2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Search of Highly Selective Cells in Convolutional Layers with Hebbian Learning

Fernando Aguilar-Canto
Computational Cognitive Sciences Laboratory-CIC
Instituto Politécnico Nacional
Mexico City, Mexico
pherjev@gmail.com

Hiram Calvo
Computational Cognitive Sciences Laboratory-CIC
Instituto Politécnico Nacional
Mexico City, Mexico
hcalvo@cic.ipn.mx

***Abstract*—Deep Convolutional Neural Networks (ConvNets) have demonstrated successful implementations in various vision tasks, including image classification, segmentation, and image captioning. Despite their achievements, concerns persist regarding the explainability of these models, often referred to as black-box classifiers. While some interpretability papers suggest the existence of object detectors in ConvNets, others refute this notion. In this paper, we address the challenge of identifying such neurons by utilizing Hebbian learning to discover the most associated neurons for a given stimulus. Our method focuses on the VGG19 and ResNet50 networks with the Dogs-vs-Cats dataset. During experimentation, we found that the most associated hidden neurons to the labels are not object detectors. Instead, they seem to encode relevant aspects of the category. By shedding light on these findings, we aim to improve the understanding and interpretability of deep ConvNets for future advancements in the field of computer vision.**

***Index Terms*—Interpretability, Convolutional Neural Networks, Hidden Semantics approaches, ResNet50, VGG19**

I. Introduction

Deep Neural Networks (DNNs) are the state-of-the-art solution in several tasks such as Image Classification [1], [2], Image Captioning [3], Machine Translation [4], Natural Language Understanding [5], among others. Nevertheless, Deep Networks are usually labeled as *black-boxes* [6], in the sense that they are less explainable solutions than other approaches, which is critical in several areas such as Medical Imaging [7], [8], Self-Driving Cars [9], Legal affairs [10], among others.

According to [11], *interpretability* and *explainability* both refers to the ability to provide understandable explanations in human terms. Different approaches have attempted to reduce the mentioned lack of interpretability/explainability, such as providing rules as explanations [12], [13], explaining hidden semantics (see Related work), using attributes as explanation [14]–[16], or by showing examples [17], [18]. As a consequence, there is diversity in the proposed interpretability methods, each of them aims to reveal different aspects of the network, and evaluation, even if it exists (see, for instance, [19]), it mainly remains in different criteria.

If DNNs are black boxes, the biological counterparts might also be considered like that. Neuroscience, however, has revealed many aspects of the inner operations of individual neurons, thanks to the development of micro-electrodes. Most experiments have been conducted in the mammal visual cortex, measuring the firing-rate response of individual neurons given a set of stimuli [20]–[22].

Similarities between DNNs, in particular, Convolutional Neural Networks (ConvNets), have been highlighted by some authors [23], although it is still debated in Computational Neuroscience [24]. One possible similarity relies on the presence of highly selective neurons found in the Infratemporal Cortex and Middle Temporal Lobe and the emergence of *object-detectors* in deep layers [25]–[27].

Nevertheless, the presence of highly selective neurons has led to the conclusion of the presence of *grandmother-cells* [28], which have been challenged by neuroscientists [29]. In the case of ConvNets, some authors have debated the actual presence of the so-called object detectors, since they have a high rate of false positives [30].

This paper addresses the problem of the existence of highly selective neurons in Deep ConvNets, by applying the stimuli-response framework previously used in Neuroscience, to verify if similar results emerge. We propose a framework to search the referred highly selective neurons by using Hebbian Learning if such neurons exist. Hebbian Learning is used to find relevant associations in the data [31]. In addition, this paper aims to evaluate critically the existence of real-object detectors using different metrics to see whether such neurons can be considered classifiers, in well-defined classification problems. In this sense, we would like to verify if Transfer Learning operates at a single neuron level.

This paper is structured as follows: Section II briefly introduces related works in explainability and discusses pertinent discussions found in the literature. Section III presents the algorithms and methods for 1) detecting the most associated units, and 2) analyzing and evaluating such units. Section IV showcases relevant quantitative and qualitative results, while Section V summarizes our findings.

