A Critical Analysis of Color Neurons in Vision Models

Fernando Aguilar-Canto, Hiram Calvo, Ricardo Menchaca Méndez
CIC IPN

TL;DR:

We propose three criteria for validating single-unit interpretations and identify five color-selective neurons across three architectures, showing that rigorous unit-level understanding is attainable.

Introduction

Skepticism in modern unit's level interpretability:

1. *Interpretation* of the results yielded by certain methods [3].
2. *Validity* of the methods themselves, usually tied to top activations [4, 1, 8, 9].
3. *Relevance* of the most interpretable units [7].
4. whether the unit level constitutes a *proper* tier of analysis in the first place [6].

To address this skepticism, we propose that any interpretation I_u of a unit u should satisfy three criteria:

1. *Completeness criterion*: The interpretation I_u must hold over a predefined and semantically coherent domain $\mathcal{D}$, not dependent on the explainer nor exclusively on synthetic stimuli.
2. *Falsifiability criterion*: There must exist an experimental procedure that could reveal I_u to be false if it were incorrect.
3. *Relevance criterion*: The unit u must be functionally relevant to the network or layer, even for simple tasks.

Methodology

Studied networks:

1. ResNet50 [5] (`layer1`)
2. ViT-B/16 [2] (`encoder.block.0.mlp[1]` and `encoder.ln`)
3. ViT-L/32 [2] (`encoder.block.0.mlp[1]`)

Datasets:

1. *Spectral dataset* of synthetic stimuli.
2. *Chromatic dataset* of natural images for the Completeness and Falsifiability criteria.
3. Smaller *probing dataset* to identify relevant units.

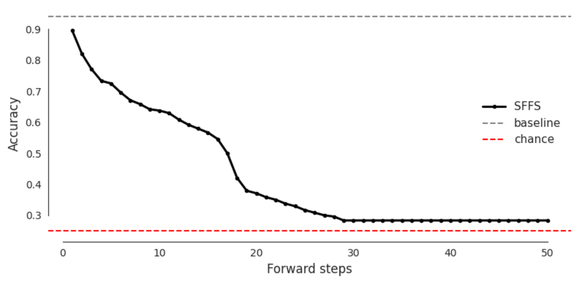
All natural images are drawn from the ILSVRC2012 ImageNet validation set, using the train/val/test splits (35,000 / 5,000 / 10,000) introduced by Oikarinen and Weng [9].

Relevance criterion

We probe the layer's capacity to solve a color classification task using a linear classifier trained on its activations. Candidate units are selected through Sequential Floating Forward Selection (SFFS) [10].

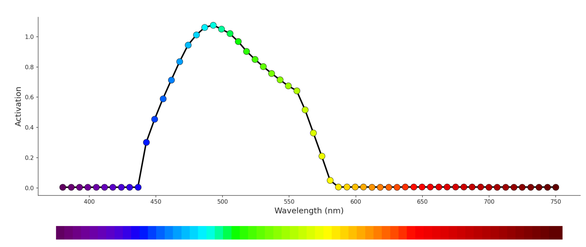
In the image below, we can observe the effect of sequential pruning with SFFS in the `layer1` of the ResNet50. SFFS ablation reduced accuracy from 0.942 to 0.283, identifying 29 relevant units. For conciseness, we analyze in detail the top-ranked unit (38).

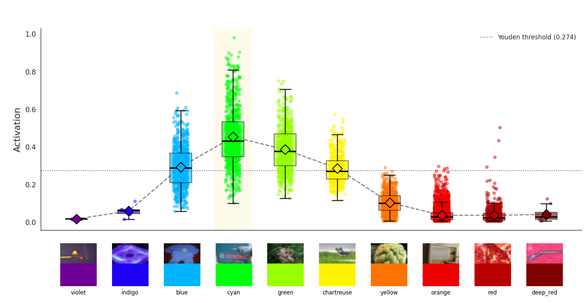
Several other units showed synthetic tuning curves similar to that of unit 38, suggesting that the blue/cyan direction is broadly relevant for the probing task. However, not all SFFS-selected units passed the subsequent criteria: units 177, 24, and 96 failed pairwise Mann–Whitney tests against the remaining color classes, indicating that causal relevance to the probe task does not guarantee chromatic selectivity.



