# ImageNet suffers from dichotomous data difficulty

*Kristof Meding*[1], Luca M. Schulze-Buschoff*[1], Robert Geirhos[12] and Felix A. Wichmann[1]*

*joint first authors

**Today, CNNs are incredibly powerful generalisation systems---but to what degree have we understood how their inductive bias influences model decisions? We here attempt to disentangle the various aspects that determine how a model decides. In particular, we ask: what makes one model decide differently from another? In a meticulously controlled setting, we find that irrespective of the network architecture or objective (e.g. self-supervised, semi-supervised, vision transformers, recurrent models) all models end up with similar decision. For the range of investigated models and their accuracies, ImageNet is dominated by trivial and impossible images (beyond label errors). Removing the "impossible" and "trivial" images allows us to see pronounced differences between models. This implies that in future comparisons of machines, much may be gained from investigating the decisive role of images and the distribution of their difficulties.**

## ImageNet is dominated by 46% "trivial" and 12% "impossible" images*

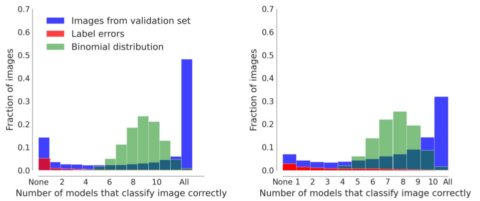
## Thus, all models make similar decisions

## Removing "impossible" and "trivial" images pronounces model differences

***Of course, this depends on the accuracy of the models. More details can be found here: https://arxiv.org/abs/2110.05922**

### Dichotomous data difficulty in a nutshell

We have tested various factors related to the inductive bias—among other aspects, architecture, optimiser, learning rate, and initialisation—and yet, on ImageNet, all models agree in the sense that they all make largely similar errors. Even radically different state-of-the-art (SOTA) models make surprisingly similar errors on the ImageNet validation set. To a certain degree, image difficulty appears dichotomous.



(a) ResNet-18 variants

(b) State-of-the-art models

Figure 1: Irrespective of model differences (e.g. architecture, hyperparameters, optimizer), most ImageNet validation images are either "trivial" (in the sense that all models classify them correctly) or "impossible" (all models make an error). For comparison, a binomial distribution of errors is shown: this is the distribution of errors expected for completely independent models if all images were equally difficult.

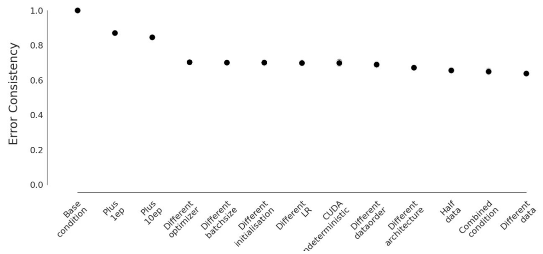
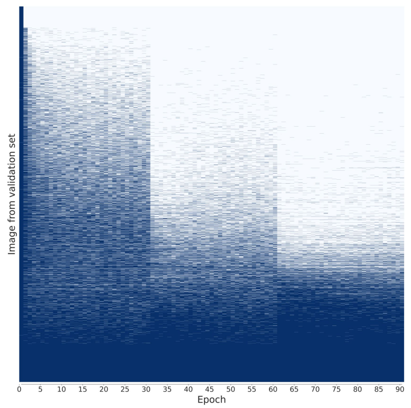


Figure 2: ResNet-18 results. Error consistencies between the different conditions and the base network on the ImageNet validation set after 90 epochs. For conditions for which multiple models were trained the mean over all models of a condition is plotted in black.



Decisions on all 50K ImageNet validation images of the ResNet-18 base condition over the epochs. Blue indicates that the respective item was falsely classified during the specific epoch, while white indicates that it was correctly classified. The items from the ImageNet validation set are ordered according to the mean accuracy the base network achieved on them over the course of the 90 epochs. Therefore, items which were classified correctly from epoch 1 are on top and items which were classified incorrectly from epoch 1 are on the bottom.

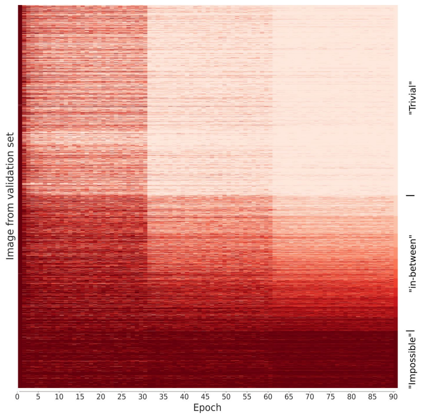


Figure 2: Decisions on all 50K ImageNet validation images of all 13 ResNet-18 networks with different inductive biases. Dark red indicates that the respective item was falsely classified by all networks. Light red indicates that the image was correctly classified by all networks. Images are ordered according to the mean accuracy across networks in the last epoch.

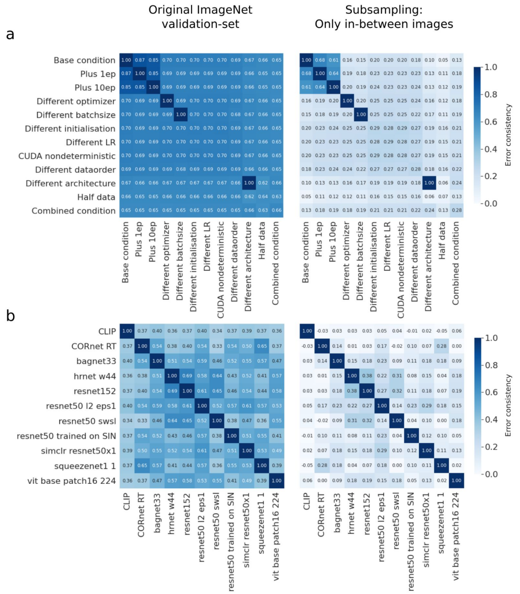


Figure 3: Error consistency on the original ImageNet test-set (left panels) and on in-between images only (right panels) for the ResNet-variants (a) and the SOTA networks (b). Error consistency around 0 indicates independent responses. A diagonal element of 1 represents that only one network for comparison was available. Clearly, removing the trivial and impossible images (right panels) shows that different network architectures are behaving differently from one another, i.e. their different processing strategies are not longer masked by DDD (left panels)—thus allowing more insights to be gained.

**Affiliation**
[1]: Neural Information Processing Group, University of Tübingen, Germany
[2]: IMPRS-IS

**Funding Acknowledgment:**
Funding was provided, in part, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 276693517, SFB 1233, TP 4 Causal inference strategies in human vision (K.M., L.S.B., and F.A.W.). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting R.G; and the German Research Foundation through the Cluster of Excellence ``Machine Learning—New Perspectives for Science'', EXC 2064/1, project number 390727645, for supporting F.A.W.