II. Related work

The search for object detectors in hidden layers corresponds to the Hidden Semantics as an Explanation approach in interpretability. The dominant method in this approach is Activation Maximization [32], which involves searching in a large domain space, but the resulting images are not always interpretable in human terms. Given neuron activity n_{ij} for neuron j in layer

i, and network parameters θ, Activation Maximization aims to find:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}}(h_{ij}(\mathbf{x}, \theta) - \lambda\Omega(\mathbf{x})), \quad (1)$$

where Ω is an optional regularizer, and λ controls the importance of the regularizer. Activation Maximization has been used in ConvNets [2], [33].

Another approach for revealing hidden semantics involves searching for the emergence of object detectors in deep convolutional layers, if such neurons exist. This stimuli-based approach was implemented by [34] and found some units with high precision in single object recognition. The neural networks in this case were trained in Scene Recognition, leading to the emergence of object detectors during the training process.

A. Literature discussion about the existence of selective cells

The existence of object detection by a single hidden neuron was supported by stimuli-based and Activation Maximization approaches [35], [36]. However, Activation Maximization has faced recent criticism [37]. Some authors challenged the idea of object detectors in hidden layers. [38] found that ConvNets can perform well without relying on single object detectors and suggested using regularization techniques for better generalization. [30] questioned the existence of such object detectors in classical ConvNets due to the lack of highly selective units with high hit-rates or low false-alarm rates. Considering these concerns, Fong *et al.* [39] proposed investigating vectorial representations instead of single neurons.

III. Methodology

Broadly speaking, this proposal aims to identify hidden neurons in convolutional layers that are highly associated with specific stimuli classes and not associated with other classes. Once these neurons are identified through Hebbian Learning (see [31]), their performance in predefined classification tasks can be evaluated. The process involves two steps: first, learning associations between neurons and stimuli classes, and then verifying their selectivity. This complete procedure combines both Hebbian Learning and symbolic techniques to achieve its objectives.

A. First step: Hebbian learning

The first step of this methodology involves selecting a layer c from a Deep Network to study its output. In ConvNets, the output of the convolutional layer with p neurons is represented by a tensor $\mathbf{T}_c \in \mathbb{R}^{\ell\times\ell\times p}$. To reduce the feature tensor to a single vector $\mathbf{u}$, the maximum values of each $\ell \times \ell$ image are taken, resulting in:

$$\mathbf{u}[k] = \max_{i,j} \mathbf{T}_c[i, j, k]. \quad (2)$$

Next, for each training example with index e (from a training dataset $(\mathbf{x}_e, y_e)_{e=1}^{n}$), the index $k_e^* = \arg\max_k \mathbf{u}[k]$ is selected. Let q represent a specific class. The set K_q is defined as:

$$K_q = \{k_e^* \mid y_e = q\}, \quad (3)$$

which contains the indexes of the given class that maximize $\mathbf{u}$. To obtain the indexes that only maximize the class q and not other classes, the set P_q is defined as:

$$P_q = K_q - \bigcup_{q' \neq q} K_{q'}. \quad (4)$$

Once vectors $\mathbf{u}$ and $\mathbf{v}$ (one-hot encoding of labels) are defined, a weight matrix $\mathbf{H}$ is trained using Hebbian learning:

$$\mathbf{H}_0 = \mathbf{0} \quad (5)$$

$$\mathbf{H}_{e+1} = \mathbf{H}_e + \mathbf{u}_e\mathbf{v}_e^{\mathrm{T}} \quad (6)$$

To retrieve the indexes that only maximize the given class, a mask m_q is defined as a vector such that $m_q[\iota] = 1$ for all $\iota \in P_q$ and zero elsewhere. The operation $\mathbf{v}^{\mathrm{T}}\mathbf{H}$ is used to find the most associated index, denoted by κ_q, for class q. However, to mitigate potential issues where κ_q could be associated with other stimuli, the following operation is performed:

$$\kappa_q = \arg\max_k \left(\mathbf{v}_q^{\mathrm{T}}\mathbf{H} \odot m_q\right)[k], \quad (7)$$

where $\mathbf{v}_q$ represents the one-hot encoding for class q.

