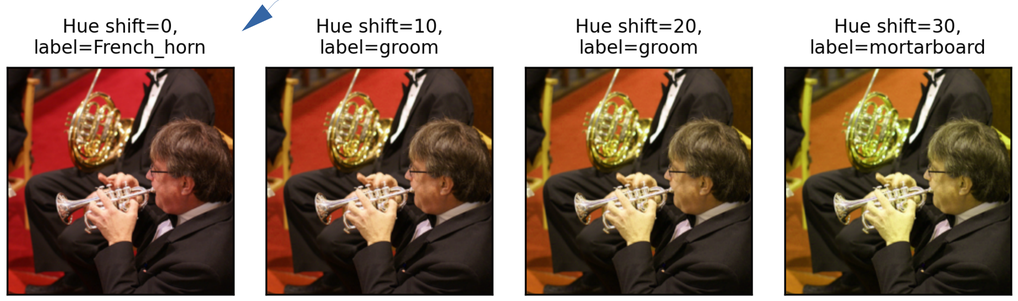
LEUVEN.AI
DTAI
Declarative Languages & Artificial Intelligence
KU LEUVEN
eavise
SCIENCE
AI

Co-authored with Elahe' Yargholi, Jan Van den Stock, Hans Op de Beeck, Joost Vennekens

Laurent Mertens
Dept. Of Computer Sciences
laurent.mertens@kuleuven.be

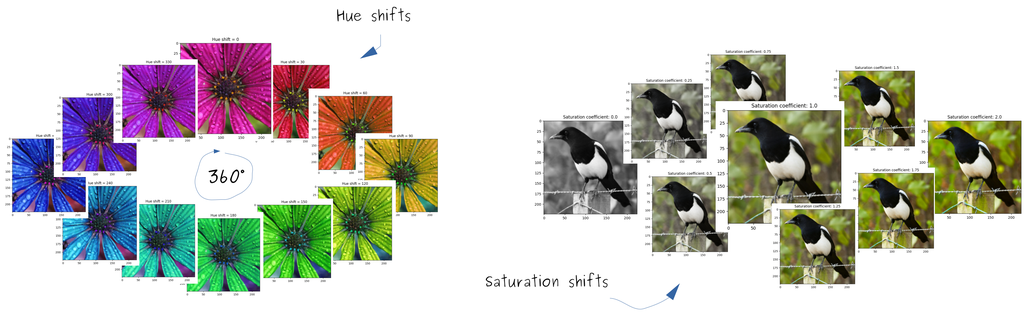
Color-Dependent Prediction Stability of Popular CNN Image Classification Architectures

Question: do you think this image represents something different than the images to its right?



If you answered "no", you disagree with AlexNet, which provided the shown labels!

Turns out that when you show **color-modified** versions of images…

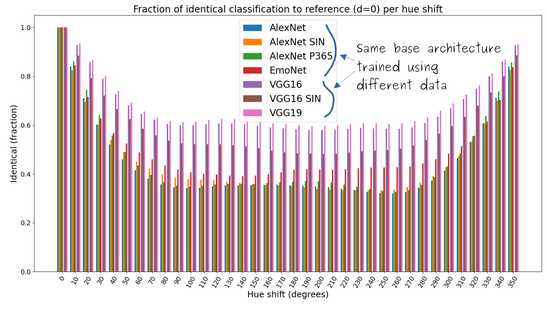


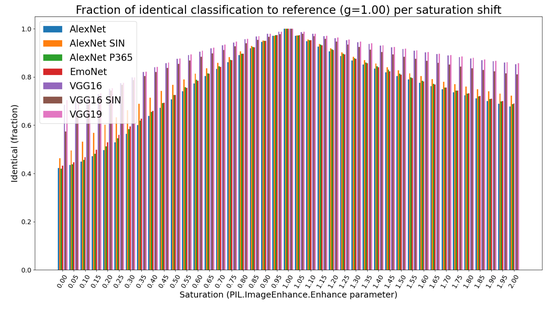
…to popular CNN models…

…they often alter their predictions!

Table 1. Overview of training and validation data per model. "⟨ModelName⟩" is a placeholder for a valid architecture, "IN-1k" = ImageNet-1k, "SIN" = Stylized ImageNet, "train" = train data, "val" = validation data.

Model	Trained on	Validated on
AlexNet, VGG16, ResNet18/50, DenseNet161	IN-1k train	IN-1k val
⟨ModelName⟩-SIN	SIN	IN-1k val
⟨ModelName⟩-P365	Places365 train	Places365 val
EmoNet	IN-1k train + EmoNet	IN-1k val





How often does the model predict the same label as for the original image, regardless of whether that label was correct?

