# [Supplementary Document]Assessing The Importance Of Colours For CNNs In Object Recognition

**Anonymous Author(s)**
Affiliation
Address
email

## 1 Jensen-Shannon Measure

To reiterate the notations used, a dataset $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, N\}$ is composed of images $x_i \in \mathbb{R}^{C \times H \times W}$ and their corresponding labels $y_i$. $\mathcal{D}^{train}$, $\mathcal{D}^{test}$ denotes the split of the dataset into train and test sets respectively. Jensen Shannon divergence between two stylisations of the same dataset is defined as:

$$JS(\mathcal{D}_1^{train}, \mathcal{D}_2^{train}) = \frac{1}{|\mathcal{D}^{train}|} \sum_{i \in \mathcal{D}^{train}} \frac{1}{|C|} \sum_{j \in C} JS(T_1(x_i)[j], T_2(x_i)[j])$$

where, $T_i$ is the transformation corresponding to $\mathcal{D}_i$ and indexing $[j]$ returns the normalised intensity histogram for the $j^{th}$ channel.

The corresponding results in table 1 show that switching channels is a gentler transformation than composing negatives in preserving original shape and texture. Greyscale is consistently with the least amount of JS divergence and can suggest as to why consistently $\text{Acc}(\mathcal{D}_G) > \text{Acc}(\mathcal{D}_I)$.

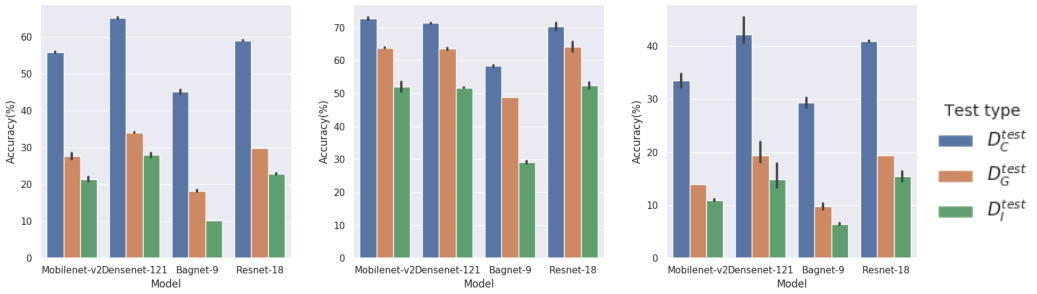
Table 1: JS measure between stylisations

| Datasets | $JS(\mathcal{D}_C^{train}, \mathcal{D}_G^{train})$ | $JS(\mathcal{D}_C^{train}, \mathcal{D}_I^{train})$ | $JS(\mathcal{D}_C^{train}, \mathcal{D}_{Neg}^{train})$ |
|---|---|---|---|
| C100 | 0.07 | 0.12 | 0.33 |
| TIN | 0.06 | 0.1 | 0.29 |
| S10 | 0.06 | 0.1 | 0.3 |
| CUB | 0.1 | 0.15 | 0.34 |
| OF | 0.09 | 0.13 | 0.32 |
| OP | 0.04 | 0.07 | 0.27 |

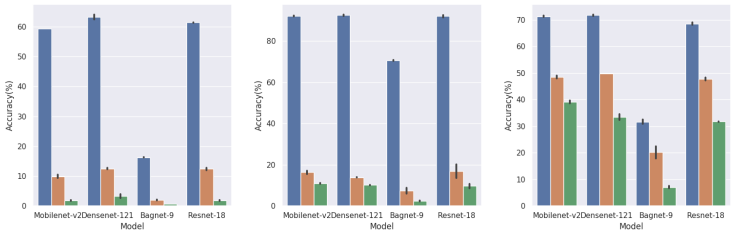
## 2 Vanilla training results on different architectures

We also report the results on different CNN architectures, figure 1 showcases our findings. The results indicate that different architectures display similar behaviour for colour importance across datasets. This supports the fact that the architecture plays no role in driving the attention of a model towards colour and the importance of colour is dependent on the task at hand.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.




(a) CIFAR-100 (b) STL-10 (c) Tiny ImageNet




(d) CUB-200-2011 (e) Oxford-Flowers (f) Oxford-IIIT Pets

Figure 1: Vanilla training results for different architectures

## 3 Evaluation on ImageNet-16

ImageNet-16 is a subset of the ImageNet dataset. It was utilised in the psychophysical studies [1, 2] for evaluating texture bias and generalisation performance of models. Approximately 200 fine grained classes of the ImageNet dataset have been merged to produce 16 broader categories. The resulting categories are: *airplane, bear, bicycle, bird, boat, bottle, car, cat, chair, clock, dog, elephant, keyboard, knife, oven and truck*. We evaluate this data under the proposed experimental setup i.e. with congruent, greyscale and incongruent images. An important thing to notice is that the original model being evaluated has been trained on the original ImageNet dataset(1000 object categories). For evaluation, if the model classifies an image to any one of the finer classes we assign it as a correct prediction. For example, if an image of a German Shepherd is classified as a Golden Retriever, we will consider it as a correct classification for Dog category.

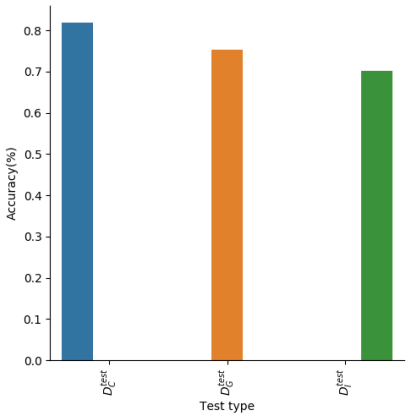



Figure 2: Resnet-50 on ImageNet-16.