B. Second step: analysis of the most associated units

κ_q represents the index of a highly associated unit in the layer with class q, excluding indexes related to other classes. Evaluation of these units' classification ability requires computing true and false positives and negatives, creating a binary confusion matrix.

Deep Networks often use Rectified Linear Units (ReLU) as activation functions, posing challenges for direct classification. Thresholds can be applied to turn selective neurons into classifiers. The threshold can be optimized using the validation set, aiming to minimize false positives and maximize precision. An algorithm is used to optimize θ:

$$\theta_{e+1} = \begin{cases} \theta_e - \alpha_1 & \text{if false positive} \\ \theta_e + \alpha_2 & \text{if false negative} \end{cases} \quad (8)$$

where $\alpha_1 = 10$, $\alpha_2 = 1$.

C. Evaluation

The following Deep ConvNets will be used for evaluation:

1) VGG19 [2] (selected layer: `block5_conv4`).
2) ResNet50 [40] (selected layer: `conv5_block3_out`).

We selected the final convolutional layers, since we are interested on highly complex patterns. Tests will be performed on the Cats-vs-Dogs dataset [41] to evaluate individual neurons. The metrics to be reported are the following: precision, accuracy, recall, specificity, and Class-Conditional Mean Activation Selectivity (CCMAS) [38].

IV. Experimental results

A. VGG19

The training of VGG19 resulted in a relatively low accuracy (0.8293) compared to the findings in [31]. Logarithm regularization seems relevant. Table I presents the main results of VGG19 network evaluation.

TABLE I: Summary of quantitative results of the most associated units of the VGG19 to the studied classes.

Metric	**B5C4[498]**	**B5C4[314]**
Class	Cat	Dog
Precision	0.7476	0.8987
Accuracy	0.5319	0.6109
Recall	0.2068	0.6389
Specificity	0.9906	0.9275
CCMAS	0.1374	0.7648

1) Neuron `block5_conv4[498]` (B5C4[498]): Unit `block5_conv4`[498] is strongly associated with the label "cat." It shows a high proportion of false negatives and a relatively low proportion of true positives (see Figure 1), and a low CCMAS. Violin plots in Figure 2 display the output distribution for both classes, indicating that a significant fraction of dogs elicit a strong response for the detected feature. Qualitative analysis suggests that this neuron is selective to "triangular ears", evident in cat photographs (Figure 4). Similarly, in dog pictures (Figure 5), the unit is selective to triangular ears of some dogs, explaining the quantitative results. Activation maximization (see Figure 3) supports this observation, revealing that the image maximizing the neuron's response consists of a set of triangles in various directions.

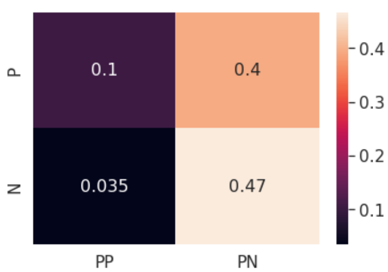

Fig. 1: Scaled confusion matrix of unit `block5_conv4`[498].

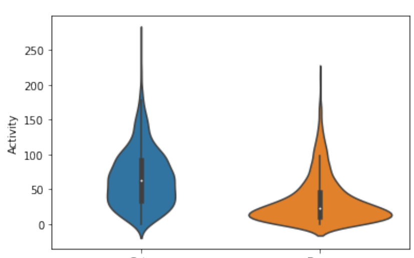

Fig. 2: Violin plot of the distributions of activations for both classes in `block5_conv4`[498].

