Abstract: Color-Dependent Prediction Stability of Popular CNN Image Classification Architectures

Laurent Mertens[1,2,3], Elahe' Yargholi[4], Jan Van den Stock[5,6], Hans Op de Beeck[4], and Joost Vennekens[1,2,3]

[1] KU Leuven, De Nayer Campus, Dept. of Computer Science J.-P. De Nayerlaan 5, 2860 Sint-Katelijne-Waver, Belgium
[2] Leuven.AI - KU Leuven Institute for AI, 3000 Leuven, Belgium
[3] Flanders Make@KU Leuven, Leuven, Belgium
[4] Department of Brain and Cognition, Leuven Brain Institute, Faculty of Psychology & Educational Sciences, KU Leuven, 3000 Leuven, Belgium
[5] Neuropsychiatry, Leuven Brain Institute, KU Leuven, 3000 Leuven, Belgium
[6] Geriatric Psychiatry, University Psychiatric Center KU Leuven, 3000 Leuven, Belgium
laurent.mertens@kuleuven.be

This is an extended abstract of a paper published at the 32nd International Conference on Artificial Neural Networks [9].

ImageNet[7] [2] is a large, publicly available, image dataset (14M+ images). Its images are organized according to the WordNet hierarchy, making it especially useful for image classification tasks, as target labels are readily available. In 2010, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [10] introduced a particular subset of images from 1000 different categories known as the ImageNet-1k dataset, with an accompanying image classification challenge.

In 2012, a Convolutional Neural Network (CNN) now widely known as AlexNet [7] convincingly won this competition. This result led to quick and widespreak adoption of CNNs to solve image classification and recognition tasks; while AlexNet was the only CNN submitted in 2012, by the next year the majority of submissions were CNN-based. Other popular architectures that were either submitted to ILSVRC or trained on the ImageNet-1k dataset, and that will be evaluated in this paper, are VGG16 [11], ResNet18 and ResNet50 [4], and DenseNet161 [5].

Despite their popularity and successes, these architectures also have weaknesses, both ethical [12, 1, 13] and technical. Our contribution [9] demonstrates a technical weakness that relates to robustness w.r.t. color changes in the input images. Although complete color invariance is not desireable, neither should a useful CNN model completely alter its predictions when small color shifts, that would not affect humans, are applied to images.

We started by investigating the effect of applying hue and saturation shifts to ImageNet-1k images on ImageNet-1k trained models, both in terms of prediction robustness—i.e., does a prediction for an altered image differ from that of the

[7] https://www.image-net.org/

original image, regardless of the correctness of that original prediction?—and accuracy. In particular, we explored the prediction stability of the popular CNN architectures AlexNet, VGG16, ResNet18 and 50, and DenseNet161. We showed that all models alter their predictions when input images have their hue shifted, considering shifts in 10° steps, with larger shifts increasing alteration frequency. Averaged over all hue shifts, relative model performance experiences a drop of 41.5%, 22.9%, 21.4%, 11.3% and 14.3% respectively for the aforementioned models, resulting in an average drop of 22.28% over all models; larger models show less sensitivity. The largest drops are observed within up to 30° shifts from reference, with performance stabilizing around the 80° mark. Saturation shifts elicit similar but more restrained behavior, with an average performance drop of only 4.0% over all models. Importantly, for both hue and saturation alterations, the prediction for images originally correctly predicted tends to be more robust than for images originally wrongly predicted.

Next, we turned our attention to EmoNet [6], an image classification model obtained by taking AlexNet trained on ImageNet-1k, replacing its last layer with a 20 node linear layer and training only this new layer on a custom dataset of 137k images annotated with one of 20 emotion labels representing the emotion elicited by the images in an observer. We showed that EmoNet inherits essentially the same behavior as its parent. EmoNet forms an interesting case, because elicited emotions form a dimension that can also reasonably be assumed to be independent of moderate color changes; a few degrees of hue shift shall not make a puppy less cute.

Following this, we looked at some of the earlier mentioned CNNs, but trained from scratch on different large datasets. In particular, we considered Stylized ImageNet [3], a dataset derived from ImageNet-1k by means of style transfer, and Places365 [14], a dataset of millions of images annotated with one of 365 scene classes. Stylized ImageNet is of particular interest, as its authors specifically constructed the dataset to obtain models that use more global ("style") rather than local ("texture") features. By comparing the effect of color-related changes on a same architecture trained on different datasets, we showed that the sensitivity to color changes appears to be an inherent property of the architecture, rather than a consequence of the training data.

Finally, we proposed to include two additional preprocessing steps in the training process, namely random hue shifts and saturation changes, which, when used to retrain existing models, were shown to improve average prediction stability for hue shifts on ImageNet-1k with 19%, 13% and 12% for AlexNet, VGG16 and ResNet18 respectively. For saturation changes, 11%, 6% and 6% improvements were obtained, in the last two cases lifting stability up to 94% and 93%. Interestingly, these retrained models retain the original model's ImageNet-1k performance, leading to the question: How exactly can several sets of convolution filters result in the same ImageNet-1k accuracy, yet show markedly different behavior when subjected to particular image transformations?

All our code and models are available through our GitLab page [8].

References

1. Crawford, K., Paglen, T.: Excavating ai: The politics of images in machine learning training sets. https://excavating.ai/, accessed: 8th March, 2023
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness (2018). https://doi.org/10.48550/ARXIV.1811.12231
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
5. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269 (2017). https://doi.org/10.1109/CVPR.2017.243
6. Kragel, P.A., Reddan, M.C., LaBar, K.S., Wager, T.D.: Emotion schemas are embedded in the human visual system. Science Advances **5**(7), eaaw4358 (2019). https://doi.org/10.1126/sciadv.aaw4358
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (may 2017). https://doi.org/10.1145/3065386
8. Mertens, L.: Gitlab repository containing the code and additional material for this paper. https://gitlab.com/EAVISE/lme/nncolorstabilityanalysis-paper
9. Mertens, L., Yargholi, E., Van den Stock, J., Op de Beeck, H., Vennekens, J.: Color-dependent prediction stability of popular cnn image classification architectures. In: Artificial Neural Networks and Machine Learning – ICANN 2022: 31st International Conference on Artificial Neural Networks (Under review)
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
12. Steed, R., Caliskan, A.: Image representations learned with unsupervised pre-training contain human-like biases. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 701–713. FAccT '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445932
13. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 547–558. FAT* '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3375709
14. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)