natureresearch

https://doi.org/10.1038/s42256-020-00257-z

Supplementary information

Shortcut learning in deep neural networks

In the format provided by the
authors and unedited

Supplementary information

A Shortcut learning across deep learning

Taken together, we have seen how shortcuts are based on dataset shortcut opportunities and discriminative feature learing that result in a failure to generalise as intended. We will now turn to specific application areas, and discover how this general pattern appears across Computer Vision, Natural Language Processing, Agent-based (Reinforcement) Learning and Fairness / algorithmic decision-making. While shortcut learning is certainly not limited to these areas, they might be the most prominent ones where the problem has been observed.

Computer Vision To humans, for example, a photograph of a car still shows the same car even when the image is slightly transformed. To DNNs, in contrast, innocuous transformations can completely change predictions. This has been reported in various cases such as shifting the image by a few pixels [49], rotating the object [48], adding a bit of random noise or blur [108, 50, 71, 109] or (as discussed earlier) by changing background [9] or texture while keeping the shape intact [38] (see Figure 4 for examples). Some key problems in Computer Vision are linked to shortcut learning. For example, transferring model performance across datasets (*domain transfer*) is challenging because models often use domain-specific shortcut features, and shortcuts limit the usefulness of unsupervised representations [110]. Furthermore, *adversarial examples* are particularly tiny changes to an input image that completely derail model predictions [8] (an example is shown in Figure 4). Invisible to the human eye, those changes modify highly predictive patterns that DNNs use to classify objects [34]. In this sense, adversarial examples—one of the most severe failure cases of neural networks—can at least partly be interpreted as a consequence of shortcut learning.

Natural Language Processing The widely used language model BERT has been found to rely on superficial cue words. For instance, it learned that within a dataset of natural language arguments, detecting the presence of "not" was sufficient to perform above chance in finding the correct line of argumentation. This strategy turned out to be very useful for drawing a conclusion without understanding the content of a sentence [15]. Natural Language Processing suffers from very similar problems as Computer Vision and other fields. Shortcut learning starts from various dataset biases such as annotation artefacts [111, 52, 112, 113]. Feature combination crucially depends on shortcut features like word length [91, 114, 15, 115], and consequently leads to a severe lack of robustness such as an inability to generalise to more challenging test conditions [116, 117, 118, 119]. Attempts like incorporating a certain degree of unsupervised training as employed in prominent language models like BERT [5] and GPT [120] did not resolve the problem of shortcut learning [15], even though it does improve few-shot generalisation [121].

Agent-based (Reinforcement) Learning Instead of learning how to play Tetris, an algorithm simply learned to pause the game to evade losing [122]. Systems of Agent-based Learning are usually trained using Reinforcement Learning and related approaches such as evolutionary algorithms. In both cases, designing a good reward function is crucial, since a reward function measures how close a system is to solving the problem. However, they all too often contain unexpected shortcuts that allow for so-called *reward hacking* [123]. The existence of loopholes exploited by machines that follow the letter (and not the spirit) of the reward function highlight how difficult it is to design a shortcut-free reward function [98]. Reinforcement Learning is also a widely used method in Robotics, where there is a commonly observed *generalisation* or *reality gap* between simulated training environment and real-world use case. This can be thought of as a consequence of narrow shortcut learning by adapting to specific details of the simulation. Introducing additional variation in colour, size, texture, lighting, etc. helps a lot in closing this gap [124, 125].

Fairness & algorithmic decision-making Tasked to predict strong candidates on the basis of their résumés, a hiring tool developed by Amazon was found to be biased towards preferring men. The model, trained on previous human decisions, found gender to be such a strong predictor that even removing applicant names would not help: The model always found a way around, for instance by inferring gender from all-woman college names [13]. This exemplifies how some—but not all—problems of (un)fair algorithmic decision-making are linked to shortcut learning: Once a predictive feature is found by a model, even if it is just an artifact of the dataset, the model's decision rule may depend entirely on the shortcut feature. When human biases are not only replicated, but worsened by a machine, this is referred to as *bias amplification* [126]. Other shortcut strategies include focusing on the majority group in a dataset while accepting high error rates for underrepresented groups [12, 127], which can amplify existing societal disparities and even create new ones over time [128]. In the dynamical setting a related problem is called *disparity amplification* [128], where sequential feedback loops may amplify a model's reliance on a majority group. It should be emphasised, however, that fairness is an active research area of machine learning closely related to invariance learning that might be useful to quantify and overcome biases of both machine and human decision making.

