Shortcut Learning

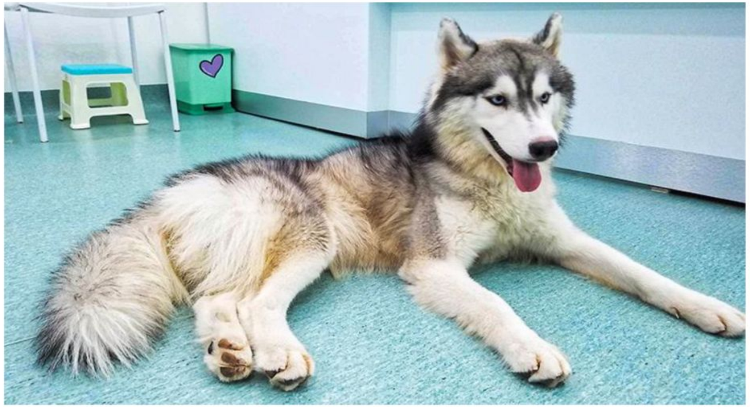
Summary by Alex Fedorov with help of Dr. Plis

May 1, 2020

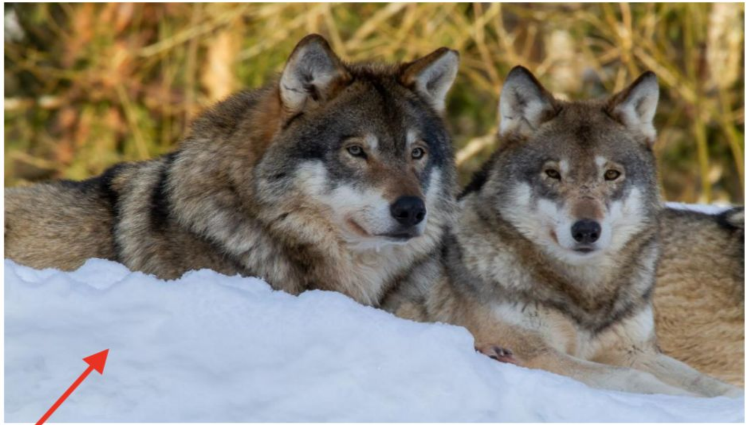
Geirhos, Robert, et al. "Shortcut Learning in Deep Neural Networks." *arXiv preprint arXiv:2004.07780* (2020).

4 + 2 = 6
SP

We trained a deep neural network ~~with superhuman performance!~~
from a bad dataset.



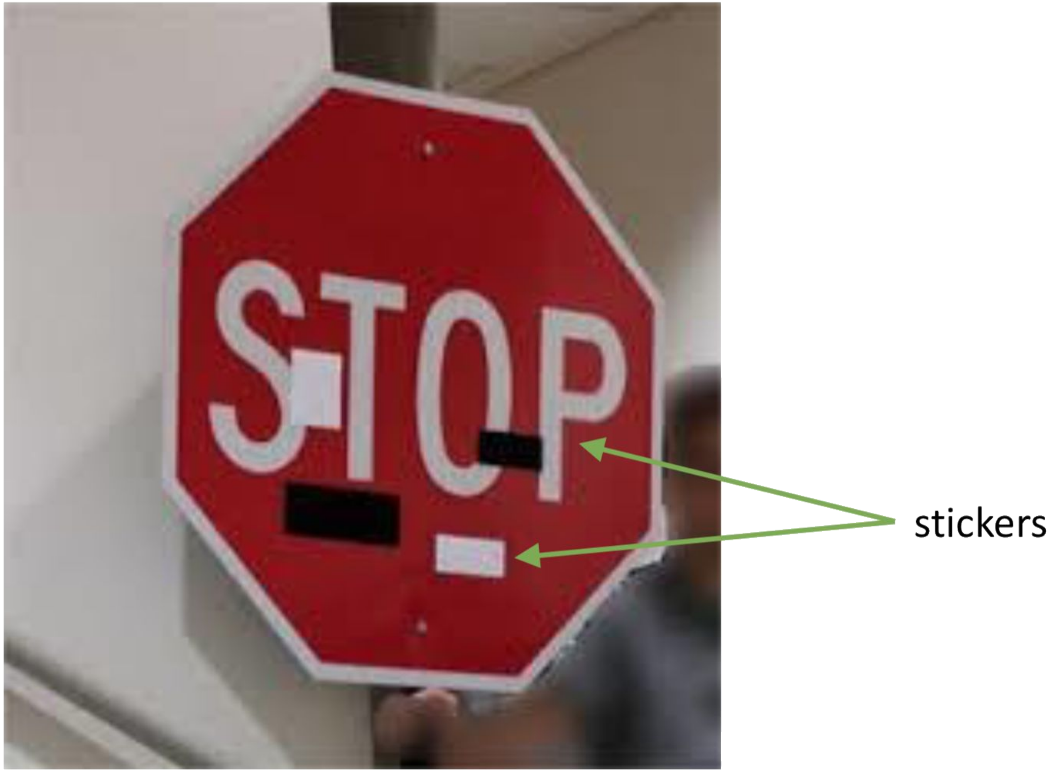
Husky



Wolf

snow was the most important feature

[Ribeiro, Singh & Guestrin ‘16]



