ImageNet suffers from dichotomous data difficulty

Kristof Meding*
University of Tübingen
Corresponding author: kristof.meding@uni-tuebingen.de

Luca M. Schulze Buschoff*
University of Tübingen

Robert Geirhos
University of Tübingen & IMPRS-IS

Felix A. Wichmann
University of Tübingen

Abstract

"The power of a generalization system follows directly from its biases" (Mitchell 1980). 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 (1.) irrespective of the network architecture or objective (e.g. self-supervised, semi-supervised, vision transformers, recurrent models) all models end up with a similar decision boundary. (2.) To understand these findings, we analysed model decisions on the ImageNet validation set from epoch to epoch and image by image. We find that the ImageNet validation set suffers from dichotomous data difficulty (DDD): For the range of investigated models and their accuracies, it is dominated by 46.0% "trivial" and 11.5% "impossible" images. Only 42.5% of the images are responsible for the differences between two models' decision boundaries. The impossible images are not driven by label errors. (3.) Finally, humans are highly accurate at predicting which images are "trivial" and "impossible" for CNNs (81.4%). Taken together, it appears that ImageNet suffers from dichotomous data difficulty. This implies that in future comparisons of brains, machines and behaviour, much may be gained from investigating the decisive role of images and the distribution of their difficulties.

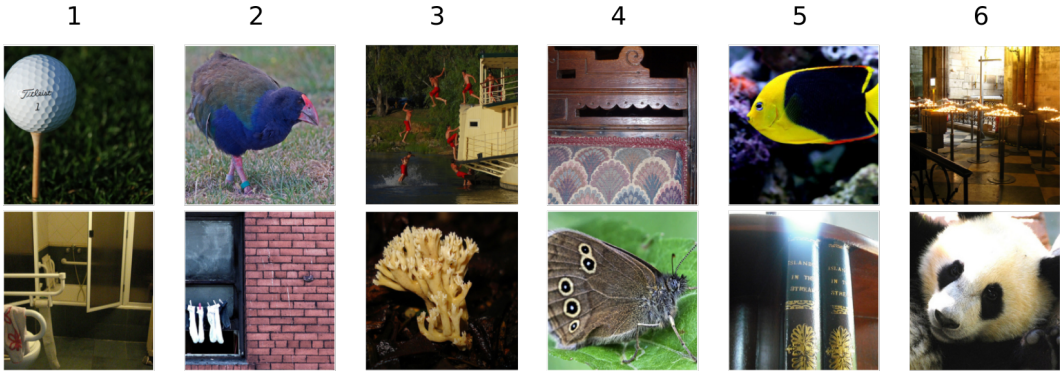


Figure 1: Can you predict which of these images are "tricky" for CNNs? Out of every of the six pairs, one image is correctly classified and one incorrectly (answers on the next page[1]). On ImageNet, image difficulty appears largely dichotomous: CNNs make highly systematic errors irrespective of inductive bias (architecture, optimiser, ...). Humans can reliably differentiate between images that are "trivially easy" and "impossibly hard" for CNNs (81% accuracy).

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.
*joint first authors.

1 Introduction

Let's play a game we call *Find those tricky images!* In Figure 1, we show pairs of images. One image is impossible for a CNN regardless of its architecture, optimiser, random seed etc.—it never gets the label correct. The other image always yields a correct classification—can you find the tricky images?

Done? We will wait. You have probably never seen these images before, and neither have CNNs seen them during training. How exactly a decision maker—be it a neural network, or a biological brain—generalises to previously unseen images is influenced by the decision maker's *inductive bias* [1]—in fact, as already recognised in 1980, "the power of a generalisation system follows directly from its biases"[2]. Commonly, the inductive bias is defined as the set of assumptions and choices that determine which hypothesis space is available to the model, before the model is exposed to data. For instance, starting from the set of all possible hypotheses, the hypothesis space of linear models (linearity is one example of a strong inductive bias) is a tiny subset. After the "choice" of the inductive bias, the dataset then influences which particular decision boundary (or concrete hypothesis) is selected from the model's hypothesis space. Finding the right inductive bias for a given problem is at the core of machine learning. Therefore, it is only consequent that a tremendous amount of work is being invested in improved architectures [3], optimisers [4], learning rate schedules [5], etc.—surely we would expect these choices to make a difference on the resulting model's decision boundary even if trained on the identical dataset. However, in the present work, we have tested various factors related to the inductive bias—among other aspects, architecture, optimiser, learning rate, and initialisation—and yet, they all agree (and perhaps even with you?) in the sense that they *all* make largely similar errors. We here investigate why this is the case. CNNs put relatively few constraints on their hypothesis space; but how much do common choices of architecture and hyperparameters (i.e. selecting a—hopefully suitable—inductive bias for a model) influence the resulting decision boundary? In other words, what makes one model decide differently from another?

This is the question we set out to investigate with a set of highly controlled experiments, investigating the similarities and differences of various ImageNet-trained models. Making progress towards answering this question is relevant in a number of different contexts. In machine learning, we would like to know where we can gain most from further research (e.g. if we were to find that the choice of optimiser does not have a strong influence on the decision boundary, then we might want to increase our research efforts on other aspects like the architecture) and at the same time, we would like to know which choices have a perhaps unwanted influence—for instance, it would be reassuring to know that changing the batch size does not have a strong influence on model errors. In neuroscience and cognitive science, when searching for candidate models of human visual object recognition, we would like to know along which dimensions models differ most, and how this relates to theories of biological object recognition. Finally, in applied settings, one would like to understand which choices matter as well, since the model's decision boundary (influenced by the model's inductive bias) determines which types of errors a model will make; for instance which items are classified as "faulty" on a production line or which types of X-ray scans are likely to be misclassified by a model.

Related work. There are (at least) three lines of research related to our work: papers proposing metrics for comparing CNNs, papers investigating the consistency of model errors, and papers pointing out challenges/problems of ImageNet. Due to space constraints, the related work section can be found in the Appendix (Section B).

2 Methods

Similarity measure For the investigation of network similarities, we mainly use the behavioural measure error consistency (κ)[7] based on Cohen's work[8]. $\kappa > 0$ represents that two decision-makers systematically make errors on the same images; $\kappa = 0$ indicates no more error overlap than what could be expected by chance alone. $\kappa < 0$ shows that two networks systematically disagree.
Network variations In our experiments, we investigated the systematic agreement between CNNs, varying not only architecture but carefully controlling for the number of epochs, optimiser, batch size, random initialisation, learning rate, hardware randomness, data order, architecture, and disjoint data sampling. Unless stated otherwise, we only changed one of the above parameters at a time. We first used systematic variations on ResNet-18 (called Res-Net 18 variants). Details can be found in

[1]The tricky (=misclassified) images are: 1. bottom, 2. bottom, 3. top, 4. top, 5. bottom, 6. top. This game is an homage to "Name that dataset" by Torralba and Efros [6].

Section B.3 in the Appendix. Later, we investigated different state-of-the-art network architectures. The implementations are obtained from `modelvshuman` [9]. Details of our used hardware, software and the psychophysical experiment can be found in Appendix B.4.

3 Results

3.1 Model inductive bias has only negligible influence on model errors due to dichotomous data difficulty (DDD)

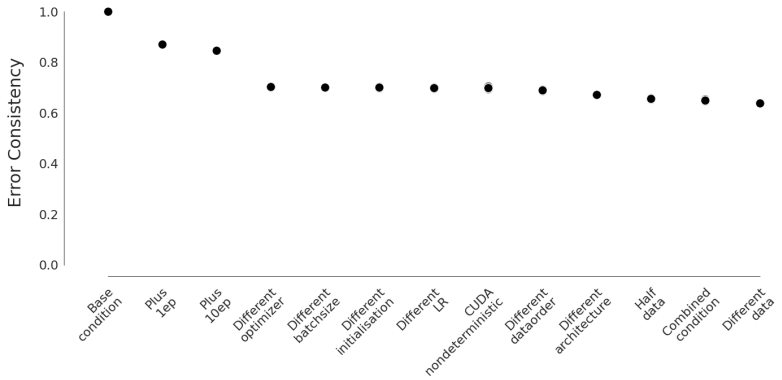


Figure 2: Error consistencies between the different conditions and the base network for ResNet-18 variants on the ImageNet validation set after 90 epochs. For conditions for which multiple models were trained. The mean over all models of a condition is plotted in black.

Figure 2 shows the result of our controlled inductive bias study on ImageNet. We plot the error consistency for all conditions compared to the base network (for error consistencies between all conditions see the matrix Figure 10 in the appendix). A high error consistency means that the networks agree way beyond what is expected by independent decision-makers. Regardless of the parameter changed (architecture, optimiser, etc.), we do not find a large difference, that is, mean error consistencies are very high indeed (around 0.7). While we also plot the single instance network error consistencies in grey, these are non-visible since the variance within the conditions is very small (except for the "cuda nondeterministic condition"). Strikingly, changes that we hypothesised would make a larger difference, e.g. different architecture, have basically the same error consistency as "minor" changes like enabling hardware randomness on the GPUs. All networks achieved similar top-1 accuracies (mean: 69.05% after 90 epochs; range: 65.87% to 71.47%; standard deviation: 1.60%, cf. Figure 7 in the Appendix). Another popular method for agreement analysis is RSA. All our results also hold here, see Figure 17 in the appendix. Additionally, switching the base architecture to VGG-11 or DenseNet-121 does not make a difference either (see Figures 8 and 9 in the Appendix).

