose to use a neural network to map from ANN activations to predicted neural recordings.

Neural Response Prediction Method

Our work proposes a new method for quantifying the neural predictivity of an ANN that is based on the overall approach proposed in section "Quantifying Brain Similarity Through Neural Predictivity." The first key idea we propose is to explicitly map ANN activations into predicted neural recordings. A non-linear function is used for this mapping in order to overcome the limitations of previous work (Schrimpf et al., 2018). A second key idea is to use a neural network to approximate this non-linear mapping from ANN activations to predicted neural responses.

Figures 1A,B present the proposed method to quantify the neural predictivity for a given ANN, and given a set of neural recordings. The method consists of the following steps:

Add NRP Network to Decode ANN Activations

The NRP network is an auxiliary structure that decodes the ANN's activations into neural response predictions that can be directly compared to neural recordings in order to compute brain similarity. The structure of the NRP network is detailed in **Figure 1B**. First, activations (layer outputs) from selected layers of the ANN are passed through a layer of neurons that we call NRP-L1. Thus, the size of the input to the NRP network is defined by the number of activations in the chosen layers from the original ANN. The layer NRP-L1 has locally dense connectivity, i.e., the activations from each layer of the ANN are processed separately. This decision was made in order to keep the number of parameters in NRP-L1 and the overall NRP network small. We then concatenate the outputs of NRP-L1 and pass them through a dense layer (NRP-L2). To enable the NRP network to model non-linear relationships, we add ReLU layers at the end of NRP-L1 and NRP-L2. The final layer in the NRP network (NRP-out) produces the predicted neural recordings. Therefore, the number of outputs of the NRP-out layer is set to be equal to the number of neural recording sites for which data is available. We evaluated the use of additional layers in the NRP network, but our experiments suggest that they do not provide improved accuracy. Overall, the NRP network forms a regression network that maps ANN activations into predicted neural responses, specifically the firing rates of the neurons at the recording sites.

Train the NRP Network

The composite network (the original ANN with added NRP layers) is trained while locking down the original ANN's weights. The training data for this composite network consists of stimuli (images) along with corresponding neural recordings from the visual cortex when the primate was presented with these stimuli. The loss function for this training is the mean squared error between the actual and predicted neural recordings. Standard gradient-based optimizers are used for this step [in our experiments, the Adam optimizer (Kingma and Ba, 2014) was found to give the best results]. A held-out set of data is used to validate the NRP network.

Network Architecture Search for the NRP Network

A key challenge faced by the proposed method arises from the limited number of neural recordings, which translates to limited training data for the NRP network. Although it is reasonable to expect this limitation to be gradually relaxed as additional experiments are performed, it is nevertheless one that must be considered in our effort. Thus, it becomes extremely important to determine an optimized configuration for the NRP network so that it has sufficient modeling capacity to predict the neural recordings, but can also be trained with the limited training data available. We address this challenge by performing a network architecture search (Elsken et al., 2019) on the NRP network. Specifically, we performed a grid search on the following hyperparameters for the NRP network: (i) ANN layers used as input to the NRP network, (ii) sizes of the NRP network layers (except NRP-out, whose outputs must match the number of neural recording sites), and (iii) learning rate.

We would like to underscore that the NRP network is not simply a part of the original ANN (e.g., more layers added to it). Instead, it should be viewed as a decoder that maps from ANN activations to a representation that can be directly compared with neural recordings. This overcomes the limitation of previous methods in scenarios where ANN representations and brain representations are not linearly related, and thus correlation metrics that assume a linear relationship are not able to accurately quantify the similarity.