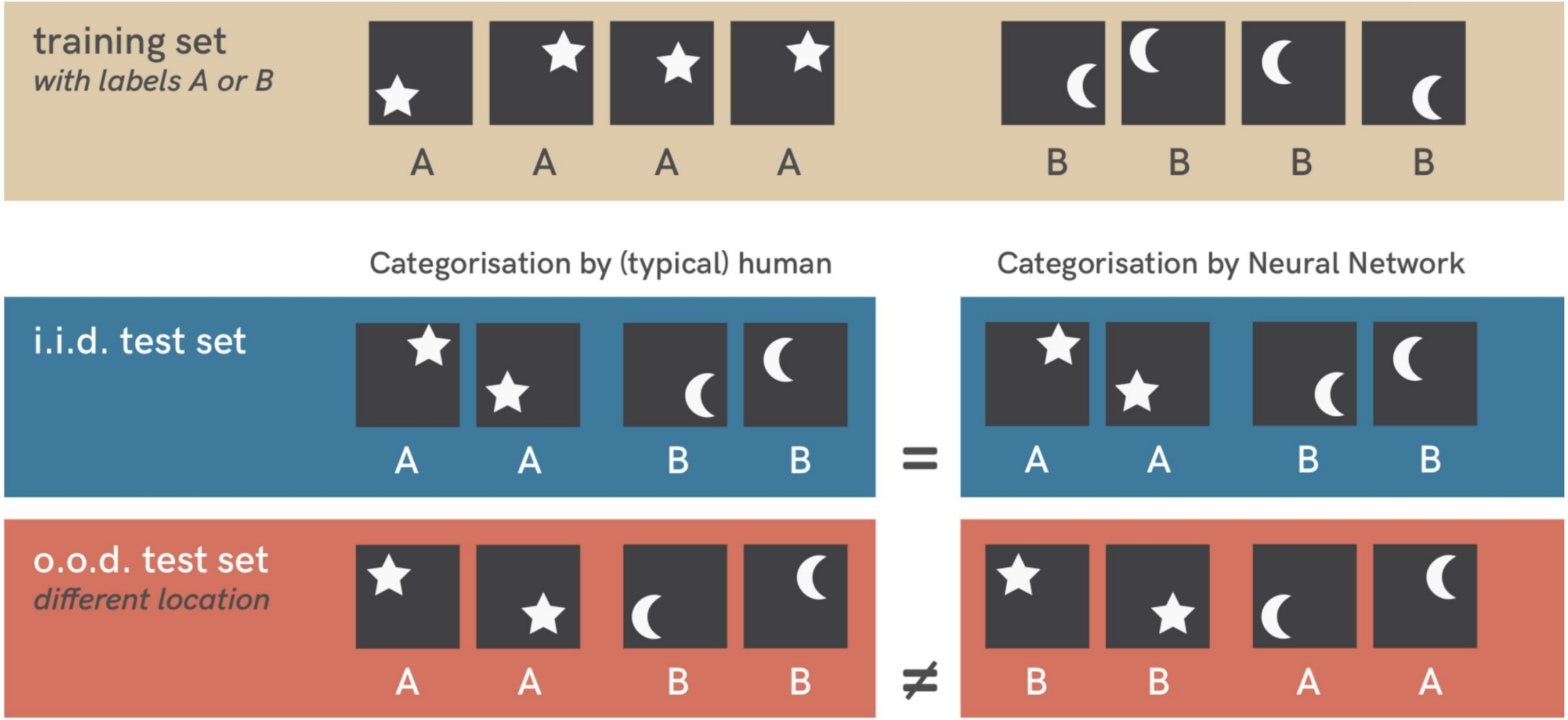
deep neural network: I see “**speed limit 45**”

why???

[Eykholt et al. ’18]

Fantastic Shortcuts

Toy example



training set
with labels A or B

Categorisation by (typical) human

Categorisation by Neural Network

i.i.d. test set

o.o.d. test set
different location

Definitions

- **Shortcuts**
  - *Decision rules* that perform well on ***i.i.d.*** test data but fail on ***o.o.d.***
    - Revealing a mismatch between intended and learned solution
- i.i.d.

https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables

  - collection of random variables is ***independent*** and ***identically*** distributed
  - if each random variable has the same probability distribution as the others and all are mutually independent
- o.o.d.

Krueger, David, et al. "**Out-of-Distribution** Generalization via Risk Extrapolation (REx)." *arXiv preprint arXiv:2003.00688* (2020).

  - Weak form
    - Ability to successfully interpolate between multiple observed distributions
  - Strong form
    - Ability to extrapolate beyond the distributions observed during training

Taxonomy

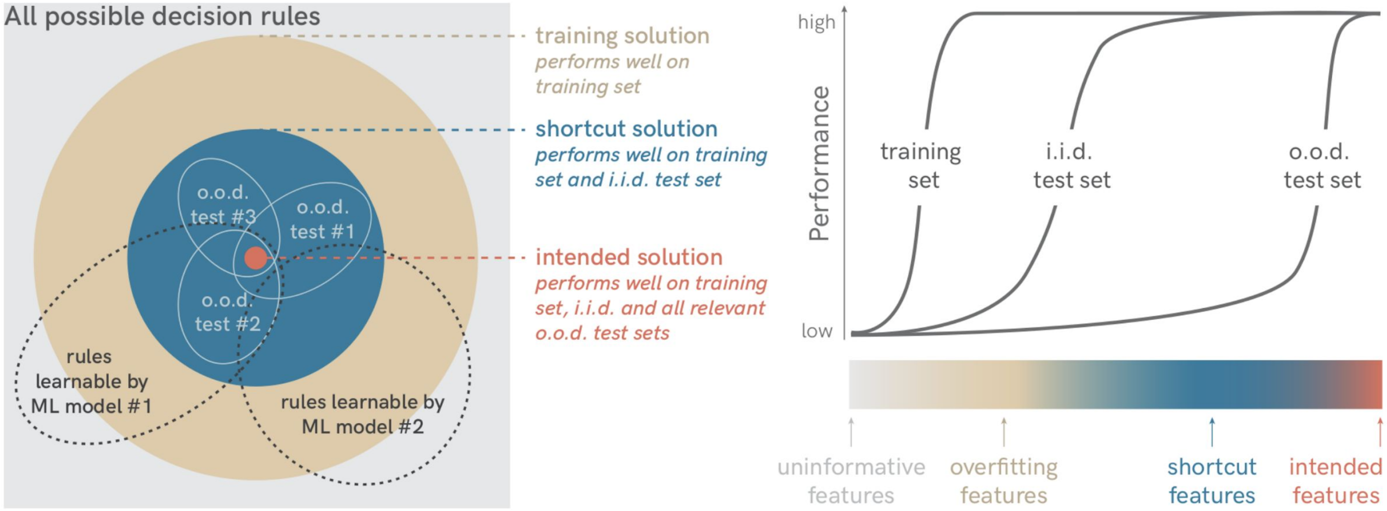

All possible decision rules
training solution
performs well on training set
shortcut solution
performs well on training set and i.i.d. test set
intended solution
performs well on training set, i.i.d. and all relevant o.o.d. test sets
o.o.d. test #3
o.o.d. test #1
o.o.d. test #2
rules learnable by ML model #1
rules learnable by ML model #2
high
low
Performance
training set
i.i.d. test set
o.o.d. test set
uninformative features
overfitting features
shortcut features
intended features

Figure 2. Taxonomy of decision rules. Among the set of all possible rules, only some solve the training data. Among the solutions that solve the training data, only some generalise to an i.i.d. test set. Among those solutions, shortcuts fail to generalise to different data (o.o.d. test sets), but the intended solution does generalise.