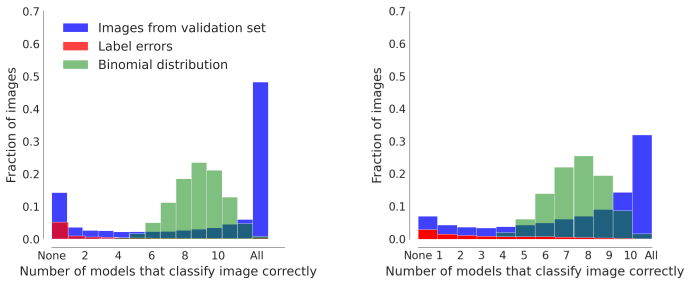
We go into a deeper analysis of why the inductive bias has such an insignificant role in Figure 4 (due to space constraints and the large figure it's not in the main paper), which plots, for the base network, whether ImageNet validation images are classified correctly (white) or incorrectly (blue) across epochs. There are three take-aways from this visualisation. (1.), one immediately notices the influence of the standard learning rate steps after 30 and 60 epochs. However, after this step, some images (bottom) are also "forgotten" (classified correctly before step but incorrectly afterwards), which contrasts with the usual expectation that a model gradually improves over time. (2.), some images are learned immediately during the very first epoch and never forgotten later (top right region), while some are never learned at all. (We will later see that this is not an effect of label errors) (3.), while accuracy usually only improves minimally from one epoch to the next (e.g. 0.04% from epoch 89 to epoch 90, or 14 additionally correctly classified images out of 50,000), on average 12.37% of the models' image classification decisions swap every epoch, corresponding to 6,184 images! (See Figure 11 in the appendix for a plot that shows the percentage of swapped labels from epoch to epoch).

This last finding becomes even more prominent in Figure 5 (again in the appendix), where we overlay the previous figure for all of the 13 networks with different hyperparameters, architectures etc. (explained in Section 2). A light red entry indicates that *all* networks correctly classify the

image, a dark red entry that all networks classify the corresponding image; shades of red indicate the cases in-between (where, e.g., some but not all networks make errors). The figure illustrates that the previous findings even hold across very different inductive biases. We observe that 48.2 % images are learned by all models regardless of their inductive bias; 14.3 % images are consistently misclassified by all models[2]; only roughly a third (37.5%) of images are responsible for the differences between two model's decisions. We call this phenomenon dichotomous data difficulty (DDD): while the inductive bias restricts the hyperparameter space for a given model, the nature of the dataset—and especially its highly non-uniform image difficulty—seems to be an important cause for the high similarity in the decisions of different networks. Model inductive bias may play a bigger role for images of intermediate difficulty (where there is substantial consistency variation across models), whereas its influence appears to be smaller for easy and hard images. As the dataset primarily consists of images that all models either classify correctly or incorrectly, all models end up with similar classification behaviour.

Is dichotomous data difficulty (DDD) only a problem for ImageNet? This is not the case: DDD is also present in CIFAR-100 and even a synthetic Gaussian dataset. As a first indication, for both of these datasets we find similarly high error consistencies between different models, just like we found for ImageNet (see Section B.1 in the appendix)

Let us consider two extreme cases in order to put these findings into context. On one end of the spectrum, if all images were equally difficult *and* if all networks were independent (i.e. their different inductive biases would result in independent decision boundaries), then we could expect a binomial distribution of model errors: out of 13 investigated models, very few images should be misclassified by all models and very few correctly classified by all models—instead, most images should be correctly classified by a handful of models. Figure 3a shows, in green, exactly this distribution expected for independent models and equally difficult data. On the extreme end of the spectrum, if the inductive bias had no influence at all and the dataset only contained "trivial" and "impossible" images, we would expect a histogram with only two "spikes": given ImageNet accuracies of 69.05% on average, one spike at "None" (30.95% for ImageNet) and one at "All" (69.05 % for ImageNet). Clearly, the empirically obtained histogram (blue) much more resembles the latter, i.e. the scenario where the (nearly) dichotomous data difficulty dominates over inductive bias. We observe that DDD on ImageNet is amplified, but not caused, by label errors ([10–12]) only have a minor influence on the "None-Bar" from our histogram in Figure 3a Hence: removing erroneous labels is beneficial and laudable, but it will not solve DDD.



(a) ResNet-18 variants

(b) State of the art models

Figure 3: Histogram showing how many models correctly classify ImageNet validation sets images in the last epoch (a) or for SOTA models (b). In blue, the density of answers from the validation set ("None": no model was correct; "All": All models were correct.). In green, samples are drawn from a binomial distribution with same mean accuracy. Additionally label errors [10] are shown in red.

3.2 State of the art models are affected by DDD

Let's do a break here. What did we find? Changing the inductive bias within *one model class* does not change the decision boundary significantly. Do our results generalize across model classes? We redo the analysis in Figure 3a with radically different models: self-supervised models (SimCLR), a semi-supervised model (SWSL), a vision transformer (ViT), a recurrent model (CORnet-RT), a very deep model (ResNet-152), a highly compressed model (SqueezeNet), an adversarially trained model (ResNet-50 with epsilon 1 L2-robustness on ImageNet by Salman et al., 2020), a bag-of-local-features model (Bagnet-33), and OpenAI's CLIP model with a transformer architecture and joint image-language training objective (11 models in total). Again, we find the same pattern, shown in

[2]Of course, these numbers change if one uses an architecture with higher top-1 accuracy, see also next section.

Figure 3b. In total, 46.0 % "trivial" images are learned by all except one model; 11.5 % "impossible" images are consistently misclassified by all except one model. (42.5%) of images are responsible for the differences between two model's decisions.

Finally, we asked whether there are some particularly easy and hard classes, or whether DDD occurs largely independent of classes (i.e. affecting images of all classes similarly). When plotting class-wise accuracies for ImageNet. Figure 6 clearly shows some classes are *very* easy to classify (e.g. up to 100% top-1 accuracy on ImageNet), while other classes are *very* difficult (e.g. down to 10% top-1 accuracy on ImageNet, for a list of top-10 easiest and hardest classes see Section B.8 in the Appendix). This means that there are *both* easy and hard images as well as easy and hard classes. Additionally, we show in the appendix that subsampling of in-between images pronounces differences between models (see Section).

3.3 Humans are highly accurate at predicting which images are difficult for CNNs

Since we found DDD to affect CNNs across datasets, we were interested to understand whether humans could identify which images were "trivial" and "impossible" for CNNs. If they can, this would mean that there is—at least to some degree—a shared notion of image difficulty between humans and CNNs. We therefore conducted a psychophysical experiment, where subjects were asked to identify which image was easier for a neural network to classify. We found that human observers were able to do so well beyond chance (50%): on average, with an accuracy of 81.36%. The accuracies of the different subjects ranged from 71.81% to 88.59%, with a standard deviation of 6.29%. The mean error consistency between the subjects was 0.5874. For all combinations of different subjects, the error consistency ranged from 0.4053 to 0.7527, with a standard deviation of 0.08783. In conclusion, even naïve human observers without machine learning experience can reliably and consistently predict which images are easy and difficult for CNNs.

4 Discussion

We investigated the influence of inductive bias on model decisions. We found that model decisions are not only determined by the inductive bias — they are also largely influenced by the dichotomous difficulty of images in common datasets. (DDD): many images are either "trivial" or "impossible". This has implications for model design. Viewed positively, results for one network generalise towards other networks with different inductive biases, and if one desires replicable results across networks, this is an advantage. This is in line with previous findings that some results transfer between different model classes, e.g. adversarial attacks [13, 14] and for the comparison of networks and humans [7]. However, if models are trained on datasets with DDD, design decisions like architectural improvements may not be able to show their full potential since the resulting models, due to DDD, have a high likelihood of ending up in a very similar regime as other (already existing) models—and might even inherit their vulnerabilities.

Previous investigations found label errors to be a problem in a number of datasets. Here we show that issues with datasets go far beyond label errors. In order to be able to improve our ability to differentiate between models (and give their inductive bias a chance to truly make a difference), we will need datasets that are more balanced with respect to image difficulty or use only in-between images. This is far from trivial since we do not know precisely what causes DDD.

Object recognition also depends on its context and surroundings. Humans can recognise objects remarkably quickly [15], but this is only true if they are effectively segmented from their background [16]. As a result one can make a real-world dataset arbitrarily trivial (or impossible) for human observers by selecting prototypical (or non-prototypical) objects, showing them from canonical (or degenerate) viewpoints and have objects segmented from (or camouflaged by) the background. Perhaps it was naïve to believe that large automatically generated datasets would somehow "get the mix right" and result in images where the difficulty within and between categories is approximately the same—or at least not so large that it dominates over inductive bias as we show in this work.

Our human experiment shows that humans can reliably identify the impossible images from ImageNet (see Figure 12 in the Appendix for more examples). Inspection of those images left us with the impression that impossible images often contain multiple objects and sometimes "unusual" objects and viewpoints (see above). From a cognitive science or neuroscience perspective DDD might thus also provide new opportunities for insights: Perhaps the impossible images are the ones which can reveal differences between humans and CNNs and are thus those which neuroscience and cognitive science should be interested in. It may be beneficial to compare humans and brains to CNNs on images selected by their difficulty.

Acknowledgments & funding disclosure

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 (L.S.B, K.M. 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. The authors declare no competing interests.