B Beyond shortcut learning

A lack of out-of-distribution generalisation can be observed all across machine learning. Consequently, a significant fraction of machine learning research is concerned with overcoming shortcut learning, albeit not necessarily as a concerted effort. Here we highlight connections between different research areas. Note that an exhaustive list would be out of the scope for this work. Instead, we cover a diverse set of approaches we find promising, each providing a unique perspective on learning beyond shortcut learning.

Domain-specific prior knowledge Avoiding reliance on unintended cues can be achieved by designing architectures and data-augmentation strategies that discourage learning shortcut features. If the orientation of an object does not matter for its category, either data-augmentation or hard-coded rotation invariance [129] can be applied. This strategy can be applied to almost any well-understood transformation of the inputs and finds its probably most general form in auto-augment as an augmentation strategy [130]. Extreme data-augmentation strategies are also the core ingredient of the most successful semi-supervised [131] and self-supervised learning approaches to date [132, 133].

Adversarial examples and robustness Adversarial attacks are a powerful analysis tool for worst-case generalisation [8]. Adversarial examples can be understood as counterfactual explanations, since they are the smallest change to an input that produces a certain output. Achieving counterfactual explanations of predictions aligned with human intention makes the ultimate goals of adversarial robustness tightly coupled to causality research in machine learning [134]. Adversarially robust models are somewhat more aligned with humans and show promising generalisation abilities [135, 136]. While adversarial attacks test model performance on model-dependent worst-case noise, a related line of research focuses on model-independent noise like image corruptions [108, 71]. Irrespective of the type of attack (e.g., adversarial attacks, image degradations or other out-of-distribution tests), model robustness is best assessed by a broad range of tests. In order to prevent models from taking a narrow defence strategy that does not generalise, evolving and adaptive tests may be necessary [137].

Domain adaptation, -generalisation and -randomisation These areas are explicitly concerned with out-of-distribution generalisation. Usually, multiple distributions are observed during training time and the model is supposed to generalise to a new distribution at test time. Under certain assumptions the intended (or even causal) solution can be learned from multiple domains and environments [138, 39, 100]. In robotics, domain randomisation (setting certain simulation parameters randomly during training) is a very successful approach for learning policies that generalise to similar situations in the real-world [124].

Fairness Fairness research aims at making machine decisions "fair" according to a certain definition [139]. Individual fairness aims at treating similar individuals similarly while group fairness aims at treating subgroups no different than the rest of the population [140, 141]. Fairness is closely linked to generalisation and causality [142]. Sensitive group membership can be viewed as a domain indicator: Just like machine decisions should not typically be influenced by changing the domain of the data, they also should not be biased against minority groups.

Meta-learning Meta-learning seeks to learn how to learn. An intermediate goal is to learn representations that can adapt quickly to new conditions [143, 144, 145]. This ability is connected to the identification of causal graphs [146, 147] since learning causal features allows for small changes when changing environments.

Generative modelling and disentanglement Learning to generate the observed data forces a neural network to model every variation in the training data. By itself, however, this does not necessarily lead to representations useful for downstream tasks [148], let alone out-of-distribution generalisation. Research on disentanglement addresses this shortcoming by learning generative models with well-structured latent representations [149]. The goal is to recover the true generating factors of the data distribution from observations [150] by identifying independent causal mechanisms [134].

C Method details of toy example

The code to reproduce our toy example (Figure 2) is available from https://github.com/rgeirhos/shortcut-perspective. Two easily distinguishable shapes (star and moon) were placed on a 200×200 dimensional 2D canvas. The training set is constructed out of 4000 images, where 2000 contain a star shape and 2000 a moon shape. The star shape is randomly placed in the top right and bottom left quarters of the canvas, whereas the moon shape is randomly placed in the top left and bottom right quarters of the canvas. At test time the setup is nearly identical, 1000 images with a star and 1000 images with a moon are presented. However, this time the position of star and moon shapes are randomised over the full canvas, i.e. in test images stars and moons can appear at any location.