Other different terms

- learning under covariate shift [16]
  - classification problems for which the training instances are governed by an input distribution that is allowed to differ arbitrarily from the test distribution
- anti-causal learning [17]
  - the effect as input and we try to predict the value of the cause variable that led to it
- dataset bias [18]
  - Selection Bias / Capture Bias / Negative Set Bias
    - obtaining data from multiple sources / data-augmentation / add negatives from other datasets
- the tank legend [19] (“dataset bias”/“data leakage”)
- the Clever Hans effect [20]
- unintended cue learning
  - Experiments conditions can benefit lab rats
- Surface learning (in school)
  - helps rather than hurts test performance on typical multiple-choice exams

Shortcuts in Computer vision

- domain transfer
  - domain-specific shortcut features
  - shortcuts limit the usefulness of unsupervised representations [54]
- adversarial examples derail the model prediction

Shortcut in NLP

- biases
  - annotation artefacts
  - like word length
- Robustness
  - inability to generalise to more challenging test conditions

Shortcut in RL

- reward hacking
- Loopholes
- generalisation or reality gap between sim and real cases
    - Introducing additional variation in colour, size, texture, lighting, etc

SHORTCUT in Fairness & algorithmic decision-making

- Gender is more important
- Bias amplification
- Underrepresentation
  - disparity amplification

FaNtastic Shortcuts and Where to find them

What makes a cow a cow?

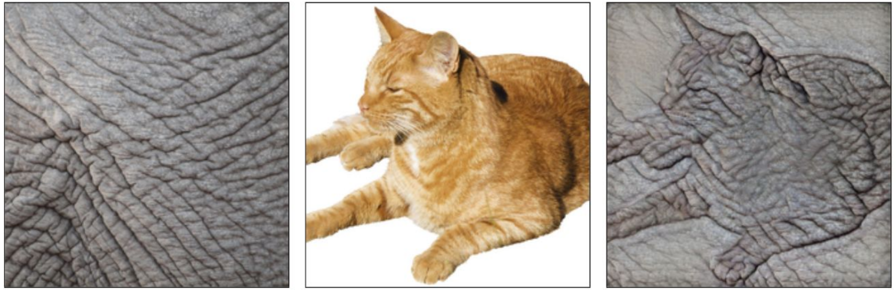
- Helps
  - Familiar background
- Confuses
  - Unexpected location
- Relationship
  - Object - Background
- **Dataset Bias**
  - **Contextual bias**
  - **High-frequency invisible patterns**
  - **Constrains the model**

Dataset

What makes a cat a cat?

- textures
  - model can ignore the shape
- feature combinations
- the decision rule to be invariant to an object shift
- severely biased to the extraction of overly simplistic features under specific dataset
- undesirable invariance
  - sometimes called excessive invariance is harmful

Decision Rule



(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

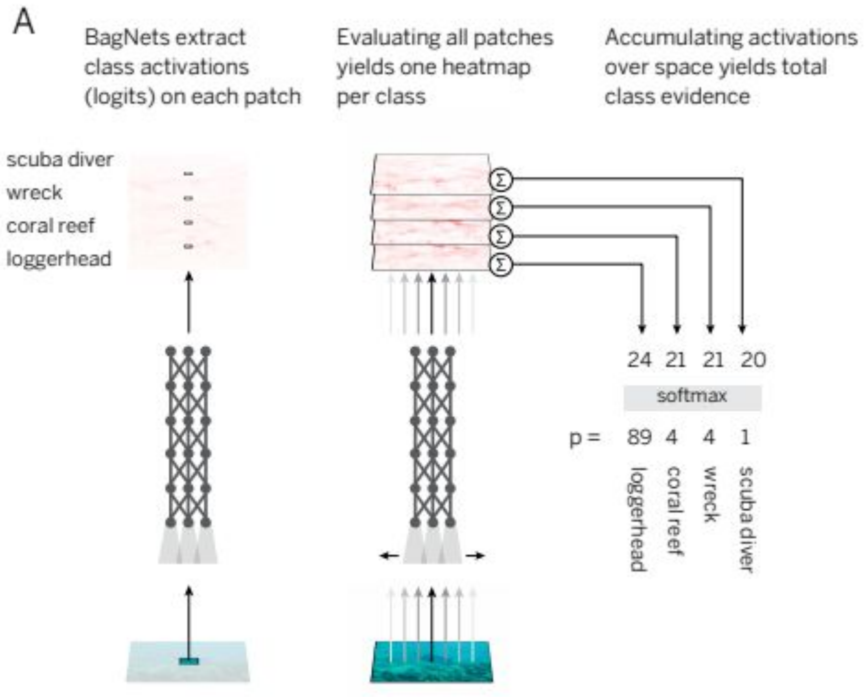
(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." ICLR2019.

Closer look to Textures: Method

- Baseline model
    - ResNet 50
- Changed receptive field to 9x9, 17x17, 33x33
- BagNet strategy


A
BagNets extract class activations (logits) on each patch
Evaluating all patches yields one heatmap per class
Accumulating activations over space yields total class evidence
scuba diver
wreck
coral reef
loggerhead
Σ
Σ
Σ
Σ
24 21 21 20
softmax
p = 89 4 4 1
loggerhead
coral reef
wreck
scuba diver