Experimental Setup

The proposed method to compute the neural predictivity of an ANN was implemented using the Tensorflow

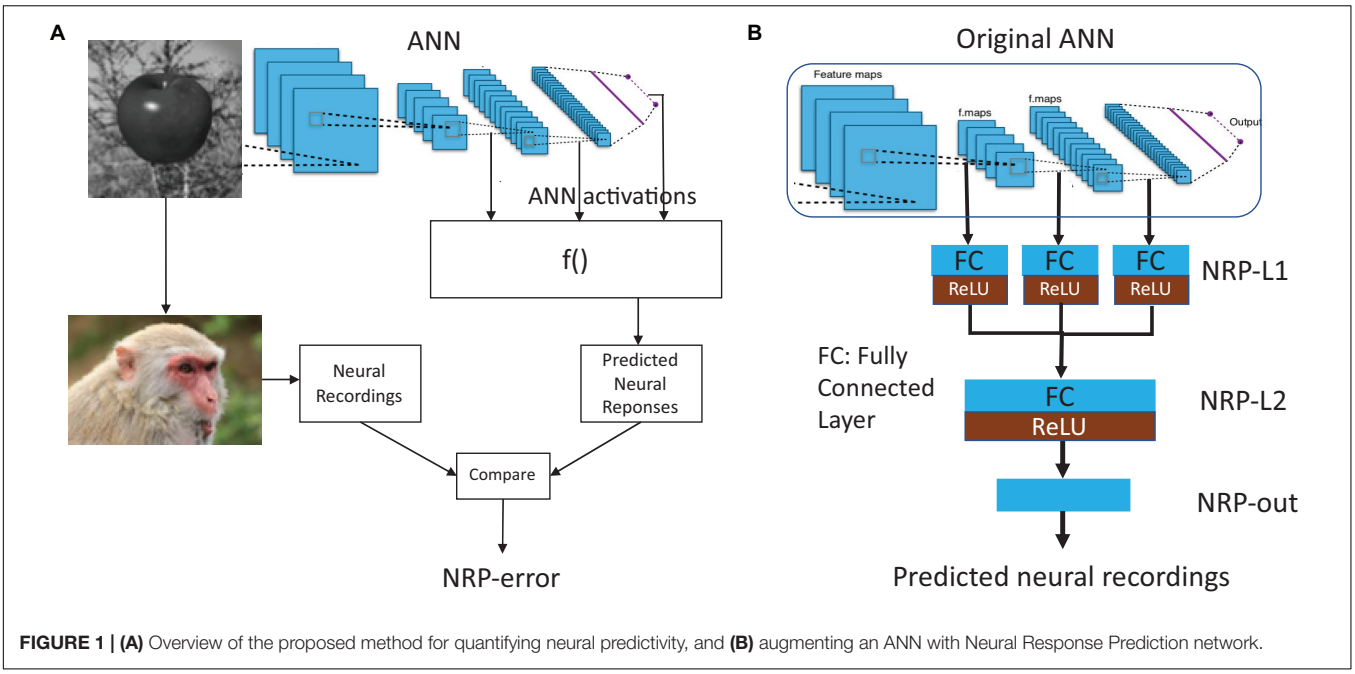

FIGURE 1 | (A) Overview of the proposed method for quantifying neural predictivity, and **(B)** augmenting an ANN with Neural Response Prediction network.

(Tensorflow, 2015) machine learning framework. NRP-errors were calculated for 8 popular image recognition ANNs that have been proposed in recent years for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015). The characteristics of these networks are described in **Table 1**.

The dataset used to train the NRP network and compute NRP-error consists of recordings from 168 neurons in the IT sub-region and 88 neurons in the V4 sub-region of the primate visual cortex (Schrimpf et al., 2018). These responses were measured when visual stimuli (3,200 images) were presented to the primates (*Rhesus macaques*) for 100 ms each immediately before these measurements were made (Schrimpf et al., 2018). Specifically, these neural recordings consist of the average neuronal firing rate for each neuron between 70 and 170 ms after the image was presented. Neuronal firing rates were normalized to the firing rates resulting from a blank gray stimulus. Note that the proposed method is generic and can be applied to recordings from additional sites or brain regions as such recordings become available. The NRP network for each ANN takes as input selected layer activations from the ANN, and produces as output predicted firing rates for each of the recording sites. The NRP-error is the mean absolute error between the predicted firing rates and neural measurements.