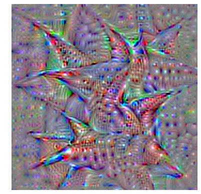

Fig. 3: Activation Maximization of the unit `block5_conv4`[498] with shapes of triangles.

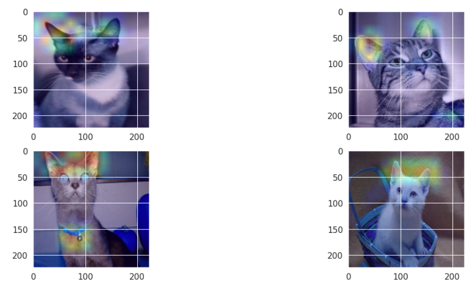

Fig. 4: Examples of heatmaps of the output of the convolution of the unit B5C4[498] in cats.

2) Neuron `block5_conv4[314]` (B5C4[314]): Unit 314 shows potential as a "dog detector," with a measured precision of 0.987. The confusion matrix (Figure 6) and violin plot (Figure 7) present a more promising performance compared to the previous neuron. The qualitative analysis indicates that the neuron is selective to cats (Figure 9) and also recognizes some dog faces (Figure 10). Activation maximization in Figure 8 appears like dog faces, further supporting its selectivity to dogs.

B. ResNet50

The first step (Hebbian learning) resulted in an accuracy of 0.9656, slightly lower than the results in [31] due to the absence of regularization. Key quantitative results are summarized in Table II.

1) Neuron `conv5_block3_out[1353]` (C5B50[1353]): Unit `conv5_block3_out`[1353] is identified as the primary candidate for a "cat detector." The optimization procedure yielded a precision of 0.9874 (see Table II), but the accuracy is impacted by a relatively large number of false negatives.

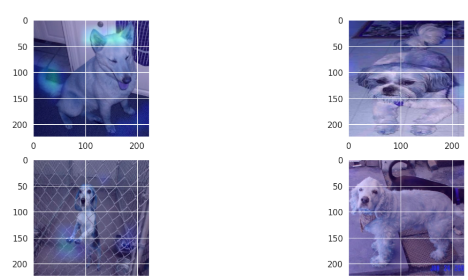

Fig. 5: Examples of heatmaps of the output of the convolution of the unit B5C4[498] in dogs.

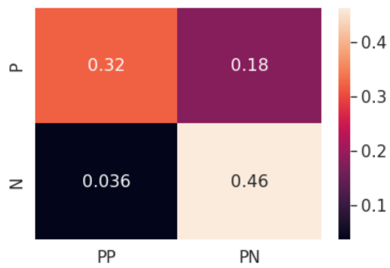

Fig. 6: Scaled confusion matrix of unit `block5_conv4`[314].

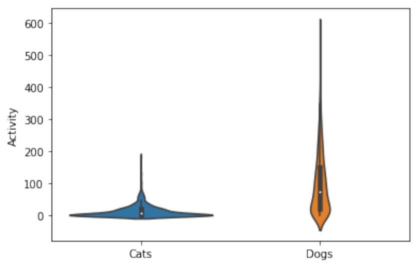

Fig. 7: Violin plot of the distributions of activations for both classes in `block5_conv4`[314].

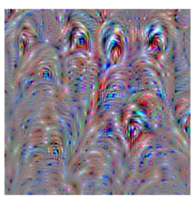

Fig. 8: Activation Maximization of the unit `block5_conv4`[314] with dog-like patterns.

CCMAS is not low, although it might not be ideal. However, the violin plot presents a more favorable outlook compared to VGG19 (see Figure 11). Nonetheless, the resulting image generated by Activation Maximization is less interpretable (Figure 12). Qualitative analysis in Figures 13 and 14 indicates that the neuron focuses on the texture of the cat, particularly in the face area.