Closer look to Textures: Class Evidence

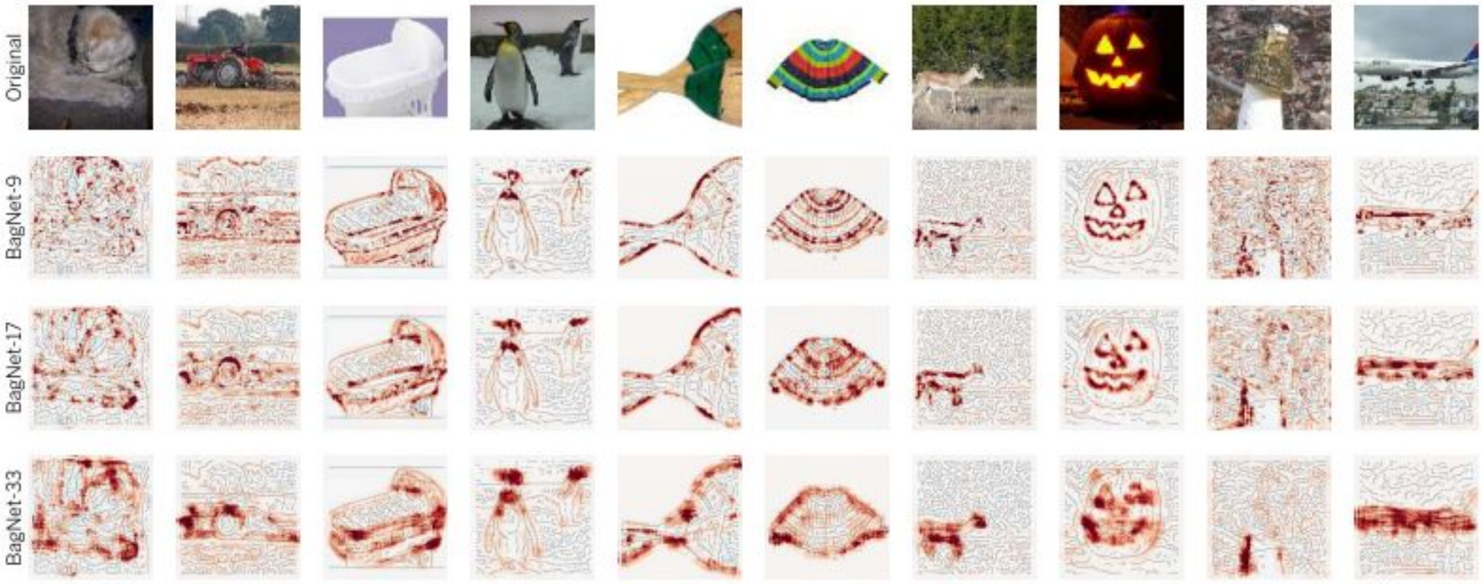


Figure 2: Heatmaps showing the class evidence extracted from of each part of the image. The spatial sum over the evidence is the total class evidence.

Closer look to Textures: Informative patches

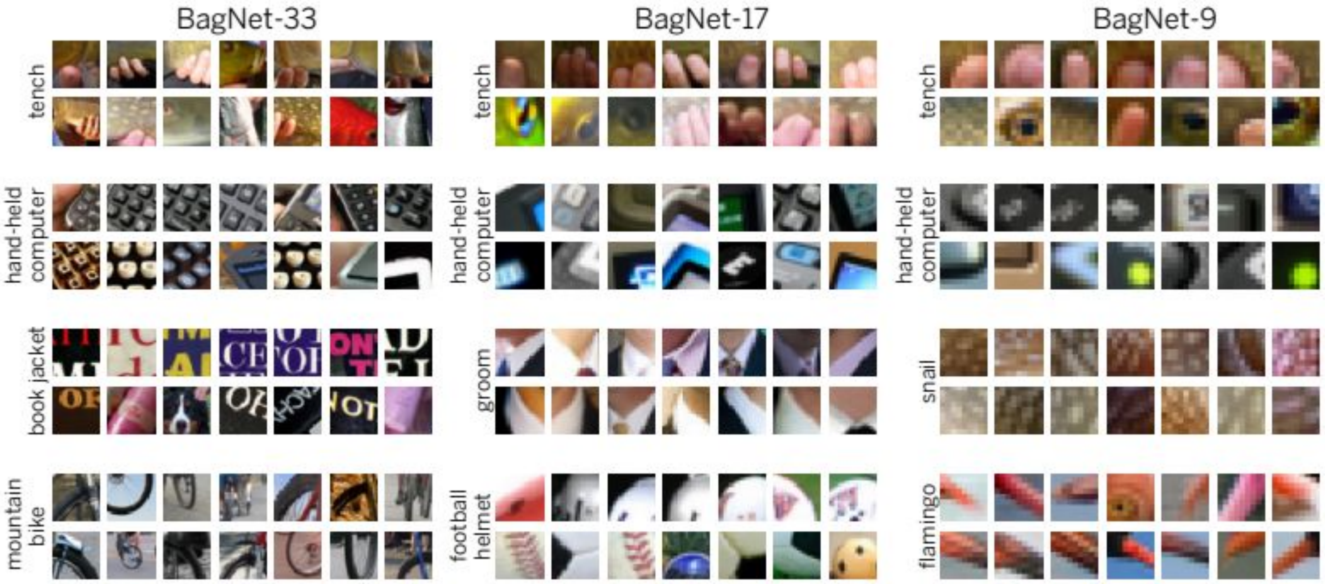


Figure 3: Most informative image patches for BagNets. For each class (row) and each model (column) we plot two subrows: in the top subrow we show patches that caused the highest logit outputs for the given class across all validation images with that label. Patches in the bottom subrow are selected in the same way but from all validation images with a *different* label (highlighting errors).

Closer look to Textures: Is bagnet behave as VGG-16

- Image Scrabbling
  - Gram Matrix
  - To keep textures, but reduce spatial information
- Same results for ResNet-50, ResNet-152, DenseNet-169
- Models memorize textures

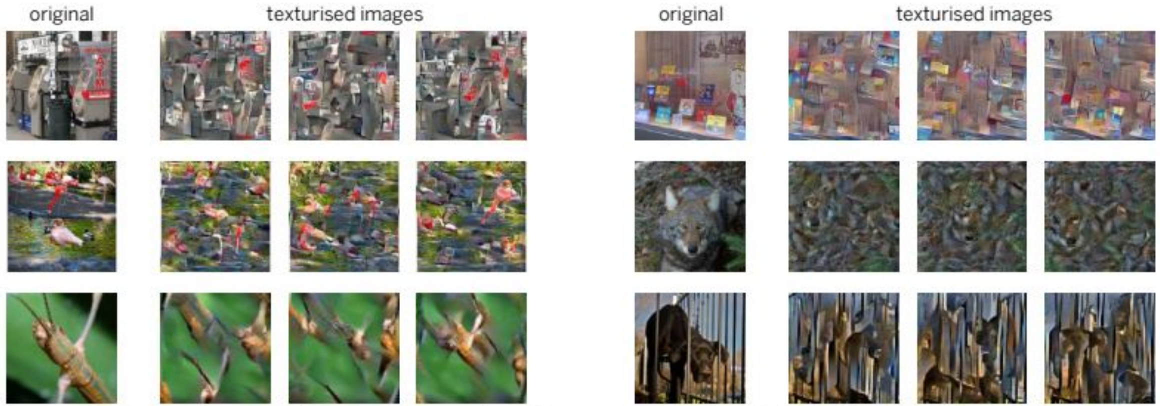
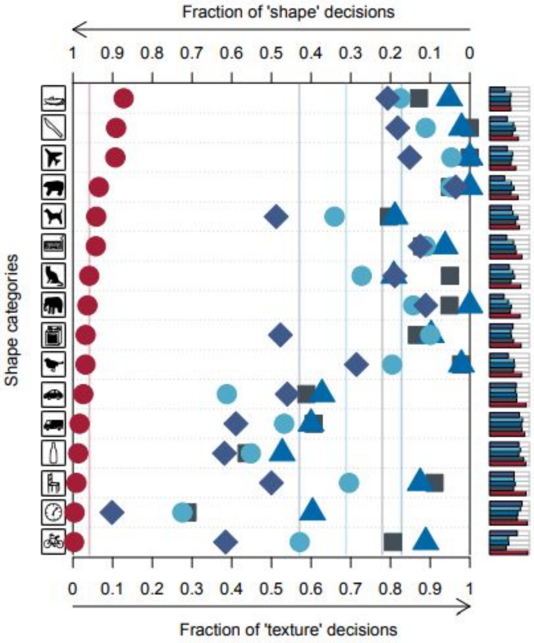


Figure 5: Examples of original and texturised images. A vanilla VGG-16 still reaches high accuracy on the texturised images while humans suffer greatly from the loss of global shapes in many images.