NRP-errors were calculated separately for the V4 and IT regions of the visual cortex. In addition to the non-linear model used to generate predicted neural recordings for the calculation of NRP-error, we also implemented a linear regression model to predict neural recordings as a representative of previous efforts.

TABLE 1 | Accuracies and NRP-errors of image recognition ANNs.

Network	Parameters	Top-5 accuracy	Top-1 accuracy	NRP-error
MobileNet	4,253,864	0.895	0.704	0.237
MobileNetV2	3,538,984	0.901	0.713	0.249
NASNetMobile	5,326,716	0.919	0.744	0.247
ResNet50	25,636,712	0.921	0.749	0.290
Xception	22,910,480	0.945	0.79	0.257
DenseNetl21	8,062,504	0.923	0.75	0.249
DenseNetl69	14,307,880	0.932	0.762	0.249
AlexNet	60,954,656	0.803	0.57	0.226

RESULTS

In this section, we discuss the results of implementing the proposed method to quantify neural predictivity of ANNs.

Table 1 presents the NRP-errors of eight different ImageNet classification ANNs. These NRP-errors were computed as the averages of the errors on the V4 and IT regions. As can be seen from the table, some of the more recent ANNs such as ResNet50 (NRP-error of 0.290) are associated with NRP-errors that are higher than older networks such as AlexNet (NRP-errors of 0.226). In fact, AlexNet achieved the lowest NRP-error, while having the lowest Top-1 accuracy, among all networks evaluated. In other words, improvements in application performance (Top-1 accuracy) have not been accompanied by increases in neural predictivity. Another observation is that deeper networks do not necessarily lead to higher neural predictivity. For example, comparing DesneNet121 and DenseNet169, we can see that the additional layers improve Top-1 accuracy but not the neural predictivity. This overall trend, illustrated in **Figure 2A**, is consistent with observations from recent efforts on quantifying brain similarity (Schrimpf et al., 2018). This is perhaps because, deeper ANNs have enabled improvements in accuracy, but have done so by adopting internal representations that are beyond and less like those used in biological systems.