We would like to thank Silke Gramer, Leila Masri and Sara Sorce for administrative and the "Cloud-masters" of the ML-Cluster at University Tübingen for technical support. We thank the group of Ludwig Schmidt at UC Berkeley for discussions during earlier stages of this project and David-Elias Künstle for helpful comments on the manuscript.

Author contributions

Motivated by previous work [7], the project was initialized by K.M. and jointly developed forward with R.G. and F.A.W. Later, L.S.B. joined this project but still at an early stage. L.S.B. wrote the code for computing and analysis and was supervised by K.M.; R.G .and F.A.W. gave input during the entire project. R.G. and F.A.W. had the idea of the psychophysical experiment. All authors planned, structured and wrote the manuscript; all figures were made by L.S.B. based on joint discussions.

References

[1] Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091*, 2020.

[2] Tom M. Mitchell. *The need for biases in learning generalizations*. 1980.

[3] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):1–74, 2021.

[4] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[5] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[6] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[7] Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33, 2020.

[8] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20 (1):37–46, 1960.

[9] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *arXiv preprint arXiv:2106.07411*, 2021.

[10] Curtis G. Northcutt, Lu Jiang, and Isaac L Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 2021.

[11] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.

[12] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From ImageNet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pages 9625–9635. PMLR, 2020.

[13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[14] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[15] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996.

[16] Felix A. Wichmann, Jan Drewes, Pedro Rosas, and Karl R. Gegenfurtner. Animal detection in natural scenes: Critical features revisited. *Journal of Vision*, 10(4):6–6, 2010.

[17] Katherine L. Hermann and Andrew K. Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *arXiv preprint arXiv:2006.12433*, 2020.

[18] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020.

[19] Liwei Wang, Lunjia Hu, Jiayuan Gu, Yue Wu, Zhiqiang Hu, Kun He, and John Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation. *arXiv preprint arXiv:1810.11750*, 2018.

[20] Katherine L. Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *arXiv preprint arXiv:1911.09071*, 2019.

[21] Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

[22] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[23] Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C. Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):1–12, 2020.

[24] Arash Akbarinia and Karl R. Gegenfurtner. Paradox in deep neural networks: Similar yet different while different yet similar. *arXiv preprint arXiv:1903.04772*, 2019.

[25] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. On the surprising similarities between supervised and self-supervised models. *arXiv preprint arXiv:2010.08377*, 2020.

[26] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

[27] Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. Model similarity mitigates test set overuse. *arXiv preprint arXiv:1905.12580*, 2019.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[29] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

[30] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[31] Kate Crawford and Trevor Paglen. Excavating AI: The politics of training sets for machine learning. `https://excavating.ai/`.

[32] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 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*, pages 547–558, 2020.

[33] Alex Krizhevsky, Geoffrey E. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[34] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[35] Duncan Riach. Determinism in deep learning (s9911). *GPU Technology Conference*, 2019.

[36] Lukas Muttenthaler and Martin N. Hebart. Thingsvision: a python toolbox for streamlining the extraction of activations from deep neural networks. *bioRxiv preprint bioRxiv:2021.03.11.434979*, 2021.

[37] Felix A. Wichmann and Frank Jäkel. *Methods in Psychophysics*, pages 1–42. John Wiley & Sons, Inc, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The central claim of our paper is that the data dichotomy dominates all inductive biases. We show this for various datasets.
   (b) Did you describe the limitations of your work? [Yes] See discussion.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We roughly used 250 GPU days for this paper. Each GPU unit on our cluster (together with CPU and RAM) consumes on average 300W. In total, this paper consumed 1800kWh. The CO2 emission in the country of the authors is roughly 400g/kW resulting in a CO2 equivalent of 720kg—this corresponds to roughly 45% of the emission of a flight from London to New York. We will compensate the amount of C02 with a certified CO2-compensation company. Furthermore, we will make sure that other researchers have access to the trained models. We can not distribute all models yet (several GBs) because of the size limit of the supplementary materials.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We ensure that we follow the ethics guidelines.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All code to reproduce our finding can be found in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See methods section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] We do not report errors bars but raw (distributional) data instead, e.g. see Figure 2.Here the overlap of multiple runs is very high (in fact not visible) so that we refrain from showing them.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See methods.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A] It is unclear which license the dataset have as CIFAR does not have any information on their web page.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Yes all code is in the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] ImageNet and CIFAR-100 are publicly available from the curators.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Recently, some issues around ImageNet were discussed e.g. by `https://www.excavating.ai`. Thus, we removed some images in our psychophysical experiment and do not show any images containing humans in this paper.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See methods.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] No potential participants risks in our experiment.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See methods.

A Appendix

A.1 Additional Figures

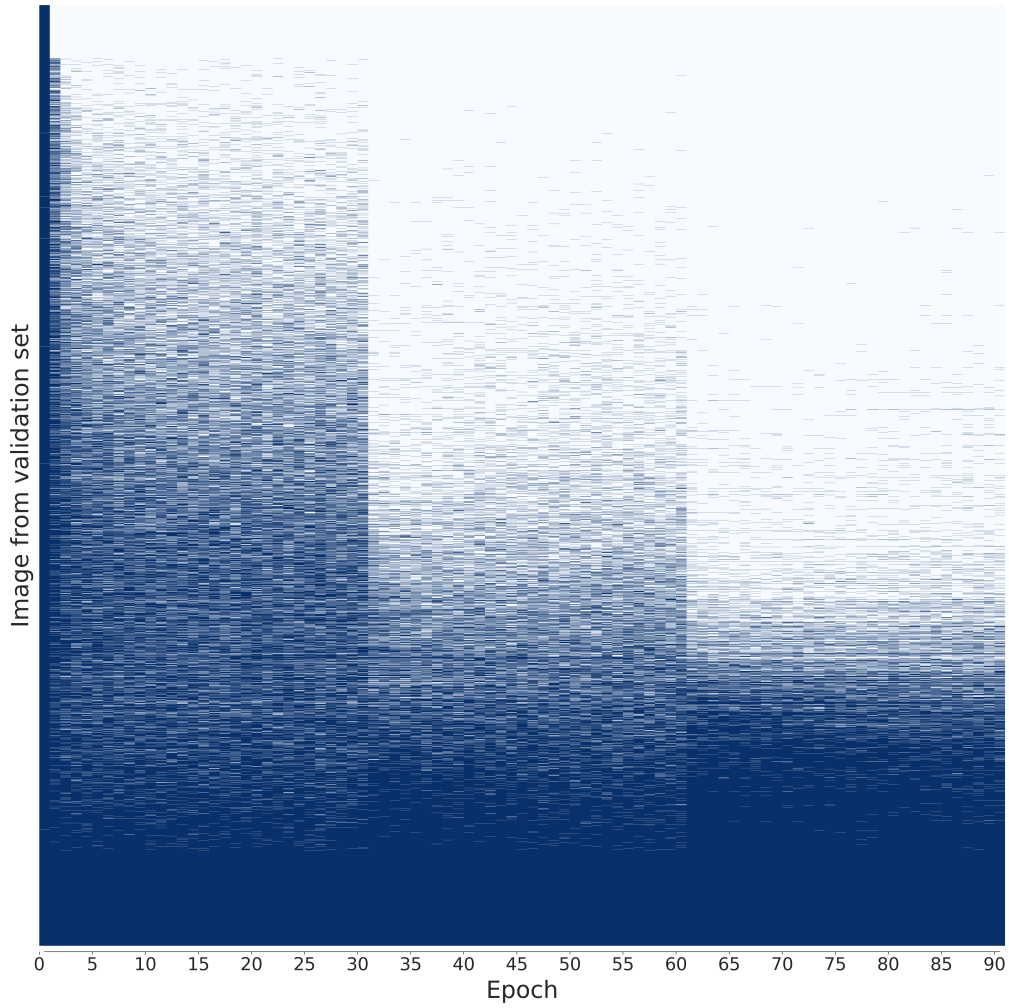


Figure 4: Decisions on all 50K ImageNet validation images of the *single* base network over the epochs. Blue indicates that the respective item was falsely classified during the specific epoch, while white indicates that it was correctly classified. The items from the ImageNet validation set are ordered 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.