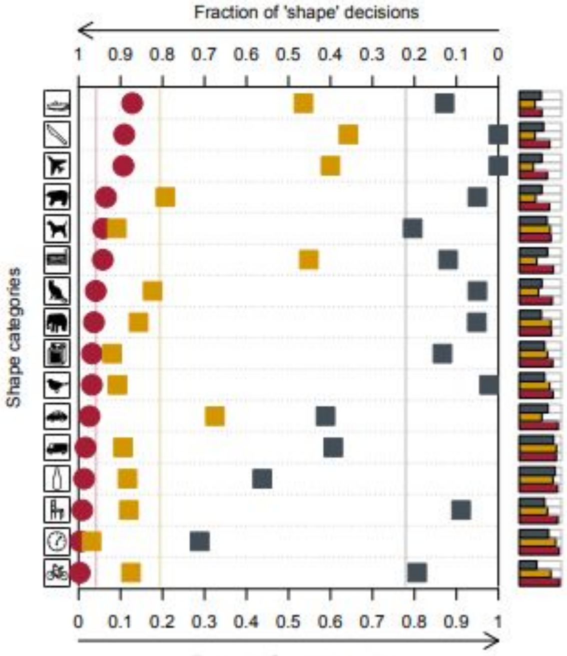
Closer look to Textures: Is bagnet behave as VGG-16

Figure 4: Classification results for human observers (red circles) and ImageNet-trained networks AlexNet (purple diamonds), VGG-16 (blue triangles), GoogLeNet (turquoise circles) and ResNet-50 (grey squares). Shape vs. texture biases for stimuli with cue conflict (sorted by human shape bias). Within the responses that corresponded to either the correct texture or correct shape category, the fractions of texture and shape decisions are depicted in the main plot (averages visualised by vertical lines). On the right side, small barplots display the proportion of correct decisions (either texture or shape correctly recognised) as a fraction of all trials. Similar results for ResNet-152, DenseNet-121 and Squeezenet1_1 are reported in the Appendix, Figure 13.



Solution: Style

Figure 5: Shape vs. texture biases for stimuli with a texture-shape cue conflict after training ResNet-50 on Stylized-ImageNet (orange squares) and on ImageNet (grey squares). Plotting conventions and human data (red circles) for comparison are identical to Figure 4. Similar results for other networks are reported in the Appendix, Figure 11.



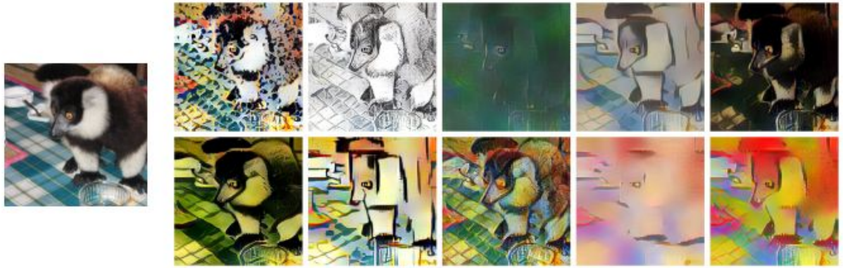
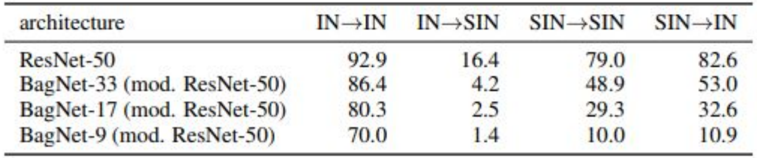


Figure 3: Visualisation of Stylized-ImageNet (SIN), created by applying AdaIN style transfer to ImageNet images. Left: randomly selected ImageNet image of class `ring-tailed lemur`. Right: ten examples of images with content/shape of left image and style/texture from different paintings. After applying AdaIN style transfer, local texture cues are no longer highly predictive of the target class, while the global shape tends to be retained. Note that within SIN, every source image is stylized only once.



architecture	IN→IN	IN→SIN	SIN→SIN	SIN→IN
ResNet-50	92.9	16.4	79.0	82.6
BagNet-33 (mod. ResNet-50)	86.4	4.2	48.9	53.0
BagNet-17 (mod. ResNet-50)	80.3	2.5	29.3	32.6
BagNet-9 (mod. ResNet-50)	70.0	1.4	10.0	10.9

Table 1: Stylized-ImageNet cannot be solved with texture features alone. Accuracy comparison (in percent; top-5 on validation data set) of a standard ResNet-50 with Bag of Feature networks (BagNets) with restricted receptive field sizes of 33×33, 17×17 and 9×9 pixels. Arrows indicate: train data→test data, e.g. IN→SIN means training on ImageNet and testing on Stylized-ImageNet.

What makes a guitar a Guitar

Generalization

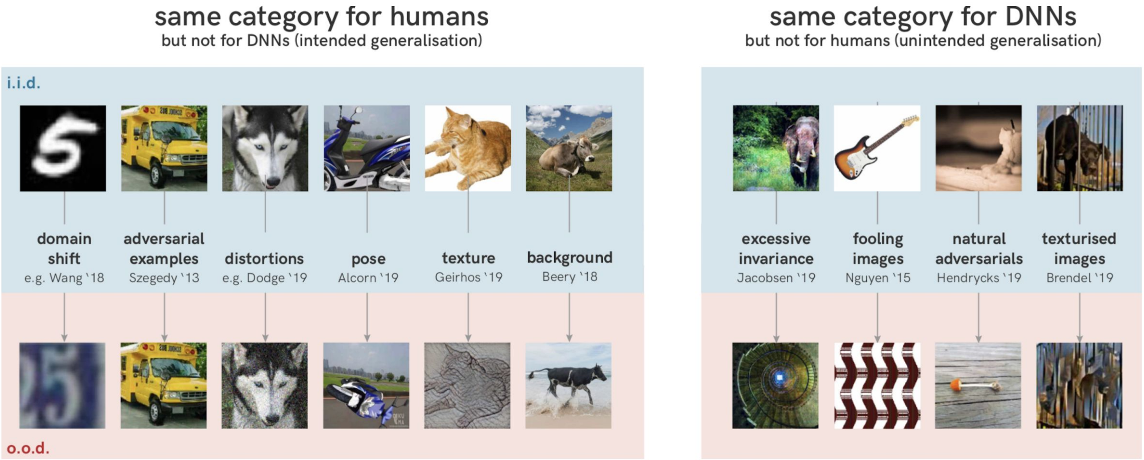


Figure 3. Both human and machine vision generalise, but they generalise very differently. Left: image pairs that belong to the same category for humans, but not for DNNs. Right: images pairs assigned to the same category by a variety of DNNs, but not by humans.