2) Neuron `conv5_block3_out[742]` *(C5B50[742]):* Neuron 742 acts as a weak candidate for a "dog detector" with

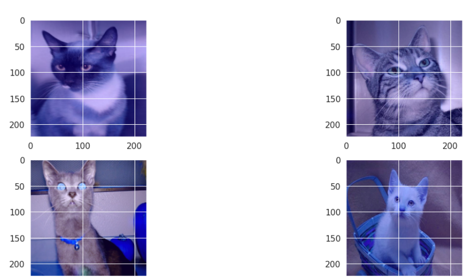

Fig. 9: Examples of heatmaps of the output of the convolution of the unit B4C5[314] in cats.

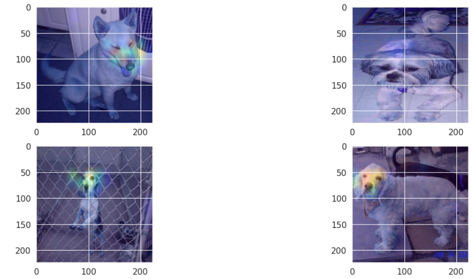

Fig. 10: Examples of heatmaps of the output of the convolution of the unit B5C4[314] in dogs.

TABLE II: Summary of quantitative results of the most associated units of the ResNet50 to the studied classes.

Metric	**C5B5O[1353]**	**C5B5O[742]**
Class	Cat	Dog
Precision	0.9874	0.8631
Accuracy	0.6332	0.5821
Recall	0.7310	0.4648
Specificity	0.9906	0.9259
CCMAS	0.6075	0.4355

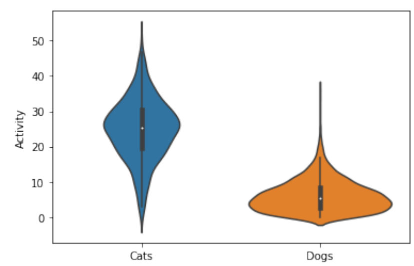

Fig. 11: Violin plot of the distributions of activations for both classes in `conv5_block3_out`[1353].

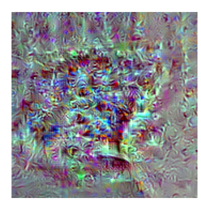

Fig. 12: Activation Maximization of the unit `conv5_block3_out`[1353]. Interpretation of the resulting image is not clear.

a precision of 0.8631. The metrics indicate that this unit is a poor "dog classifier". Increasing the threshold could enhance precision but also raise the number of false negatives, as we observe a relatively low CCMAS. The output distribution is shown in Figure 15. Similar to neuron 1353, the result of Activation Maximization is less interpretable (Figure 16). Qualitative analysis in Tables 17 and 18 helps to understand why this unit performs poorly. All the cats in Table 17 have relatively low output values, while Table 18 indicates that only one out of four dogs produces a significant output. However,

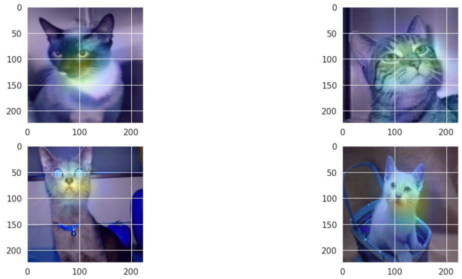

Fig. 13: Examples of heatmaps of the output of the convolution of the unit C5B5O[1353] in cats.

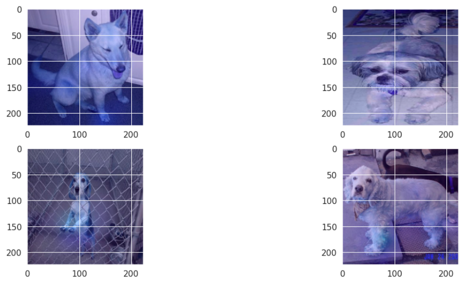

Fig. 14: Examples of heatmaps of the output of the convolution of the unit C5B5O[1353] in dogs.

the activation for this example is considerably high. This suggests that neuron 742 classifies well for a subset of dog images.