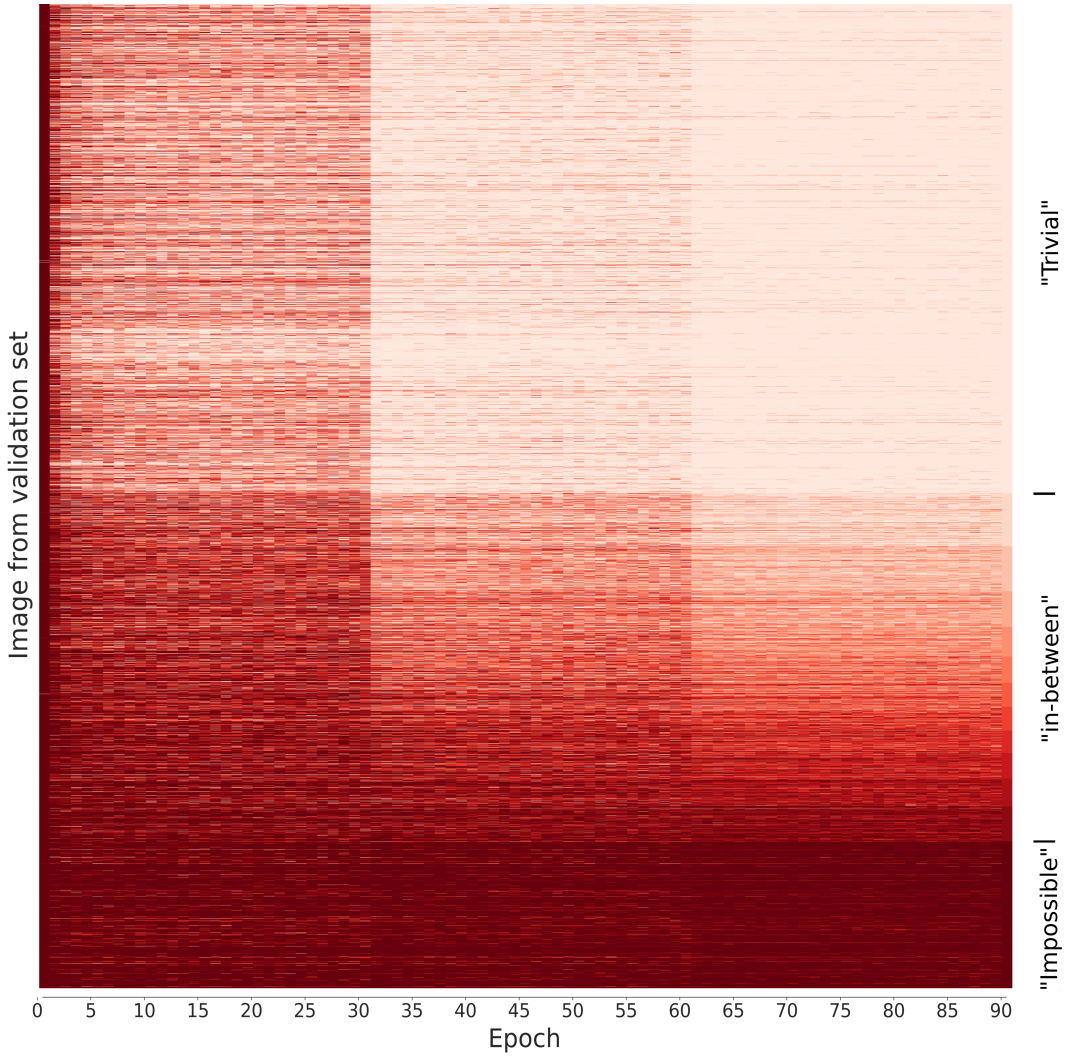


Figure 5: Decisions on all 50K ImageNet validation images of *all* 13 networks with different inductive biases (architectures, ...). Dark red indicates that the respective item was falsely classified by all networks. Light red indicates that the image was correctly classified by all networks. Images are ordered according to the mean accuracy across networks in the last epoch.

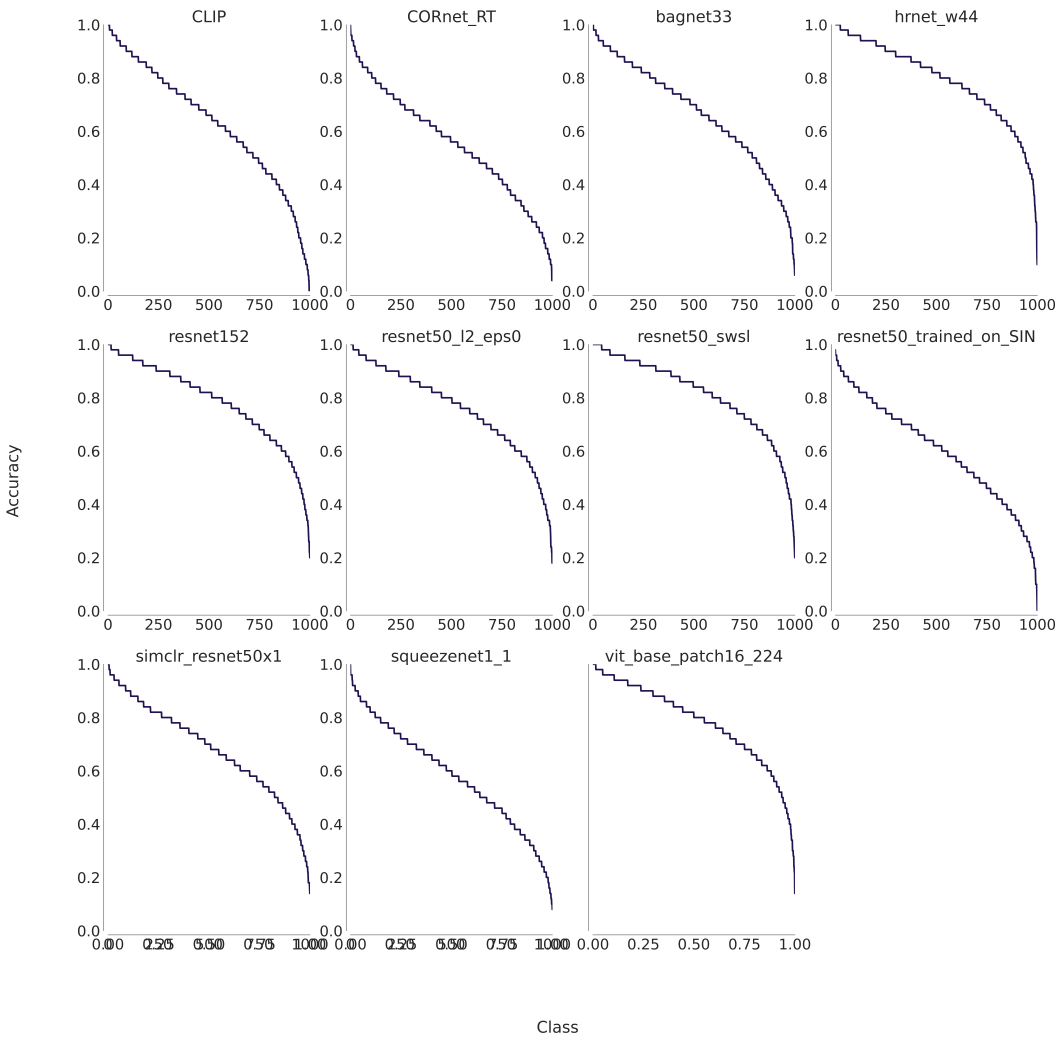


Figure 6: Class-wise accuracy per model. Shown is the decreasing accuracy for all classes in the validation sets.

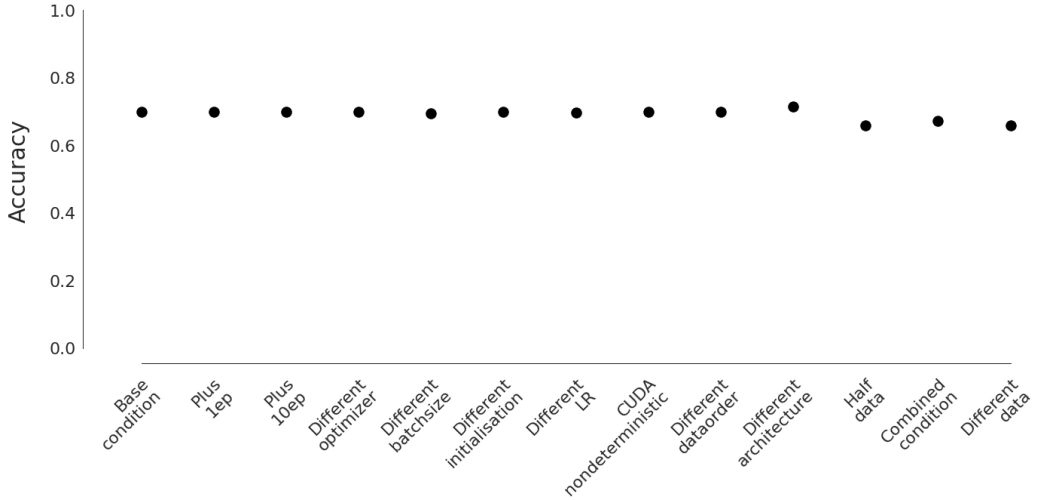


Figure 7: Accuracies of the different conditions and the base network on the ImageNet validation set after 90 epochs. The mean over all models of a condition is displayed here.

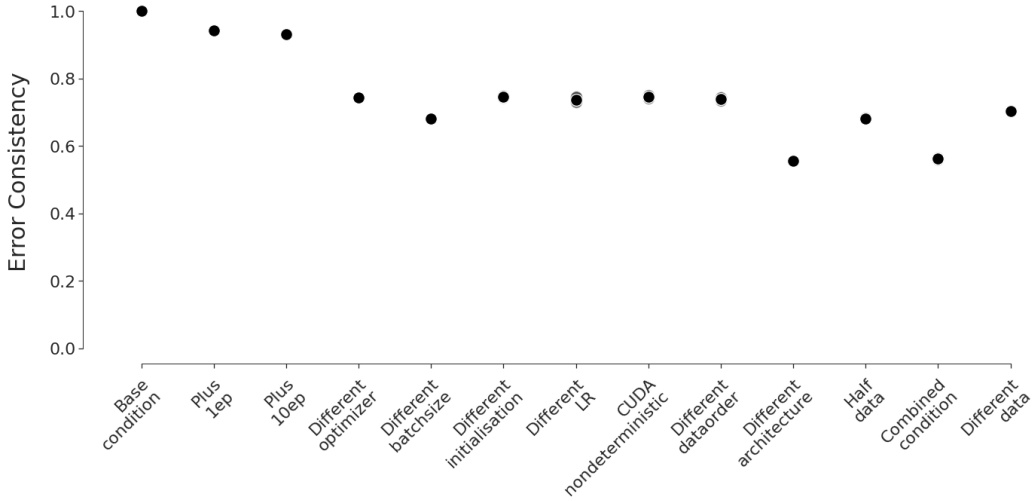


Figure 8: Error consistencies on the ImageNet validation set with VGG-11 as the base network. All variations performed are the same as outlined in the Methods section. In this case, the different architecture is an AlexNet. The conditions are ordered by the mean error consistency on the ImageNet validation set for ResNet-18 as the base network (see Figure 2). For conditions in which multiple models were trained, the model-wise error consistencies are plotted with a lower opacity compared to the mean over all models for the conditions.