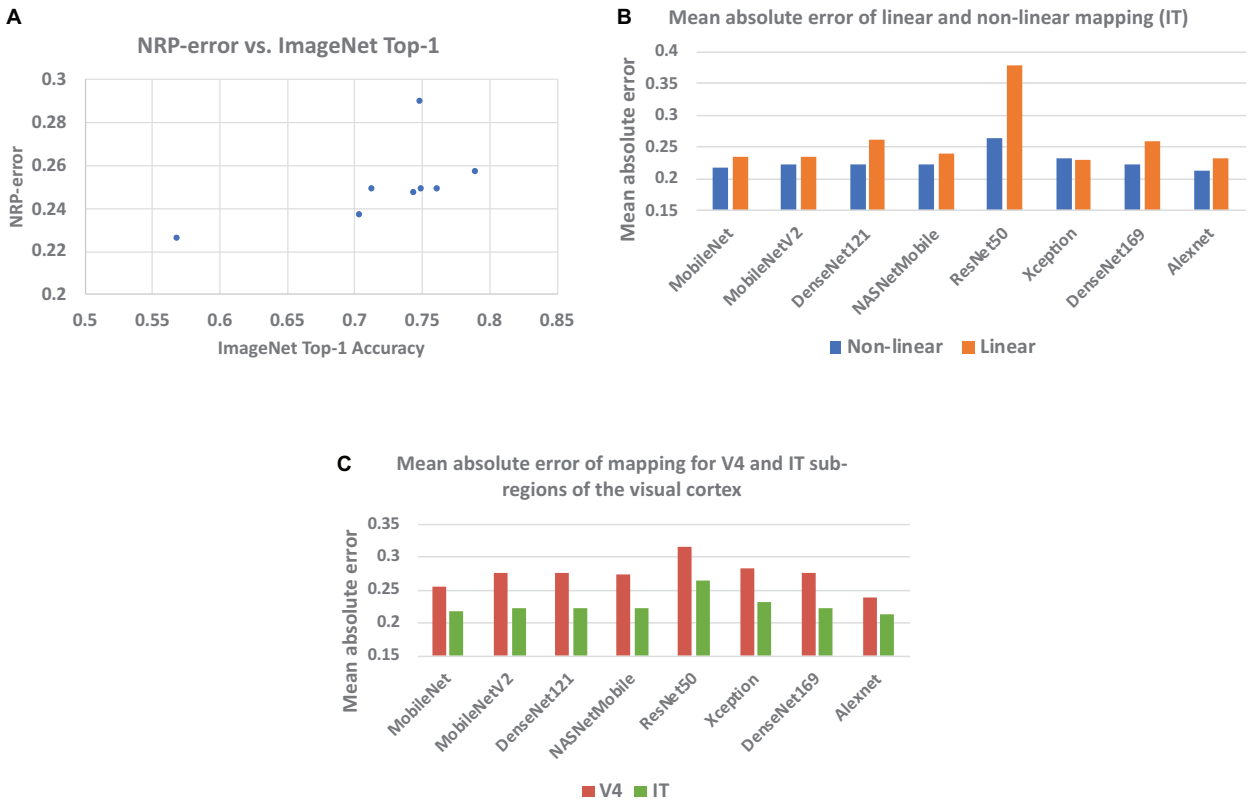

FIGURE 2 | (A) NRP-error vs. Top-1 accuracy for image recognition ANNs listed in **Table 1**, **(B)** mean absolute prediction error of proposed (non-linear) and baseline (linear) methods, and **(C)** prediction error for V4 and IT regions.

Necessity of Non-linear Mapping Function

A key feature of our work is the use of a non-linear mapping function (approximated by a neural network) to map ANN activations to predicted neural recordings in the calculation of NRP-error. This is in contrast to prior efforts, which use the Pearson correlation coefficient, effectively assuming a linear relationship between ANN activations and neural responses. In order to demonstrate the necessity of a non-linear mapping function, we also implemented a linear regression model to predict neural recordings from ANN layer activations. **Figure 2B** compares the mean absolute errors obtained from the proposed method as well as the linear regression model for the IT region. As can be seen from **Figure 2B**, our results show that a non-linear mapping function from ANN activations to predicted neural recordings significantly decreases the error of neural prediction and can hence be considered a superior predictor of an ANN's neural predictivity. The results for V4 also lead to the same conclusion. For example, in the case of ResNet-50, the mean absolute error of the linear and non-linear models are 0.379 and 0.265, respectively. This is explained by the facts that ANN layers are non-linear transformations and there is no layer-to-layer correspondence between most ANN and brain layers, making a non-linear function more suitable to model the mapping between ANN activations and neural recordings.

To further establish the inability of a linear model to capture the relationship between ANN activations and neural recordings, we computed the R^2 values of the linear regression models. We found the R^2 values to be less than 0.2 in all cases, which indicates that a linear model is unable to capture the relationship between ANN activations and neural recordings.

NRP-Errors for V4 and IT Sub-Regions

In order to compare the neural predictivities for the V4 and IT sub-regions of the visual cortex, NRP-errors were computed separately for both sub-regions. From the results, we observe that ANN activations predict IT neural recordings with a higher accuracy than V4 neural recordings (**Figure 2C**). This suggests that ANNs use representations that have a higher level of similarity with later visual cortex sub-regions (such as IT).