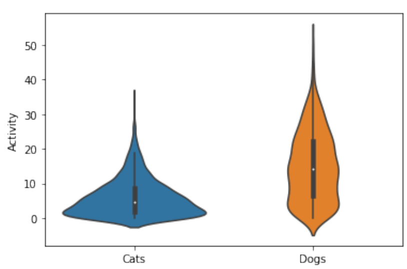

Fig. 15: Violin plot of the distributions of activations for both classes in `conv5_block3_out`[742]

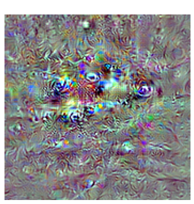

Fig. 16: Activation Maximization of the unit `conv5_block3_out`[742]. Interpretation here is left to the reader.

V. Conclusions

This research aims to use the associations yielded by Hebbian learning to find systematically highly selective cells in the inner convolutional layers. In this context, Hebbian learning is instrumentally used not to solve a fixed classification problem, but to find the most associated neurons to a given class of stimuli.

At least one neuron (C5B5O[1353]) was found to be a highly selective cell to cat signals, with a precision of 0.9874. Qualitative results show that, indeed, the neuron seems to be selective to cat areas, in particular, to cat's faces, and less activated to dog's faces. Although the accuracy can be improved, it was not the objective to optimize. As a consequence, the accuracy of the unit was found to be low, but this is not considered to be a problem since other cells might be used to take appropriate decisions.

The preliminary results show that individual neurons are indeed selective but to a subset of given classes, instead of being complete object detectors. As previously discussed, the network needs to eliminate the dependency on a single neuron for classification, and the same scenario might hold in Neuroscience. In the case of the dogs-vs-cats dataset, the analysis of neuron 742 of the last convolutional layer indicates that the unit is selective to only a proper subset of the class dog, whereas in the case of cats, the found cell seems to be a better classifier. This situation might be due to the higher diversity of images of dogs.

In this sense, we do not see the problems found in [30], since even when the classification of the unit has a high number of false negatives, such misclassified examples might be classified correctly by using other neurons. Therefore, the results do not support the idea of a one-hot encoding in convolutional layers, but a sparse and localist encoding.

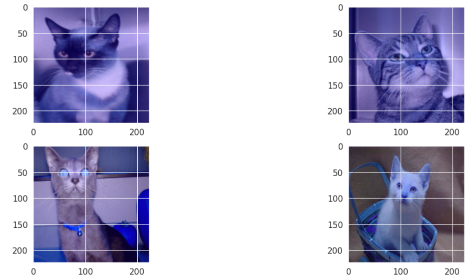

Fig. 17: Examples of heatmaps of the output of the convolution of the unit C5B5O[742] in cats.

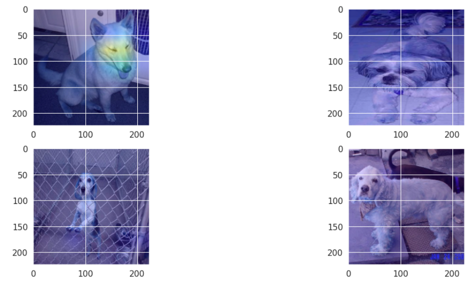

Fig. 18: Examples of heatmaps of the output of the convolution of the unit C5B5O[742] in dogs.

ACKNOWLEDGMENT

The authors wish to thank the support of the Instituto Politécnico Nacional (COFAA, SIP-IPN, Grant SIP 20230140) and the Mexican Government (CONAHCyT, SNI).

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[2] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013, unpublished.

[3] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[5] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," *arXiv preprint arXiv:1909.10351*, 2019, unpublished.

[6] D. Castelvecchi, "Can we open the black box of ai?" *Nature News*, vol. 538, no. 7623, p. 20, 2016.

[7] X. Li, Y. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, L. H. Staib, P. Ventola, and J. S. Duncan, "Braingnn: Interpretable brain graph neural network for fmri analysis," *Medical Image Analysis*, vol. 74, p. 102233, 2021.

[8] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Computers in biology and medicine*, vol. 140, p. 105111, 2022.