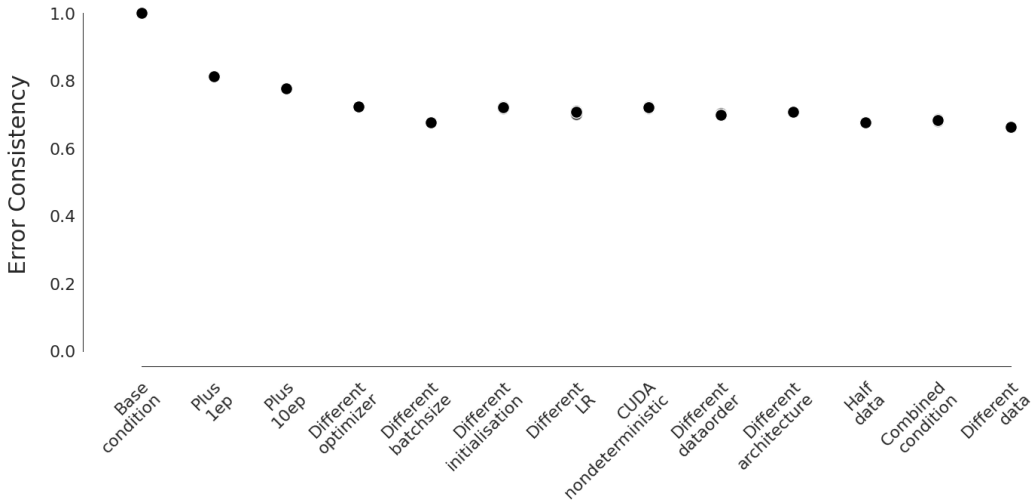


Figure 9: Error consistencies on the ImageNet validation set with DenseNet-121 as the base network. All variations performed are the same as outlined in the Methods section. In this case, the different architecture is a ResNet-50. The conditions are ordered by the mean error consistency on the ImageNet validation set for ResNet-18 as the base network (see Figure 2). For conditions in which multiple models were trained, the model-wise error consistencies are plotted with a lower opacity compared to the mean over all models for the conditions.

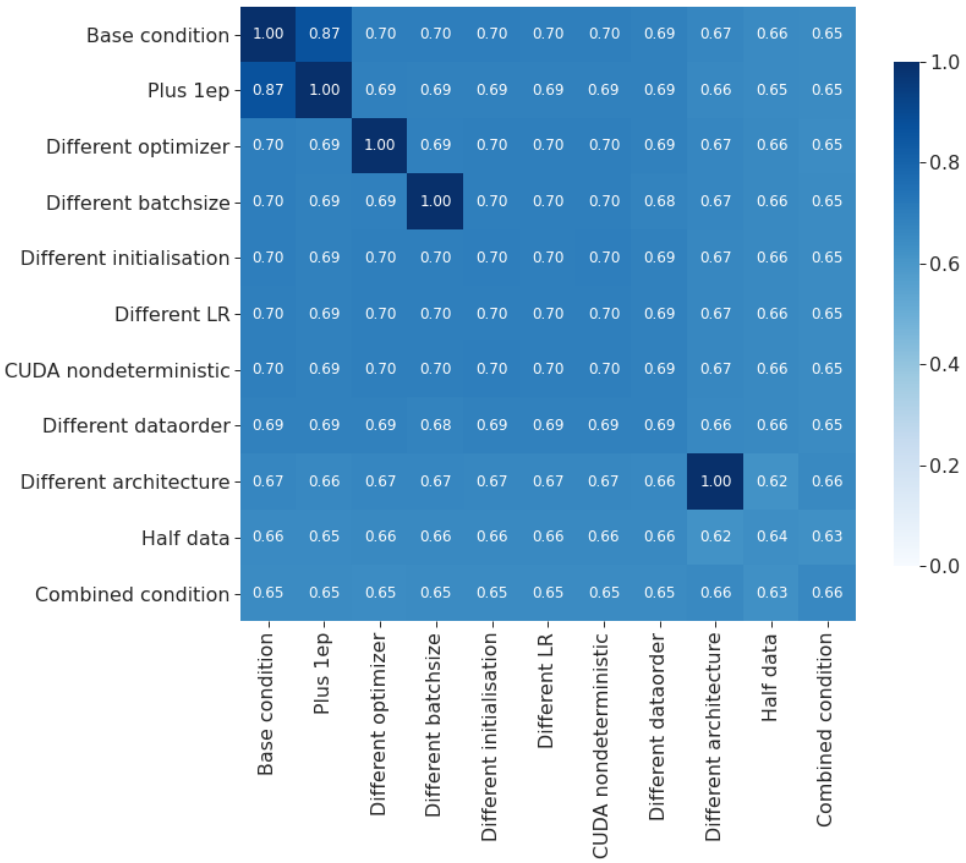


Figure 10: Error consistencies between all conditions on the ImageNet validation set with ResNet-18 as the base network.

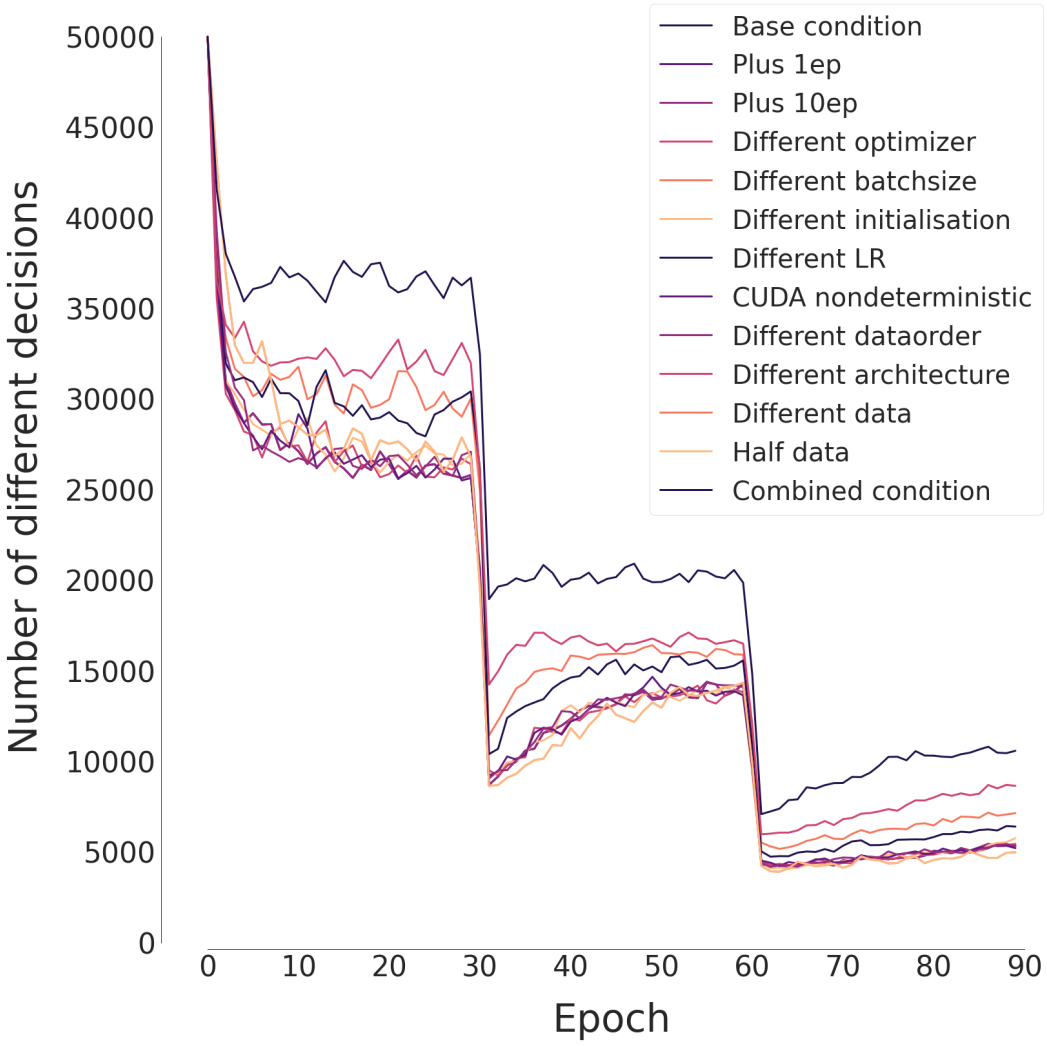


Figure 11: Lineplot showing the number of decisions that change from the current to the following epoch. For epoch 0, this means that the number of decisions that are different between epoch 0 and epoch 1 are shown. For conditions in which multiple model instances were trained, only the last instance is shown for the sake of simplicity.