Relationship Between Neural Predictivity and Layer Sizes of NRP Network

To overcome the small amount of training data (neural recordings) available for the NRP network, a suitable configuration must be determined so that it has sufficient modeling capacity to predict the neural recordings but can also be trained with the limited training data available. In order to address this, we performed a network architecture search on the NRP network by varying the sizes and number of intermediate

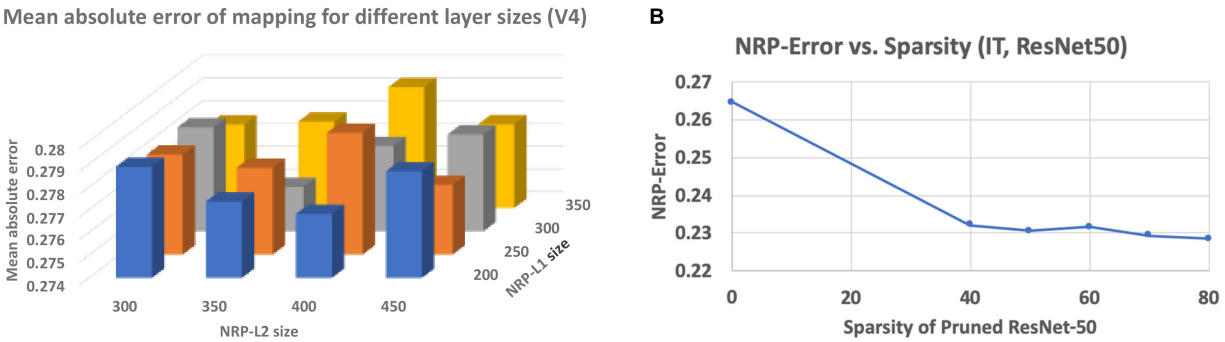

FIGURE 3 | (A) Impact of NRP network layer sizes on prediction error, **(B)** impact of network pruning on NRP-error.

layers to find a suitable configuration. Representative results obtained for the MobileNet ANN are presented in **Figure 3A**. We found that using two intermediate layers in the NRP network before the output layer is sufficient to model the mapping from ANN activations to predicted neural recordings. We also found that there is a sweet-spot of layer sizes for the NRP network that minimizes the average mean absolute error of mapping across all networks.

Impact of Network Pruning on Neural Predictivity

Finally, we investigate the impact of a popular ANN optimization technique, namely network pruning, on neural predictivity. We consider the ResNet50 ANN and applied state-of-the-art pruning algorithms (Zmora et al., 2019) to derive pruned models with varying levels of sparsity. We define sparsity as the percentage of weights that are zero-valued. We generated pruned models of the ResNet50 ANN with sparsity varying from 0 to 80%. We then applied the proposed method to the pruned models to compute the corresponding NRP-errors, and the results are presented in **Figure 3B**. The results suggest that pruning leads to a clear decrease in NRP-error, indicating a positive effect on neural predictivity. We believe this is due to the fact that pruning removes "extraneous" information from the ANN, making it easier to map its activations to the neural recordings.

DISCUSSION

Despite the rapid advances made in the field of deep learning over the past decade, biological brains still have much to teach us in the quest to build more energy-efficient and robust artificial intelligence. A key step toward drawing inspiration from biological brains is to quantify the similarity between them and their artificial counterparts. Our work takes the approach of quantifying similarity through neural predictivity, or the ability to predict neural responses from a biological brain given the internal information of ANNs. Since this is the goal of our work, we discuss closely related efforts and place our own effort in their context. We also discuss possible future directions, both in terms of improving our work and its applications.

Related Work

A recent effort that quantifies neural predictivity is Brain-Score (Schrimpf et al., 2018). Brain-Score specifically focuses on evaluating ANNs that perform core object recognition tasks, and provides a quantitative framework to compare image classification ANNs with measurements from the visual cortex of primates (firing rates for specific neurons when the primate is presented with the stimulus). It consists of a behavior sub-score and neural predictivity sub-scores for various regions of the visual cortex (V1, V2, V4, and IT). The behavior sub-score quantifies how similar the ANN's predictions are to those made by the primate when both are presented with the same stimulus. The neural predictivity sub-scores capture how well the ANN's activations correlate to the neural recordings from each region of the visual cortex. These sub-scores are computed as the Pearson correlation coefficient between ANN layer outputs and neural firing rates for that region.

Through the use of the Pearson correlation coefficient, Brain-Score implicitly assumes a linear relationship between ANN activations and neural firing rates. However, since ANN layers are non-linear transformations, there is no evidence to support this assumption. Moreover, there is no layer-to-layer correspondence between most ANN and brain layers, making the likelihood of a linear relationship even less likely.