How to pet these fantastic Shortcuts

Interpreting Results

Carefully

- Distinguishing datasets and underlying abilities
- Morgan's Canon
- Testing (surprisingly) strong baselines

Distinguishing datasets and underlying abilities

- verify how closely this test performance measures the underlying ability one is actually interested in
  - ImageNet dataset [75] was intended to measure the ability “object recognition”, but DNNs seem to rely mostly on “counting texture patches” [36].
- same strategy assumption is paralleled by deep learning

Morgan's Canon

- Anthropomorphism
  - **the tendency of humans to attribute human-like psychological characteristics to nonhumans on the basis of insufficient empirical evidence**
- Morgan’s Canon:
  - “In no case is an animal activity to be interpreted in terms of higher psychological processes if it can be fairly interpreted in terms of processes which stand lower on the scale of psychological evolution and development"
  - Never attribute to high-level abilities that which can be adequately explained by shortcut learning

Testing (surprisingly) strong baselines

- to test whether a baseline model exceeds expectations even though it does not use intended features
  - using nearest neighbours
  - Hays, J. & Efros, A. A. Scene completion using millions of photographs. ACM Transactions on Graphics (TOG)
  - Hays, J. & Efros, A. A. IM2GPS: estimating geographic information from a single image.
- we must not confuse performance on a dataset with the acquisition of an underlying ability.

Detecting shortcuts

towards o.o.d. generalization test

- Making o.o.d. generalisation tests a standard practice
- Designing good o.o.d. Tests
- O.O.D. benchmarks

Making o.o.d. generalisation tests a standard practice

- i.i.d. is “the big lie in machine learning”
  - Ghahramani, Z. Panel of workshop on advances in Approximate Bayesian Inference (AABI) 2017 (2017).
- o.o.d. tests
  - Borowski, J. et al. The notorious difficulty of comparing human and machine perception. (2019).
  - Lake,B.M.,Ullman,T.D.,Tenenbaum,J.B. & Gershman,S.J. Building machines that learn and think like people. arXiv preprint arXiv:1604.00289 (2016).
  - Chollet, F. The measure of intelligence.
  - Crosby, M., Beyret, B. & Halina, M. The Animal-AI Olympics.
  - Juliani, A. et al. Obstacle tower: A generalization challenge in vision, control, and planning. arXiv preprint arXiv:1902.01378 (2019).

Designing good o.o.d. tests

- distribution shift
- good o.o.d. tests
    - clear distribution shift
    - well-defined intended solution
    - current models struggle
- Winograd Schema Challenge

O.O.D. benchmarks

- adversarial attacks
- ARCT with removed shortcuts
- Cue conflict stimuli - conflicting texture, shape
- ImageNet-A - images which are usually wrong
- ImageNet-C - 15 image corruptions
- Object-Net - distangle background, rotation and viewpoint
- PACS - domain-based
- Shift-Mnist, biased-CelebeA, unfair dSprites - adding correlations in the training data

Why Fantastic shortcuts learned

Shortcuts: why are they learned?

- The principle of least effort
  - some words are easier to say: plane - airplane
- Inductive bias

https://en.wikipedia.org/wiki/Inductive_bias

  - The inductive bias (also known as learning bias) of a learning algorithm is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered.

Inductive biases: architecture

- convolutions make it harder to use location
    - d'Ascoli, S., Sagun, L., Bruna, J. & Biroli, G. Finding the needle in the haystack with convolutions: On the benefits of architectural bias. arXiv preprint arXiv:1906.06766 (2019).
    - It helps with denoising and inpainting
        - Ulyanov, D., Vedaldi, A. & Lempitsky, V. Deep image prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 9446–9454 (2018).
- attention layers
    - Understand context by modeling relationships
    - Vaswani, A. et al. Attention is all you need. In Advances in Neural Information Processing Systems, 5998–6008 (2017).
- ReLU activations can lead to unexpected effects like unwarranted confidence
    - Hein, M., Andriushchenko, M. & Bitterwolf, J. Why ReLU networks yield highconfidence predictions far away from the training data and how to mitigate the problem. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 41–50 (2019).

Inductive biases: Training data

- don't disappear even adding more data
- modifying images is better
  - vulnerability
    - Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards deep learning models resistant to adversarial attacks.
  - texture bias
    - Geirhos, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.

Inductive biases: loss function

- Use all available information
  - Jacobsen,J.-H.,Behrmann,J.,Zemel,R.&Bethge,M.Excessive invariance causes adversarial vulnerability.
- Disentangle intended features
  - Heinze-Deml, C. & Meinshausen, N. Conditional variance penalties and domain shift robustness.
  - Arjovsky,M.,Bottou,L.,Gulrajani,I.& Lopez-Paz,D.Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019).

Inductive biases: Optimization

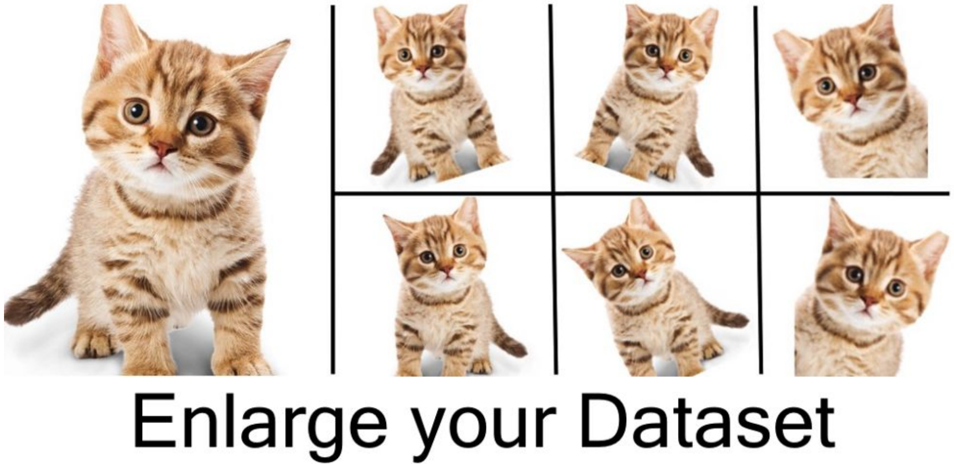
- Large learning rate to learn simple patterns
- Small learning rate for complex patterns and memorization