Completeness criterion

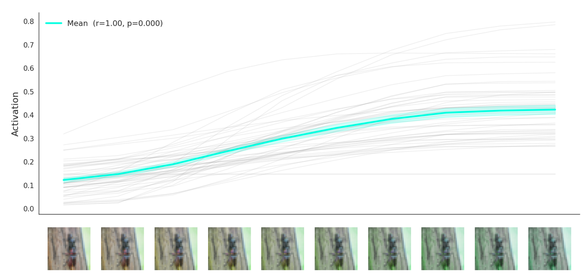
Once the relevant units for the color classification task have been identified in a given layer, we validate the Completeness criterion by constructing tuning curves from both the spectral (first image) and chromatic datasets (second image). The selected unit (38), showed clear selectivity for the cyan band. The Kruskal–Wallis test confirmed significant variation across bands ($H = 1616.46$, $p < 0.001$). Cyan achieved the highest CCMAS (0.639) and d' (2.485), with significantly higher activation than eight of nine alternative bands.

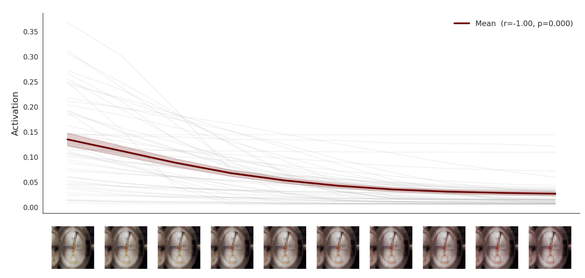




Falsifiability criterion

We draw a random sample of natural images from the spectral dataset and apply a gradual color shift to each image, transforming its pixels toward the hypothesized preferred wavelength λ. The shift is applied in HSV space. In this case (unit 38) we applied shift towards the preferred direction (fist image) producing an increase on the activation while a shift towards unpreferred direction (second image) produce a general decrease on the neural activation.





Discussion

In this paper, we applied the proposed framework to critically analyze color neurons, arguably among the most basic visual features in computer vision. We found units that satisfy all criteria: three excitatory unimodal neurons (green, orange, cyan), one bimodal neuron (blue–red), and one inhibitory neuron (suppressed by blue–cyan). However, we note that the strict 0.75 precision/recall threshold for a 'strong detector', while a useful reference, proved to be overly conservative for the present setting, where perceptual distances between color classes are not uniform. As a main finding, we consider that we have evidence that these units are correctly interpreted. However, the shift experiment generalizes across randomly sampled images from the test set, suggesting that the interpretation may extend beyond monochromatic stimuli.

Two clarifications are in order. First, validating one interpretation does not imply it is the only valid interpretation of the unit; a neuron may encode additional features beyond color. Second, the interpretation scheme presented here is not final and could be refined. Finally, these findings suggest that some neurons behave as functions of their input in ways that can be formalized as testable hypotheses.

Limitations: The color shift intervention may produce out-of-distribution samples. The validation domain $\mathcal{D}$ is restricted to images with a discernible predominant color. Relevance is assessed at the layer level via a color classification probe. Several spectral bands at the extremes of the visible spectrum contain very few samples in the test set.

References

[1] Dilyara Bareeva et al. "Manipulating Feature Visualizations with Gradient Slingshots". In: *arXiv preprint arXiv:2401.06122* (2024).

[2] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[3] Ella M Gale et al. "Are there any 'object detectors' in the hidden layers of CNNs trained to identify objects or scenes?" In: *Vision Research* 176 (2020), pp. 60–71.

[4] Robert Geirhos et al. "Don't trust your eyes: on the (un) reliability of feature visualizations". In: *arXiv preprint arXiv:2306.04719* (2023).

[5] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[6] Jing Huang et al. "Rigorously Assessing Natural Language Explanations of Neurons". In: *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. 2023, pp. 317–331.

[7] Ari S Morcos et al. "On the importance of single directions for generalization". In: *arXiv preprint arXiv:1803.06959* (2018).

[8] Geraldin Nanfack et al. "Adversarial Attacks on the Interpretation of Neuron Activation Maximization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 5. 2024, pp. 4315–4324.

[9] Tuomas Oikarinen and Tsui-Wei Weng. "Linear Explanations for Individual Neurons". In: *arXiv preprint arXiv:2405.06855* (2024).

[10] Pavel Pudil, Jana Novovičová, and Josef Kittler. "Floating search methods in feature selection". In: *Pattern recognition letters* 15.11 (1994), pp. 1119–1125.