Our work extends the state-of-the-art through two key ideas. First, it advocates the use of an explicit (non-linear) mapping function to predict neural responses from ANN activations in order to quantify neural predictivity. A second key idea is the use of a neural network (known to be a universal function approximator) to approximate the mapping function itself. Our experiments clearly support the merit of these proposals by demonstrating an improved ability to predict neural responses.

Future Work

One possible direction to build upon our effort would be to collect and incorporate additional neural recordings into the dataset used. A dataset with additional recording locations and more input images would allow us to train larger (and potentially more accurate) NRP networks without the risk of over-fitting. Since internal representations are greatly influenced by training, it would also be interesting to study whether networks trained

with bio-plausible learning rules (e.g., STDP) yield higher neural predictivity than ANNs trained with gradient-descent. Finally, building upon a recent result that using brain-like representations in the early layers of an ANN can lead to higher robustness, it would be interesting to study whether there is a relationship between an ANN's neural predictivity and its robustness to noise and adversarial perturbations.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

AA implemented the methods, performed the experiments, and analyzed the data. SS and KR provided mentorship and inputs into all aspects of the research. SS assisted with experiments. All authors contributed to writing the manuscript.

FUNDING

This work was supported in part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Martin Schrimpf and James DiCarlo from the Department of Brain and Cognitive Sciences at the Massachusetts Institute of Technology for providing them with the neural recordings used in this work, and for their valuable suggestions.

REFERENCES

AI and Compute, (2018). Available online at: https://openai.com/blog/ai-and-compute/ (accessed August 15, 2020).

Bi, G. Q., and Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* 18, 10464–10472.

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*. Hoboken, NJ: Wiley-Interscience.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., and DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *bioRxiv* doi: 10.1101/2020.06.16.154542

Deepmind AlphaGo, (2017). Available online at: https://deepmind.com/research/case-studies/alphago-the-story-so-far (accessed August 15, 2020).

Elsken, T., Metzen, J. H., and Hutter, F. (2019). Neural architecture search: a survey. *J. Mach. Learn. Res.* 20, 1–21.

Fukushima, K. (1980). Neocognitron, a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernet.* 36, 193–202.

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258.

Hubel, D. H., and Weisel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

IBM, (2011). *A Computer Called Watson*. Available online at: https://www.ibm.com/ibm/history/ibm100/us/en/icons/watson/ (accessed August 15, 2020).

Kingma, D. P., and Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. Available online at: https://arxiv.org/abs/1412.6980 (accessed August 15, 2020).

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324.

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Netw.* 10, 1659–1671. doi: 10.1016/S0893-6080(97)00011-7

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., and Ma, S. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comp. Vis.* 115, 211–252.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., et al. (2020). Artificial neural networks accurately predict language processing in the brain. *bioRxiv* doi: 10.1101/2020.06.26.174482

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., et al. (2018). Brain-score: which artificial neural network for object recognition is most brain-like? *bioRxiv* doi: 10.1101/407007

Sharmin, S., Panda, P., Sarwar, S. S., Lee, C., Ponghiran, W., and Roy, K. (2019). A comprehensive analysis on adversarial robustness of spiking neural networks. *Int. Joint Conf. Neural Netw.* 63, 3493–3500.

Tensorflow, (2015). An end-to-end open source machine learning platform. Available online at: https://www.tensorflow.org/ (accessed August 15, 2020).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, New York, NY.

Venkataramani, S., Roy, K., and Raghunathan, A. (2016). "Efficient embedded learning for IoT devices," in *Proceedings of the Asia and South Pacific Design Automation Conference (ASP-DAC)*, Macao, 308–311.

Zmora, N., Jacob, G., Zlotnik, L., Elharar, B., and Novik, G. (2019). Neural network distiller: a python package for DNN compression research. *arXiv*. Available online at: https://arxiv.org/abs/1910.12232 (accessed August 15, 2020).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.