[9] J. Kim, A. Rohrbach, Z. Akata, S. Moon, T. Misu, Y.-T. Chen, T. Darrell, and J. Canny, "Toward explainable and advisable model for self-driving cars," *Applied AI Letters*, vol. 2, no. 4, p. e56, 2021.

[10] A. Bibal, M. Lognoul, A. De Streel, and B. Frénay, "Legal requirements on explainability in machine learning," *Artificial Intelligence and Law*, vol. 29, no. 2, pp. 149–169, 2021.

[11] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[13] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," *Advances in neural information processing systems*, vol. 31, 2018.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[15] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[16] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[17] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International conference on machine learning*. PMLR, 2017, pp. 1885–1894.

[18] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[19] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[20] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, p. 574, 1959.

[21] ——, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, p. 106, 1962.

[22] ——, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[23] R. M. Cichy and D. Kaiser, "Deep neural networks as scientific models," *Trends in cognitive sciences*, vol. 23, no. 4, pp. 305–317, 2019.

[24] M. H. Herzog and A. M. Clarke, "Why vision is not both hierarchical and feedforward," *Frontiers in computational neuroscience*, vol. 8, p. 135, 2014.

[25] R. Desimone, T. D. Albright, C. G. Gross, and C. Bruce, "Stimulus-selective properties of inferior temporal neurons in the macaque," *Journal of Neuroscience*, vol. 4, no. 8, pp. 2051–2062, 1984.

[26] R. Desimone, "Face-selective cells in the temporal cortex of monkeys," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 1–8, 1991.

[27] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, "Invariant visual representation by single neurons in the human brain," *Nature*, vol. 435, no. 7045, pp. 1102–1107, 2005.

[28] J. S. Bowers, "What is a grandmother cell? and how would you know if you found one?" *Connection Science*, vol. 23, no. 2, pp. 91–95, 2011.

[29] R. Q. Quiroga, G. Kreiman, C. Koch, and I. Fried, "Sparse but not 'grandmother-cell'coding in the medial temporal lobe," *Trends in cognitive sciences*, vol. 12, no. 3, pp. 87–91, 2008.

[30] E. M. Gale, N. Martin, R. Blything, A. Nguyen, and J. S. Bowers, "Are there any 'object detectors' in the hidden layers of cnns trained to identify objects or scenes?" *Vision Research*, vol. 176, pp. 60–71, 2020.

[31] F. J. Aguilar Canto, "Convolutional Neural Networks with Hebbian-Based Rules in Online Transfer Learning," in *Mexican International Conference on Artificial Intelligence*. Springer, 2020, pp. 35–49.

[32] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[33] F. Wang, H. Liu, and J. Cheng, "Visualizing deep neural network by alternately image blurring and deblurring," *Neural Networks*, vol. 97, pp. 162–172, 2018.

[34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.

[35] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015, unpublished.

[36] A. Nguyen, J. Yosinski, and J. Clune, "Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks," *arXiv preprint arXiv:1602.03616*, 2016, unpublished.

[37] R. Geirhos, R. S. Zimmermann, B. Bilodeau, W. Brendel, and B. Kim, "Don't trust your eyes: on the (un) reliability of feature visualizations," *arXiv preprint arXiv:2306.04719*, 2023, unpublished.

[38] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, and M. Botvinick, "On the importance of single directions for generalization," *arXiv preprint arXiv:1803.06959*, 2018, unpublished.

[39] R. Fong and A. Vedaldi, "Net2Vec: Quantifying and Explaining how Concepts are Encoded by Filters in Deep Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8730–8738.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] J. Elson, J. J. Douceur, J. Howell, and J. Saul, "Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization," in *Proceedings of 14th ACM Conference on Computer and Communications Security (CCS)*. Association for Computing Machinery, Inc., October 2007. [Online]. Available: https://www.microsoft.com/en-us/research/publication/asirra-a-captcha-that-exploits-interest-aligned-manual-image-categorization/

[42] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020, unpublished.