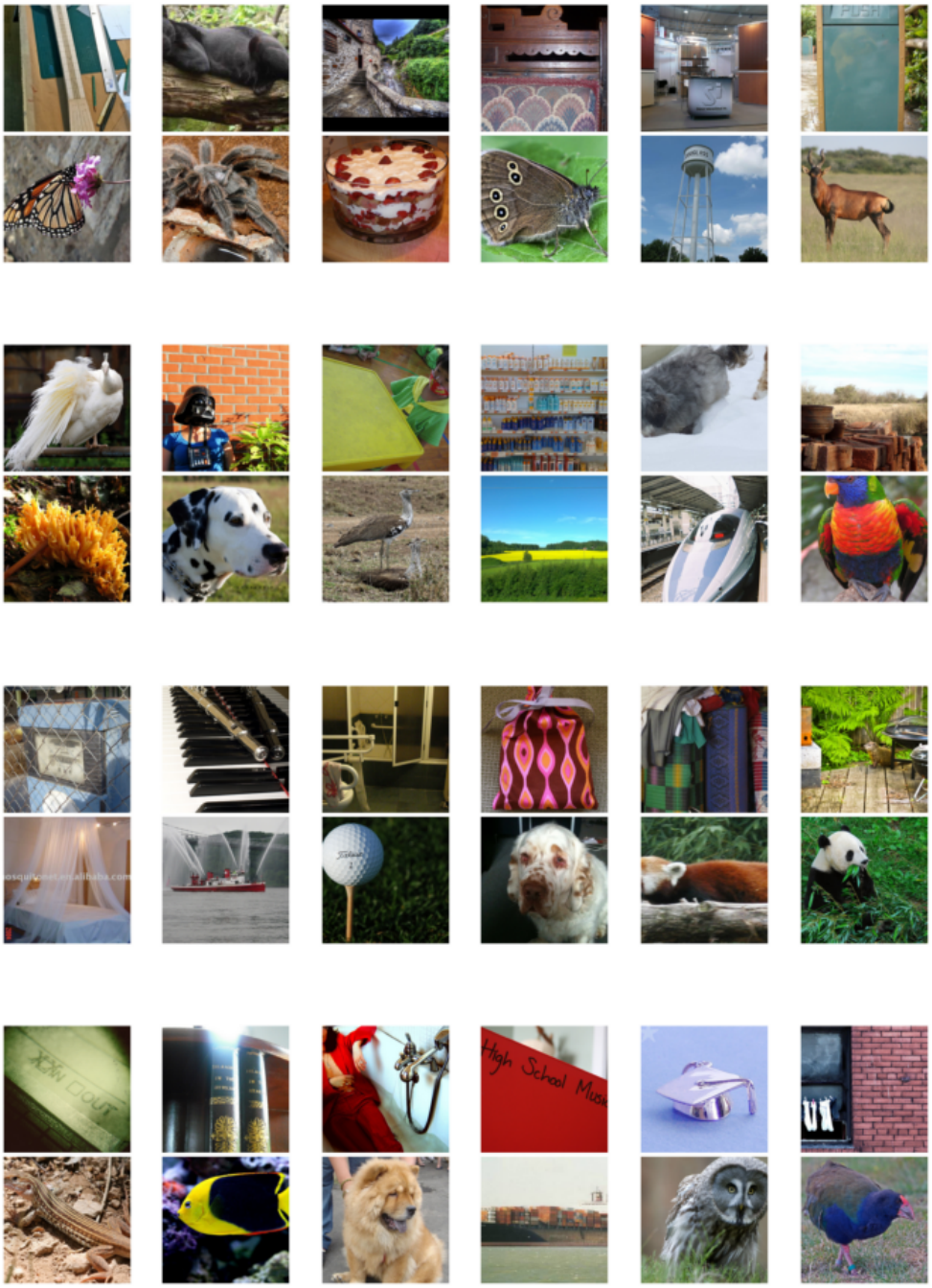


Figure 12: Pairs of impossible (top) and trivial images (bottom) from ImageNet.

B Related work

Metrics for CNN comparisons Given the scientific, practical and engineering implications of model inductive biases, it is not entirely surprising that a number of studies investigated differences between neural networks. For this purpose, the standard metric is accuracy. Some studies also focus on learned features and decision boundaries [e.g. 17–20], or internal representations [21, 22]. Using representational similarity analysis (RSA) and most similar to our work, [23, 24] investigated whether different CNNs yield correlated representations and found that many neural networks show differences on a representational level. How intermediate representations are related to classification behaviour largely remains unclear. In order to compare networks on a behavioural level directly, metrics such as *error consistency* can be used. Error consistency (measured by κ) assesses the degree of agreement between two decision-makers [7, 25].

Consistent model errors Tramèr et al. [26] observe that the decision boundaries of two models are highly similar, an issue that is related to the transferability of adversarial examples between models. Additionally, it has been shown that standard vanilla models systematically agree on their errors both on IID (independent and identically distributed) data [27] and OOD (out-of-distribution) data [7]. It is unclear whether, if at all, there is a connection between model inductive bias, dataset difficulty and consistent model errors.

Problems of ImageNet The ImageNet dataset [28] has numerous issues. Next to those affecting most datasets—such as dataset bias [6]—a number of problems have been identified. Very recently, Northcutt et al. [29] showed that around 6% of ImageNet validation images suffer from label errors. Additionally, many images simply require more than a single label since multiple objects are present, and the distinctions between classes seem rather arbitrary at times [11, 12]. Even when trying to replicate the original ImageNet labeling procedure in order to create a new test set, models trained on ImageNet have an accuracy drop of 11–14% on this new test set [30]. Finally, ImageNet labels are based on the WordNet hierarchy, which contains many problematic categories. For instance, many categories in the "person" subtree have labels ranging from outdated to outrageous and racist [31, 32].

B.1 Dichotomous data difficulty similarly afflicts other datasets, not just ImageNet

Is dichotomous data difficulty (DDD) only a problem for ImageNet? We here show that this is not the case on two different datasets. CIFAR-100 [33] and the third dataset ("Gaussian noise") was generated by ourselves to investigate the effect of training on a dataset that does not contain any "natural image structure". It was generated by drawing pixel-wise uncorrelated Gaussian noise for each of the three RGB-channels. The dataset consisted of 100 classes with 20000 train and 50 test images per class. The i-th class has a mean of 128 and a standard deviation of $\sigma = i$, which is how classes can be identified by a model.

Is DDD also present in CIFAR-100 and even a synthetic Gaussian dataset? As a first indication, for both of these datasets we find similarly high error consistencies between different models, just like we found for ImageNet (see Figure 13).

If our hypothesis is true that dichotomous data difficulty is underlying this pattern (rather than, e.g., idiosyncrasies of natural images), we should be able to replicate this result pattern for an entirely artificial dataset where we introduce easy and difficult images by design, but otherwise use random noise. To this end, we constructed a dataset consisting entirely of Gaussian noise, where each of the 100 classes has a fixed standard deviation, i.e. there are very easy and very hard classes.[3] And indeed, training models on this Gaussian data set leads to a very similar result pattern as for natural data sets like ImageNet and CIFAR-100 (shown in panel (b) and (c) of Figure 14). This is a strong indication —together with the imbalanced class accuracies in Figure 15— that highly consistent model errors are a result of DDD and not an artefact of natural images.

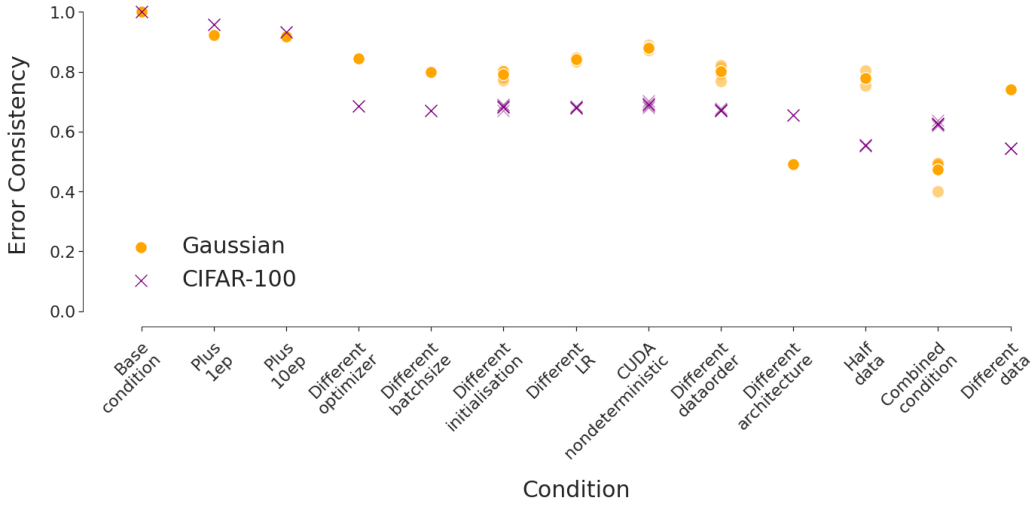
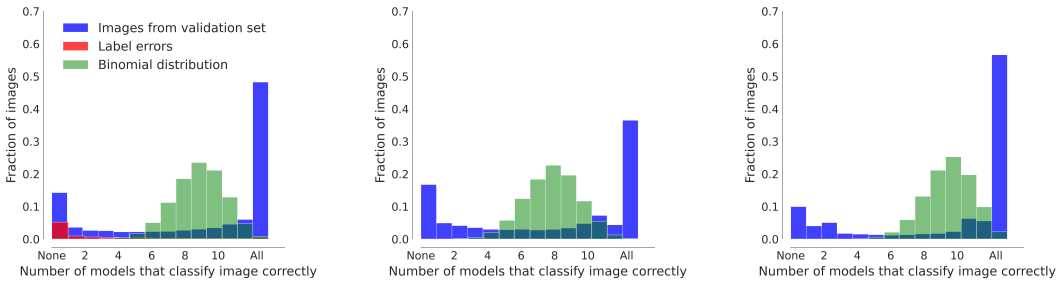