We train two classifiers on this dataset: a fully connected network as well as a convolutional network. The classifiers are trained for five epochs with a batch size of 100 on the training set and evaluated on the test set. The training objective is standard crossentropy loss and the optimizer is Adam with a learning rate of 0.00001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 1e-08$. The fully connected network was a three-layer ReLU MLP (multilayer perceptron) with 1024 units in each layer and two output units corresponding to the two target classes. It reaches 100% accuracy at training time and approximately chance-level accuracy at test time (51.0%). The convolutional network had three convolutional layers with 128 channels, a stride of 2 and filter size of 5×5 interleaved with ReLU nonlinearities, followed by a global average pooling and a linear layer mapping the 128 outputs to the logits. It reaches 100% accuracy on train and test set.

D Image rights & attribution

Figure 1 consists of four images from different sources. The first image from the left was taken from [14] with permission of the author. The second image from the left was generated by ourselves. The third image from the left is from ref. [17]. It was released under the CC BY 4.0 license as stated here: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683 and adapted by us from Figure 2B of the corresponding publication. The image on the right is Figure 1 from ref. [118]. It was released under CC BY 4.0 license as stated here: https://www.aclweb.org/anthology/D17-1215/(at the bottom) and retrieved by us from.

The image from Section “Dataset shortcut opportunities” was adapted from Figure 1 of ref. [9] with permission from the authors (image cropped from original figure by us). The image from Section “Decision rule (shortcuts from discriminative learning)” was adapted from Figure 1 of ref. [38] with permission from the authors (image cropped from original figure by us). The image from Section “Generalisation reveals shortcuts” was adapted from Figure 1 of ref. [45] with permission from the authors (image cropped from original figure by us).

Figure 4 consists of a number of images from different sources. The first author of the corresponding publication is mentioned in the figure for identification. The images from ref. [8] were released under the CC BY 3.0 license as stated here: https://arxiv.org/abs/1312.6199 and adapted by us from Figure 5a of the corresponding publication (images cropped from original figure by us). The images from ref. [50] were adapted from Figure 1 of the corresponding paper with permission from the authors (images cropped from original figure by us). The images from ref. [48] were adapted from Figure 1 of the corresponding paper with permission from the authors (images cropped from original figure by us). The images from ref. [38] were adapted from Figure 1 of the corresponding paper with permission from the authors (images cropped from original figure by us). The images from ref. [41] were adapted from Figure 1 of the corresponding paper with per- mission from the authors (images cropped from original figure by us). The images from ref. [36] were adapted from Figure 5 of the corresponding paper with permission from the authors (images cropped from original figure by us). The images from ref. [9] were adapted from Figure 1 of the corresponding paper with permission from the authors (images cropped from original figure by us). The images from ref. [45] were adapted from Figure 1 and Figure 2 of the corresponding paper with permission from the authors (images cropped from original figures by us).

References

108. Geirhos, R. *et al.* Generalisation in humans and deep neural networks. In *Proc. Adv.* NeurIPS (2018).

109. Michaelis, C. *et al.* Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv:1907.07484* (2019).

110. Minderer, M., Bachem, O., Houlsby, N. & Tschannen, M. Automatic shortcut removal for self-supervised representation learning. *arXiv:2002.08822* (2020).

111. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D. & Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question An- swering. In *Proc. IEEE CVPR*, 6904–6913 (2017).

112. Kaushik, D. & Lipton, Z. C. How much reading does reading comprehension re- quire? A critical investigation of popular benchmarks. *arXiv:1808.04926* (2018).

113. Geva, M., Goldberg, Y. & Berant, J. Are we modeling the task or the annota- tor? An investigation of annotator bias in natural language understanding datasets. *arXiv:1908.07898* (2019).

114. Kavumba, P. *et al.* When choosing plausible alternatives, Clever Hans can be clever. arXiv:1911.00225 (2019).

115. McCoy, R. T., Pavlick, E. & Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in Natural Language Inference. *arXiv:1902.01007* (2019).

116. Agrawal, A., Batra, D. & Parikh, D. Analyzing the behavior of visual question answering models. *arXiv:1606.07356* (2016).

117. Belinkov, Y. & Bisk, Y. Synthetic and natural noise both break neural machine translation. *arXiv:1711.02173* (2017).