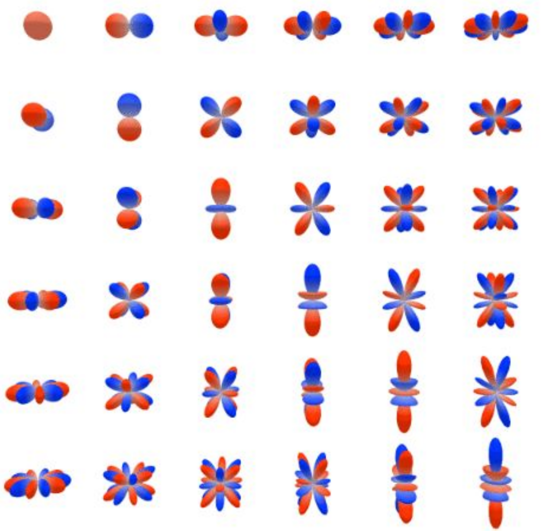
- Wu, L., Zhu, Z. & E, W. Towards understanding generalization of deep learning: Perspective of loss landscapes. arXiv preprint arXiv:1706.10239 (2017).
- De Palma, G., Kiani, B. T. & Lloyd, S. Deep neural networks are biased towards simple functions. arXiv preprint arXiv:1812.10156 (2018).
- Valle-Perez, G., Camargo, C. Q. & Louis, A. A. Deep learning generalizes because the parameter-function map is biased towards simple functions. In International Conference on Learning Representations (2019).
  24
- Sun, K. & Nielsen, F. Lightlike neuromanifolds, Occam’s Razor and deep learning. arXiv preprint arXiv:1905.11027 (2019).
- Arpit, D. et al. A closer look at memorization in deep networks. In International Conference on Machine Learning (2017).
- Li, Y., Wei, C. & Ma, T. Towards explaining the regularization effect of initial large learning rate in training neural networks. arXiv preprint arXiv:1907.04595 (2019).

Beyond fantastic shortcuts

DOMAIN-SPECIFIC PRIOR KNOWLEDGE

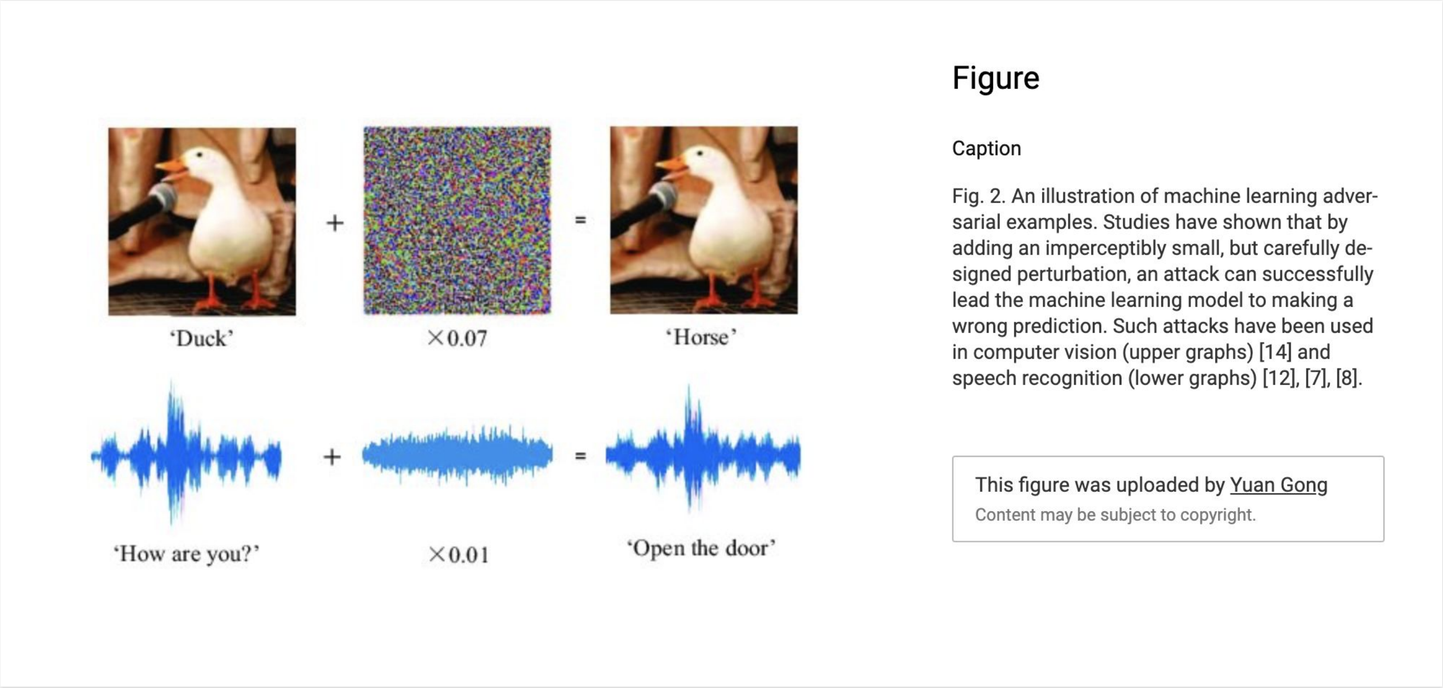
- data-augmentation
- hard-coded rotation invariance
  - Cohen, T. & Welling, M. Group equivariant convolutional networks. In International Conference on Machine Learning





ADVERSARIAL EXAMPLES AND ROBUSTNESS

- worst-case generalisation



Figure

Caption

Fig. 2. An illustration of machine learning adversarial examples. Studies have shown that by adding an imperceptibly small, but carefully designed perturbation, an attack can successfully lead the machine learning model to making a wrong prediction. Such attacks have been used in computer vision (upper graphs) [14] and speech recognition (lower graphs) [12], [7], [8].

This figure was uploaded by Yuan Gong

Content may be subject to copyright.