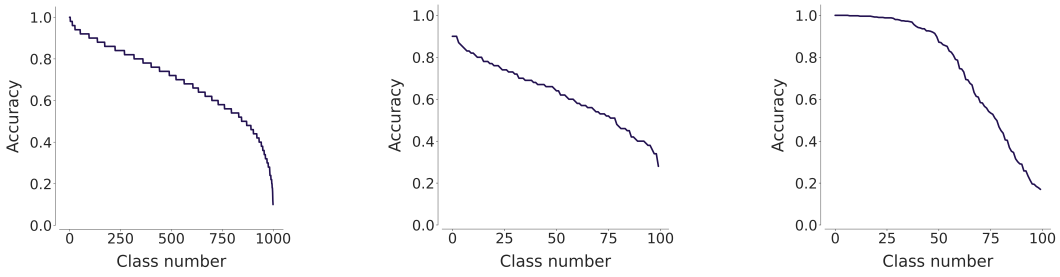
Figure 13: Error consistencies between the different conditions and the base network for the validation sets of CIFAR-100 and our Gaussian dataset. The conditions are ordered by the mean error consistency on the ImageNet validation set (see Figure 2). For conditions in which multiple models were trained, the model-wise error consistencies are plotted with a lower opacity compared to the mean over all models for the conditions.

[3] We constructed the dataset with a decreasing KL-divergence between classes. Thus some classes are easier than others. In fact, we show in the Appendix (Figure 18) that the KL divergence is a very good predictor for class accuracies.



(a) ImageNet (b) CIFAR-100 (c) Gaussian

Figure 14: Histogram showing how many models correctly classify validation sets images in the last epoch. In blue, the densities of how many items were answered correctly are shown. "None" indicates that no models classified the items correctly (impossibles), while for "All" items were classified correctly by all models ("trivial images"). For the sake of simplicity, only the last model was used for conditions where multiple models were trained. In green, samples are drawn from a binomial distribution with n equal to the number of models and p equal to the mean accuracy over the models. Additionally for ImageNet, the distribution of 5000 label errors as identified by the cleanlab package are shown in red [10].



(a) ImageNet (b) CIFAR-100 (c) Gaussian

Figure 15: Class-wise accuracy per dataset. Shown is the decreasing accuracy for all classes in the validation sets and for the fully trained base network.

B.2 Dataset subsampling according to Dichotomous Data Difficulty reveals differences between models

So far we have seen that models agree despite markedly different choices of architecture, training objectives, and many other aspects. While we hypothesized DDD (a dataset issue) to be the cause, an alternative explanation would be that models simply agree irrespective of the choice of data difficulty. In order to differentiate between these two competing hypotheses we performed an experiment where we removed both the "trivial" and the "impossible" images from the validation dataset. If model agreement is indeed caused by DDD, then we should find much stronger differences between different models (as indicated through lower error consistency scores). The results are presented in Figure 16: Indeed, model differences are now much more pronounced, in many cases the consistency between different models even approaches zero, indicating that some networks make truly independent decisions, i.e. have learned independent decision boundaries whilst being similarly accurate. This shows that the high agreement between different models (as observed e.g. by [7, 9] and [27]) is a result of dataset DDD problems, not that inductive bias does not matter much. Please note that the reduced consistency is not trivially caused by the removal of impossible and trivial images: Even when removing extreme images (all models correct/incorrect), two models could agree or disagree on the remaining images of intermediate difficulty (error consistency is calculated pair-wise).

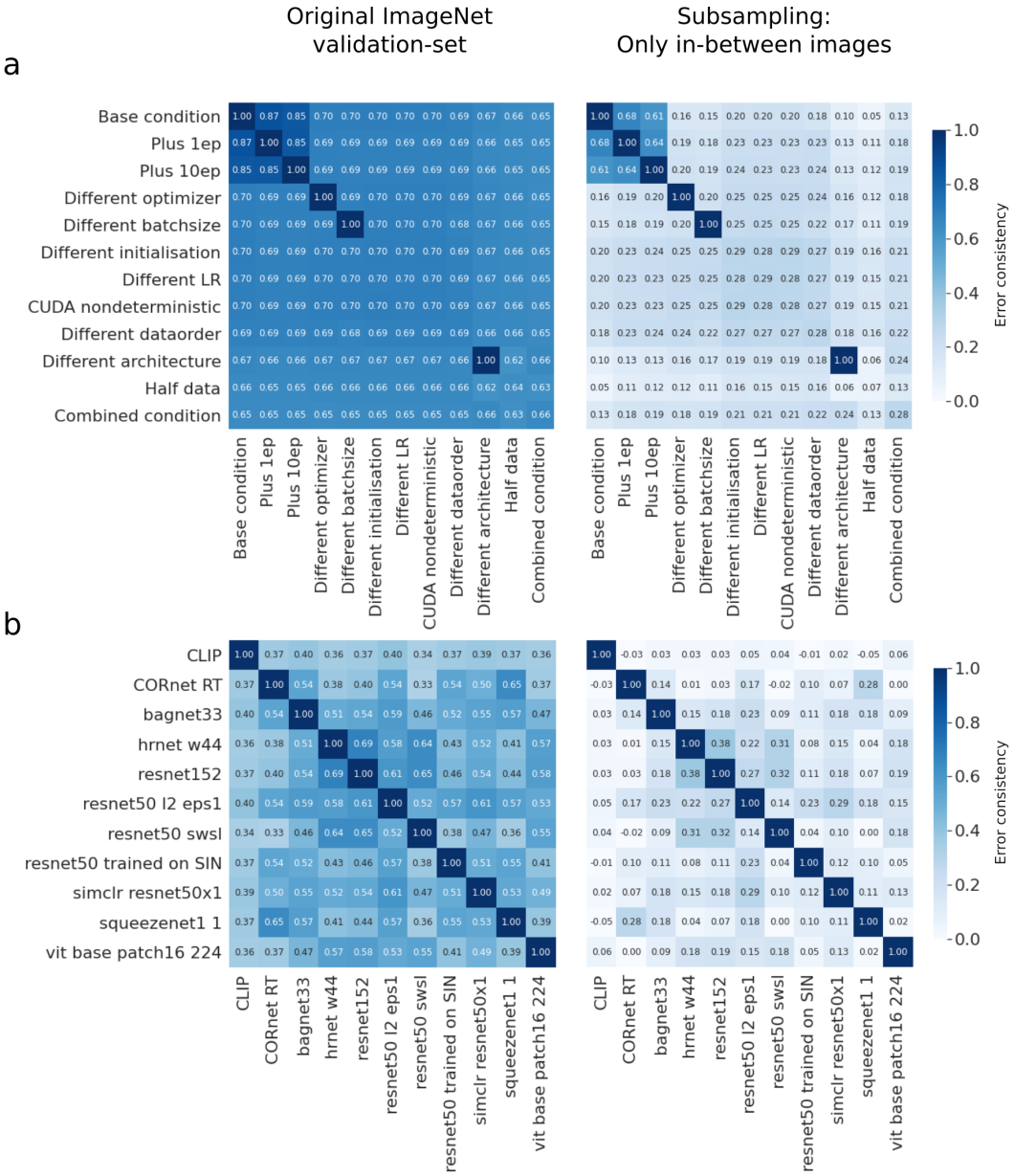


Figure 16: Error consistency on the original ImageNet test-set (left panel) and on the in-between images only (right panel) for the ResNet-variants (a) and the SOTA networks (b). Error consistency around 0 indicates independent responses. A diagonal element of 1 represents that only one network for comparison was available, otherwise the within condition consistency is calculated, see Section 2.

B.3 Variations of inductive biases

Our systematic variations are:

- *Base condition*: a standard ResNet-18 trained on ImageNet in PyTorch[4] was used as the baseline network for all comparisons; one instance trained.
- *Plus 1ep*: a network was trained for one additional epoch compared to the base network; one instance trained.
- *Plus 10ep*: a network was trained for ten additional epochs compared to the base network; one instance trained.
- *Different optimizer*: a network was trained using SGD with Nesterov momentum[34] instead of vanilla SGD; one instance trained.
- *Different batch size*: for this condition we split the batch size in half (128 instead of the 256). This was done by drawing the same batches from the data loader and splitting them in half. We then input the halves sequentially into the model, effectively doubling the number of gradient updates; one instance trained.
- *Different initialisation*: networks were varied in the initialisation of their layer weights by choosing a different random seed for each network; five instances trained.
- *Different learning rate*: the networks were trained using initial learning rates varying from 0.148 to 0.152 instead of the default learning rate of 0.1. We narrowed the range such that they still reach the same accuracy level; five instances trained.
- *CUDA non-deterministic*: Training networks without CUDA determinism is the standard procedure. However, graphic card operations are not necessarily deterministic, e.g. functions like `reduce_sum` [35]. This non determinism might not influence accuracy but may influence agreement between instances; five instances trained.
- *Different dataorder*: networks were trained with the exact same training data, however the order of the samples was varied for each model by choosing a different random seed before initialisation of the data loader; five instances trained.
- *Different architectures*: we trained a DenseNet-121 as a different architecture. Due to hardware constraints, we had to use a batch size of 64 for this condition; one instance trained.
- *Half data*: the network was trained on only half of the data but compared to the base condition with all data; one instance trained.
- *Combined condition*: for this condition, we combined multiple conditions. Here, we trained networks of the different architecture condition with training data in a different order, using SGD with Nesterov momentum, varying learning rates from 0.148 to 0.152, and different initialisations for each network; 5 instances trained.
- *Different data*: two networks were trained on the first and second half of the ImageNet training set respectively. Thus two different, *disjoint* training datasets were used—of course from the same distribution (ImageNet). For this condition we compared the networks to each other instead of comparing against the base condition.