	AlexNet	AlexNet SIN	AlexNet P365	EmoNet	VGG16	VGG16 SIN	ResNet18	ResNet18 P365	ResNet50	ResNet50 SIN	DenseNet161	DenseNet161 P365
Equal d_{all}	$.431^{-.15}$	$.441^{-.14}$	$.447^{-.15}$	$.489^{-.12}$	$.659^{-.10}$	$.581^{-.12}$	$.659^{-.10}$	$.614^{-.11}$	$.800^{-.06}$	$.632^{-.10}$	$.759^{-.07}$	$.662^{-.09}$
Equal $\lvert d\rvert \leq 30$	$.718^{-.10}$	$.709^{-.09}$	$.747^{-.09}$	$.724^{-.09}$	$.861^{-.05}$	$.803^{-.06}$	$.847^{-.05}$	$.830^{-.06}$	$.910^{-.03}$	$.826^{-.06}$	$.897^{-.04}$	$.847^{-.05}$
Equal g_{all}	$.746^{-.15}$	$.786^{-.13}$	$.756^{-.15}$	$.758^{-.15}$	$.883^{-.08}$	$.857^{-.09}$	$.874^{-.09}$	$.842^{-.10}$	$.950^{-.03}$	$.885^{-.08}$	$.946^{-.05}$	$.863^{-.09}$
Equal $g \in]0.5, 1.5[_{\setminus\{1\}}$	$.863^{-.07}$	$.883^{-.06}$	$.872^{-.07}$	$.870^{-.07}$	$.943^{-.03}$	$.924^{-.04}$	$.937^{-.03}$	$.919^{-.04}$	$.972^{-.01}$	$.940^{-.03}$	$.977^{-.01}$	$.929^{-.04}$
Top1 d_0,g_1	.566	.400	-	-	.716	.522	.697	-	.803	.602	.771	-
Top1 d_{all}	$.331^{-.09}$	$.263^{-.06}$	-	-	$.552^{-.07}$	$.409^{-.05}$	$.548^{-.06}$	-	$.713^{-.04}$	$.488^{-.05}$	$.661^{-.05}$	-
OL+ d_{all}	$.528^{-.15}$	$.579^{-.14}$	-	-	$.732^{-.10}$	$.713^{-.10}$	$.743^{-.09}$	-	$.857^{-.05}$	$.752^{-.09}$	$.823^{-.06}$	-
OL− d_{all}	$.305^{-.14}$	$.349^{-.14}$	-	-	$.475^{-.12}$	$.437^{-.13}$	$.465^{-.12}$	-	$.568^{-.09}$	$.451^{-.12}$	$.541^{-.10}$	-
Top1 g_{all}	$.505^{-.06}$	$.382^{-.02}$	-	-	$.688^{-.04}$	$.511^{-.02}$	$.668^{-.04}$	-	$.798^{-.01}$	$.591^{-.02}$	$.762^{-.02}$	-
OL+ g_{all}	$.855^{-.14}$	$.846^{-.14}$	-	-	$.917^{-.15}$	$.874^{-.14}$	$.915^{-.15}$	-	$.938^{-.15}$	$.897^{-.15}$	$.934^{-.15}$	-
OL− g_{all}	$.253^{-.17}$	$.345^{-.15}$	-	-	$.379^{-.16}$	$.361^{-.11}$	$.359^{-.14}$	-	$.451^{-.09}$	$.362^{-.10}$	$.399^{-.10}$	-
O.P. d_{all}	58.5^{122}	40.6^{96}	29.7^{53}	3.7^{3}	28.0^{76}	26.5^{73}	25.3^{71}	13.0^{29}	31.6^{119}	24.5^{70}	19.1^{61}	9.9^{24}
O.P. $\lvert d\rvert \leq 30$	10.9^{37}	6.7^{23}	4.9^{13}	2.1^{2}	5.2^{18}	4.4^{16}	5.0^{18}	2.8^{6}	6.4^{45}	4.2^{14}	4.0^{15}	2.6^{6}
O.P. g_{all}	11.9^{40}	19.8^{55}	20.0^{44}	2.3^{2}	5.3^{20}	4.3^{17}	4.9^{18}	3.8^{10}	5.5^{23}	11.9^{36}	2.9^{10}	2.9^{8}
O.P. $g \in]0.5, 1.5[_{\setminus\{1\}}$	2.5^{5}	18.9^{52}	19.4^{42}	1.4^{1}	1.6^{2}	1.5^{2}	1.6^{2}	1.5^{1}	5.5^{22}	11.9^{36}	1.2^{1}	1.4^{1}

- Behavior seems **independent of training data.**
- Behavior is **inherited** through transfer learning.
- Larger architectures appear less sensitive.
- Effect is more pronounced for images for which the model originally made a wrong prediction.
- Different label is more than just "flipping second and first place"

(And yes, additional pre-processing during training alleviates the problem…)

Code and extra plots available at **HTTPS://GITLAB.COM/EAVISE/LME/NNCOLORSTABILITYANALYSIS-PAPER!**