Domain adaptation, -generalisation and -randomisation

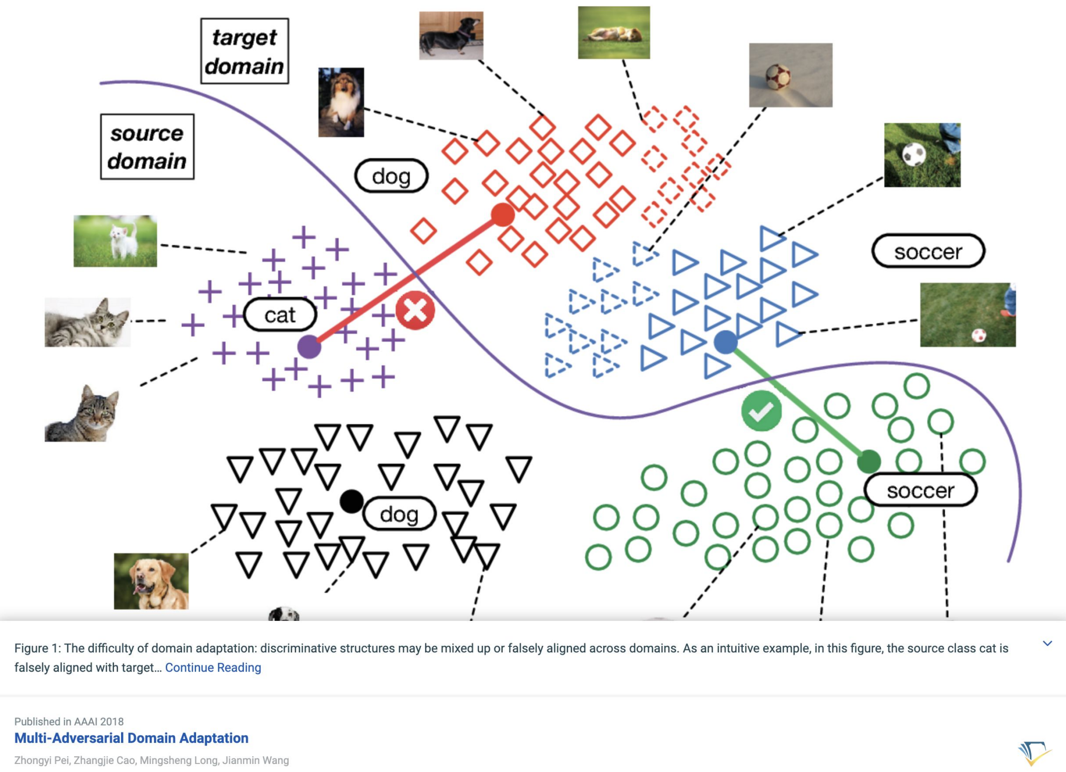


Figure 1: The difficulty of domain adaptation: discriminative structures may be mixed up or falsely aligned across domains. As an intuitive example, in this figure, the source class cat is falsely aligned with target… Continue Reading

Published in AAAI 2018

Multi-Adversarial Domain Adaptation

Zhongyi Pei, Zhangjie Cao, Mingsheng Long, Jianmin Wang

Meta-learning

- 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).
- Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. & Lillicrap, T. Meta-learning with memory-augmented neural networks. In International Conference on Machine Learning, 1842–1850 (2016).
- [Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In International Conference on Machine Learning (2017).

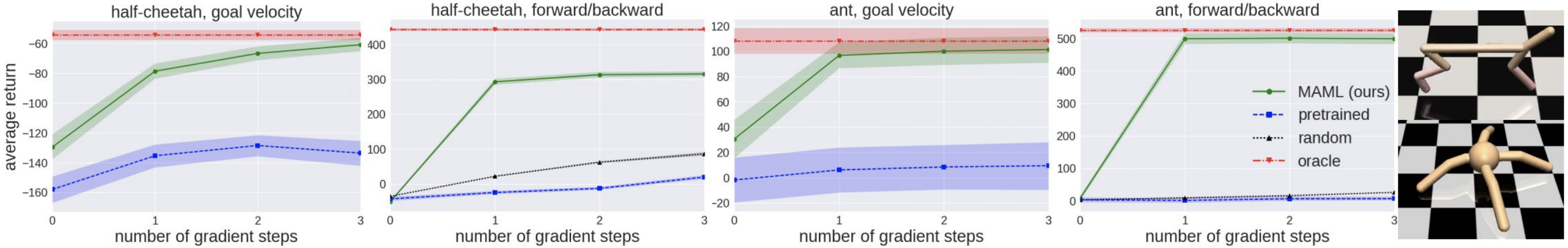


Figure 5. Reinforcement learning results for the half-cheetah and ant locomotion tasks, with the tasks shown on the far right. Each gradient step requires additional samples from the environment, unlike the supervised learning tasks. The results show that MAML can adapt to new goal velocities and directions substantially faster than conventional pretraining or random initialization, achieving good performs in just two or three gradient steps. We exclude the goal velocity, random baseline curves, since the returns are much worse (< -200 for cheetah and < -25 for ant).

Generative modelling and disentanglement

- Fetaya, E., Jacobsen, J.-H., Grathwohl, W. & Zemel, R. Understanding the limitations of conditional generative models.
- Beta-VAE: Learning basic visual concepts with a constrained variational framework.
- Hyva ̈rinen, A. & Oja, E. Independent component analysis: algorithms and applications. Neural Networks
- Scholkopf, B. Causality for machine learning. arXiv preprint arXiv:1911.10500 (2019).

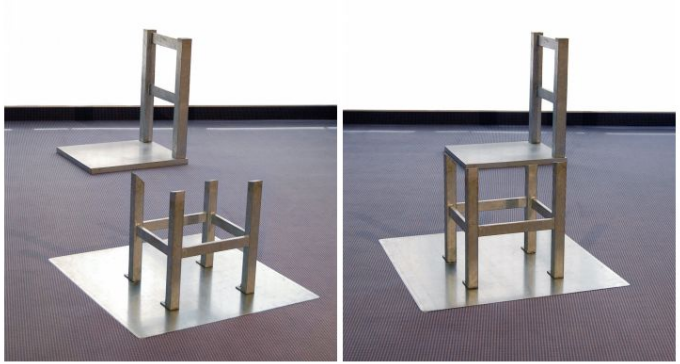


Figure 1: *Beuchet chair*, made up of two separate objects that appear as a chair when viewed from a special vantage point violating the independence between object and perceptual process. (Image courtesy of Markus Elsholz, reprinted from Peters et al. (2017).)

Fairness

- Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214-226 (2012).
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. Learning fair representations. In International Conference on Machine Learning, 325-333 (2013).
- Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems, 3315-3323 (2016).
- Kusner, M. J., Loftus, J., Russell, C. & Silva, R. Counterfactual fairness. In Advances in Neural Information Processing Systems, 4066-4076 (2017).

1. shortcut learning is ubiquitous
2. Interpreting results carefully
3. Testing o.o.d. Generalisation
4. Understanding what makes a solution easy to learn

Thank you.