Figure 2 shows the results for ImageNet-16. The ordering of accuracies is in line to other datasets reported in our paper. The differences in accuracies are significant yet comparatively low when compared to other datasets. Geirhos et al. [1] have already shown that for these categories the decisions are mostly made by textures.

## 4 Training details

We utilised Pytorch framework for all of our experiments. The models and datasets are readily available via. torchvision. For the datasets which are not available through torchvision, the details on how to acquire them are provided in the *README.md* document of the shared code. Missing key-values in table $i$ can be found via. recursive search in table $i-1$. All the fgvc datasets utilise similar training hyper-parameters and hence we have provided only the details for CUB-200.

Table 2: Training details for toy experiment

| Key | Value | Note |
|---|---|---|
| Model | Resnet-18 | – |
| Data Augmentation | Random(rotation, horizontal flip) | – |
| $\lvert\mathcal{D}^{train}\rvert$ | 999 | Generated uniformly at random |
| $\lvert\mathcal{D}^{test}\rvert$ | 999 | Generated uniformly at random |
| Image size | $224 \times 224$ | – |
| Batch size | 32 | – |
| Epochs | 15 | – |
| Optimiser | SGD | – |
| Learning rate(LR) | 0.01 | Halved at LR decay epoch |
| LR decay epoch | 7 | – |

Table 3: Training details for CIFAR-100

| Approach | Key | Value |
|---|---|---|
| Common | Models | Bagnet-9, Resnet-18, Densenet-121, Mobilenet-v2 |
| | Image size | $32 \times 32$ |
| | Train aug. | Random(rotation, horizontal flip), standardisation |
| | Test aug. | Standardisation |
| | Batch size | 128 |
| | Optimiser | SGD |
| | LR decay rate | 0.5 |
| Vanilla | Epochs | 200 |
| | LR | 0.1 |
| | Train aug. | Common |
| | LR decay epochs | $[50, 100, 150]$ |
| Fine-tuning | Pre-trained weights | ImageNet |
| | LR | 0.01 |
| | Epochs | 30 |
| | LR decay epochs | $[15]$ |
| Color augmentation | Pre-trained weights | ImageNet |
| | Train aug. | Common + Random color jitters |
| | LR | 0.01 |
| | Epochs | 30 |
| | LR decay epochs | $[15]$ |
| Incongruent Training | Pre-trained weights | ImageNet |
| | Train aug. | Common + Random (color jitters, channel switching) |
| | Finetuned with | Random(rotation, horizontal flip) |
| | LR | 0.01 |
| | Epochs | $30, 30$ |
| | LR decay epochs | $[15], [15]$ |

Table 4: Training details for STL-10

| Approach | Element | Value |
|---|---|---|
| Common | Image size | $96 \times 96$ |
| | Batch size | $64$ |
| Vanilla | Epochs | $200$ |
| | LR | $0.1$ |
| | LR decay epochs | $[40, 80, 120, 160]$ |
| Fine-tuning | Epochs | $50$ |
| | LR | $0.01$ |
| | LR decay epochs | $[15, 30, 45]$ |
| Color augmentation | Epochs | $50$ |
| | LR | $0.01$ |
| | LR decay epochs | $[15, 30, 45]$ |
| Incongruent Training | Epochs | $50, 50$ |
| | LR | $0.01$ |
| | LR decay epochs | $[15, 30, 45]$, $[15, 30, 45]$ |

Table 5: Training details for Tiny ImageNet

| Approach | Element | Value |
|---|---|---|
| Common | Image size | $64 \times 64$ |
| | Batch size | $128$ |
| Vanilla | Epochs | $150$ |
| | LR | $0.1$ |
| | LR decay epochs | $[30, 60, 90, 120]$ |
| Fine-tuning | Epochs | $50$ |
| | LR | $0.01$ |
| | LR decay epochs | $[15, 30, 45]$ |
| Color augmentation | Epochs | $50$ |
| | LR | $0.01$ |
| | LR decay epochs | $[15, 30, 45]$ |
| Incongruent Training | Epochs | $50, 50$ |
| | LR | $0.01$ |
| | LR decay epochs | $[15, 30, 45]$, $[15, 30, 45]$ |

# References


[1] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bygh9j09KX`.

[2] Robert Geirhos, David H. J. Janssen, Heiko H. Schütt, Jonas Rauber, Matthias Bethge, and Felix A. Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker, 2017.

Table 6: Training details for CUB-200

| Approach | Key | Value |
|---|---|---|
| Common | Image size | $224 \times 224$ |
| | Train aug. | Random(rotation, horizontal flip, crop), standardisation |
| | Test aug. | center crop(224), standardisation |
| | Batch size | 32 |
| | Optimiser | SGD |
| | LR decay rate | 0.5 |
| Vanilla | Epochs | 200 |
| | LR | 0.1 |
| | Train aug. | Common |
| | LR decay epochs | $[50, 100, 150]$ |
| Fine-tuning | Pre-trained weights | ImageNet |
| | LR | 0.01 |
| | Train aug. | Common |
| | Epochs | 40 |
| | LR decay epochs | $[15, 30]$ |
| Color augmentation | Pre-trained weights | ImageNet |
| | LR | 0.01 |
| | Train aug. | Common + Random color jitters |
| | Epochs | 40 |
| | LR decay epochs | $[15, 30]$ |
| Incongruent Training | Pre-trained weights | ImageNet |
| | Train aug. | Common + Random (color jitters, channel switching) |
| | Finetuned with | Common aug. |
| | LR | 0.01 |
| | Epochs | $40, 40$ |
| | LR decay epochs | $[15, 30], [15, 30]$ |