118. Jia, R. & Liang, P. Adversarial examples for evaluating reading comprehension systems. *arXiv:1707.07328* (2017).

119. Glockner, M., Shwartz, V. & Goldberg, Y. Breaking NLI systems with sentences that require simple lexical inferences. *arXiv:1805.02266* (2018).

120. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* **1** (2019).

121. Brown, T. B. *et al.* Language models are few-shot learners (2020). 2005.14165.

122. Murphy VII, T. The first level of Super Mario Bros. is easy with lexicographic orderings and time travel. *Proc. SIGBOVIK* 112 (2013).

123. Amodei, D. *et al.* Concrete problems in AI safety. *arXiv:1606.06565* (2016).

124. Tobin, J. *et al.* Domain randomization for transferring deep neural networks from simulation to the real world. In *Proc. IEEE/RSJ IROS*, 23–30 (IEEE, 2017).

125. Akkaya, I. *et al.* Solving Rubik's Cube with a robot hand. *arXiv:1910.07113* (2019).

126. Zhao, J., Wang, T., Yatskar, M., Ordonez, V. & Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv:1707.09457* (2017).

127. Rich, A. S. & Gureckis, T. M. Lessons for artificial intelligence from the study of natural stupidity. *Nat. Mach. Int.* **1**, 174 (2019).

128. Hashimoto, T. B., Srivastava, M., Namkoong, H. & Liang, P. Fairness without demographics in repeated loss minimization. *arXiv:1806.08010* (2018).

129. Cohen, T. & Welling, M. Group equivariant convolutional networks. In *Proc. ICML*, 2990–2999 (2016).

130. Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V. & Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proc. IEEE CVPR*, 113–123 (2019).

131. Berthelot, D. *et al.* Mixmatch: A holistic approach to semi-supervised learning. arXiv:1905.02249 (2019).

132. Hjelm, R. D. *et al.* Learning deep representations by mutual information estimation and maximization. *arXiv:1808.06670* (2018).

133. Oord, A. v. d., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv:1807.03748* (2018).

134. Schölkopf, B. Causality for machine learning. *arXiv:1911.10500* (2019).

135. Schott, L., Rauber, J., Bethge, M. & Brendel, W. Towards the first adversarially robust neural network model on MNIST. In *Proc. ICLR* (2019).

136. Engstrom, L. *et al.* Learning perceptually-aligned representations via adversarial robustness. *arXiv:1906.00945* (2019).

137. Tramer, F., Carlini, N., Brendel, W. & Madry, A. On adaptive attacks to adversarial example defenses. arXiv preprint arXiv:2002.08347 (2020).

138. Peters, J., Bühlmann, P. & Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *J. R. STAT. SOC. B.: Ser. B Meth.* **78**, 947–1012 (2016).

139. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through awareness. In Proc. 3rd Inn. Theor. Comput. Sc. Conf., 214–226 (2012).

140. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. Learning fair representations. In *Proc. ICML*, 325–333 (2013).

141. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. In *Proc. Adv. NeurIPS*, 3315–3323 (2016).

142. Kusner, M. J., Loftus, J., Russell, C. & Silva, R. Counterfactual fairness. In *Proc. Adv. NeurIPS*, 4066–4076 (2017).

143. Schmidhuber, J. Evolutionary principles in self-referential learning. On learning how to learn: The meta-

meta-. hook.) Diploma thesis, Institut f. Informatik, Tech. Univ. Munich **1**, 2 (1987).

144. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. & Lillicrap, T. Meta-learning with memory-augmented neural networks. In *Proc. ICML*, 1842–1850 (2016).

145. Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. ICML* (2017).

146. Bengio, Y. *et al.* A meta-transfer objective for learning to disentangle causal mechanisms. arXiv:1901.10912 (2019).

147. Ke, N. R. *et al.* Learning neural causal models from unknown interventions. arXiv:1910.01075 (2019).

148. Fetaya, E., Jacobsen, J.-H., Grathwohl, W. & Zemel, R. Understanding the limitations of conditional generative models. In *Proc. ICLR* (2020).

149. Higgins, I. *et al.* Beta-VAE: Learning basic visual concepts with a constrained variational framework. *Proc. ICLR* (2017).

150. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Networks* **13**, 411–430 (2000).