[4] See `https://github.com/pytorch/examples/tree/master/imagenet`: batch size of 256, 90 epochs, the SGD optimizer and an initial learning rate of 0.1 that was divided by 10 every 30 epochs.

B.4 Software, hardware and psychophysical experiment

Software, Hardware and data The networks were trained on GeForce RTX 2080 Ti GPUs with CUDA Version 11.1, CPU cores and 32 GB RAM shared between the cores. All code was written in PyTorch using Python 3 and the code to reproduce our findings is available in the supplementary material. For the RSA analysis, we used the thingsvision toolkit [36]. We used three data sets: ImageNet [28], CIFAR-100 [33] and the third dataset ("Gaussian noise") was generated by ourselves to investigate the effect of training on a dataset that does not contain any "natural image structure". It was generated by drawing pixel-wise uncorrelated Gaussian noise for each of the three RGB-channels. The dataset consisted of 100 classes with 20000 train and 50 test images per class. The i-th class has a mean of 128 and a standard deviation of $\sigma = i$, which is how classes can be identified by a model.

Psychophysical experiment In order to test whether humans can infer which images are easy and hard for CNNs, we conducted a psychophysical two-alternative forced choice experiment [37]. In the experiment, observers were instructed to indicate by button press which image of an image pair they believe to be more difficult for a network to classify correctly. Images were chosen from the ImageNet validation set such that the image pairs consisted of one image which all networks with different inductive bias classified correctly and another image which all networks misclassified (see also Figure 12). Stimuli were non-normalized images of size 224×224 px. Observers performed 149 self-paced trials. Overall, nine observers (mean age = 34.6yrs, 2 female, 7male) participated. Two observers were entirely naïve to CNN research, a further four were naïve to the purpose of the experiments, but knew about CNNs. Subjects received monetary compensation of 10 € per hour. The total duration of the experiment was 30 minutes.

B.5 Additional Networks

Control experiments. To ensure that our findings generalize across different architectures and different datasets, we reran our main experiment with a number of variations:

First, we tested different architectures; *ImageNet with Densenet-121 as base network*: Using a Denenet-121 as base network with a slightly altered training paradigm using only 30 instead of 90 epochs—to reduce the environmental impact of our study—and a batch size of 64 due to GPU RAM limitations. A ResNet-50 was used as comparison architectures.

Imagenet with VGG-11 as base network: For VGG-11 as the base network, we used a starting learning rate of 0.1 as according to the standard PyTorch implementation. Again, we only trained networks in this paradigm for 30 epochs and with learning rate steps every 10 epochs. Additionally, we used an AlexNet as different architecture.

Second, in addition to ImageNet and our Gaussian dataset we used another dataset, namely *CIFAR-100*: Again, we followed the standard ResNet-18 PyTorch implementation with the modification that we only used a total of 30 epochs to reduce the environmental impact of our study.

B.6 Control experiment Representational Similarly Analysis

Additionally, to check whether our results are reproducible outside of a behavioural measure, we applied the tool representational similarly analysis (RSA). RSA is a method that originated in the brain sciences. It quantifies whether the inner representation—here the activation of kernels by single images—is similar across networks [21, 23]. An RSA between two networks yields a correlation index between -1 and 1, indicating anti-correlation, no correlation (0) and perfect correlation respectively. It is important to note that the correlation values from RSA and κ from error consistency are not comparable, although they have the same limits.

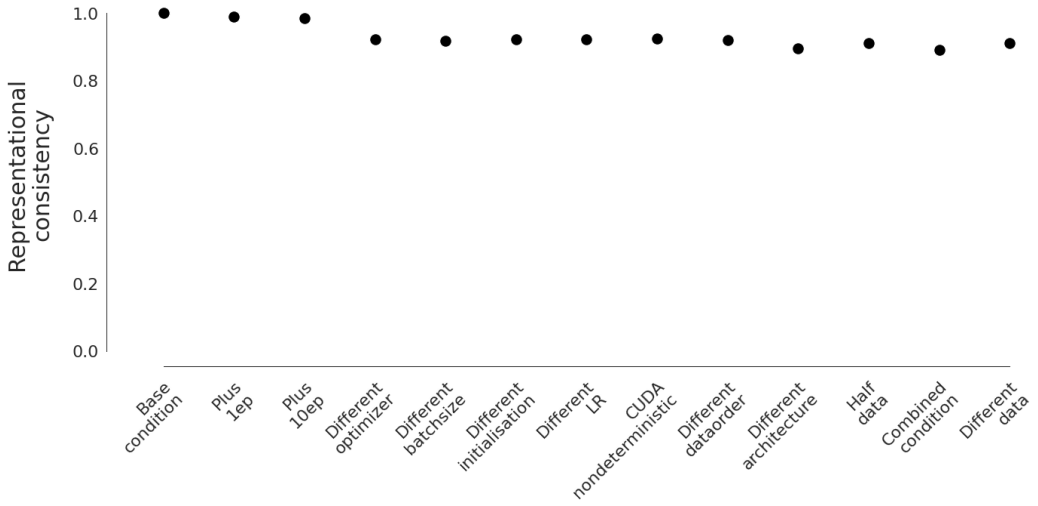


Figure 17: Correlations between the last fully connected layers of the different conditions and the base network on the ImageNet validation set after 90 epochs. For conditions in which multiple models were trained, only the first model was used.

B.7 KL divergence

We constructed a third dataset ("Gaussian noise"). It was generated by drawing pixel-wise uncorrelated Gaussian noise for each of the three RGB-channels. The dataset consisted of 100 classes with 20000 train and 500 test images per class. The i-th class has a mean of 128 and a standard deviation of $\sigma = i$, which is how classes can be identified by a ML model. With this procedure, the KL-Divergence

$$KL(Class_i, Class_{i+1}) = log(\frac{\sigma_{i+1}}{\sigma_i}) + \frac{\sigma_i^2}{2 \cdot \sigma_{i+1}^2} - \frac{1}{2} \tag{1}$$

between class i and $i-1$ is decreasing, see Figure 18.

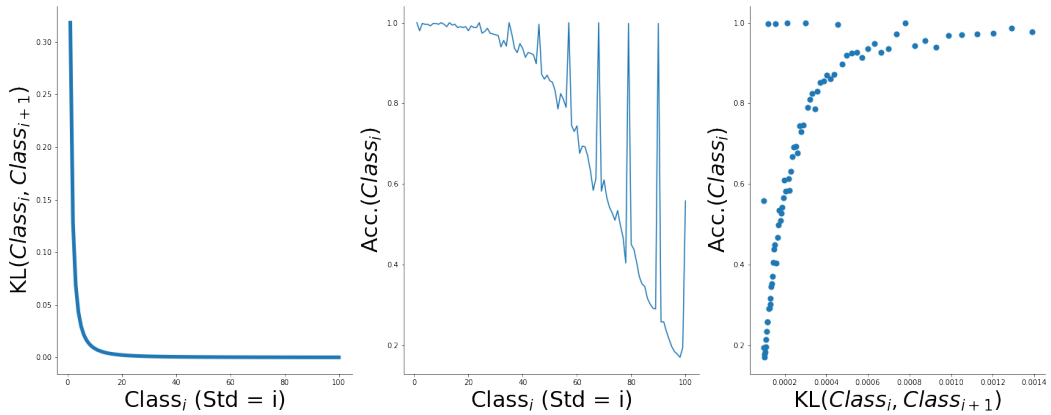


Figure 18: Kl-Divergence vs. accuracy for the Gaussian dataset. (Left) KL-divergence between Class$_i$ and Class$_{i+1}$. (Centre) Acc. of Class$_i$. (Right) Scatterplot between KL-divergence and accuracy. For the last plot we skip the first 20 classes (with accuracy close to 1) for better visibility.

B.8 ImageNet classes

Highest accuracy	Lowest accuracy
'earthstar'	'screen, CRT screen'
'yellow lady's slipper, yellow lady-slipper, Cypripedium calceolus, Cypripedium parviflorum'	'velvet'
'proboscis monkey, Nasalis larvatus'	'sunglass'
'Leonberg'	'ladle'
'freight car'	'tiger cat'
'echidna, spiny anteater, anteater'	'notebook, notebook computer'
'African hunting dog, hyena dog, Cape hunting dog, Lycaon pictus'	'hook, claw'
'limpkin, Aramus pictus'	'cleaver, meat cleaver, chopper'
'hamster'	'letter opener, paper knife, paperknife'
'three-toed sloth, ai, Bradypus tridactylus'	'spatula'

Table 1: Table displaying the ten classes, for which the base network achieved with highest and lowest accuracies respectively. Items are in a descending order, so that 'earthstar' has the highest accuracy and 'screen, CRT screen' has the lowest accuracy.