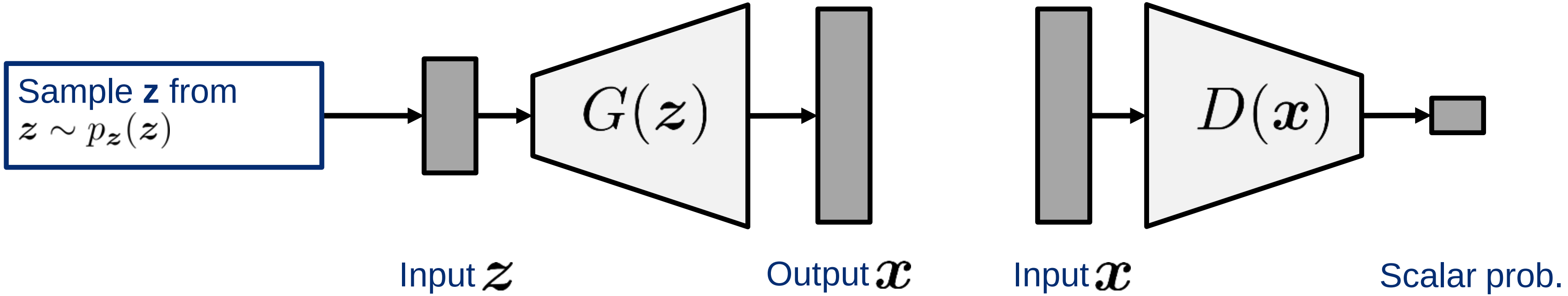
EN.601.482/682 Deep Learning

Interpretability, Generalization and Domain Gaps

Mathias Unberath, PhD
Assistant Professor
Dept of Computer Science
Johns Hopkins University

Generative Adversarial Networks



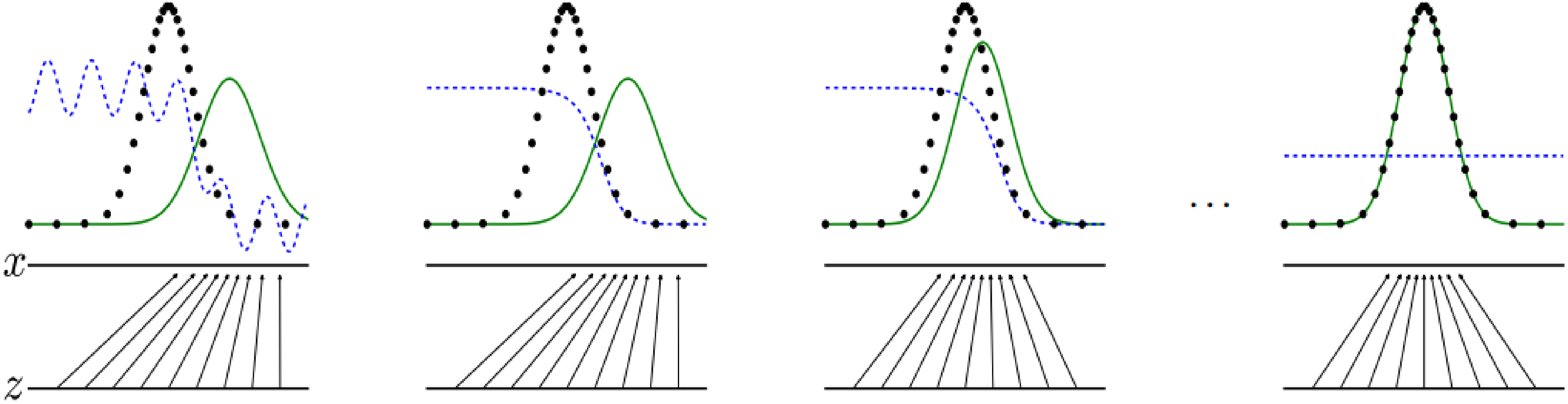
$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G((z))))]$$

GANs: A two-player minimax game with value function V(D,G)

Generator: Generates images in an attempt to fool the discriminator

Discriminator: Tries to differentiate between real and generated images

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. NeurIPS (pp. 2672-2680).



- Near convergence: $p_g(G)$ is similar to p_{data}, and $D(\boldsymbol{x})$ is partially accurate
- Inner loop: $D(\boldsymbol{x})$ is trained to better discriminate, converging to $D^\star(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}$
- After G update: Gradient of $D(\boldsymbol{x})$ has guided $G(\boldsymbol{z})$ to be more likely classified as “real”
- After multiple iterations, $p_g(G) = p_{\text{data}}$ and $D(\boldsymbol{x}) = \frac{1}{2}$

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. NeurIPS (pp. 2672-2680).

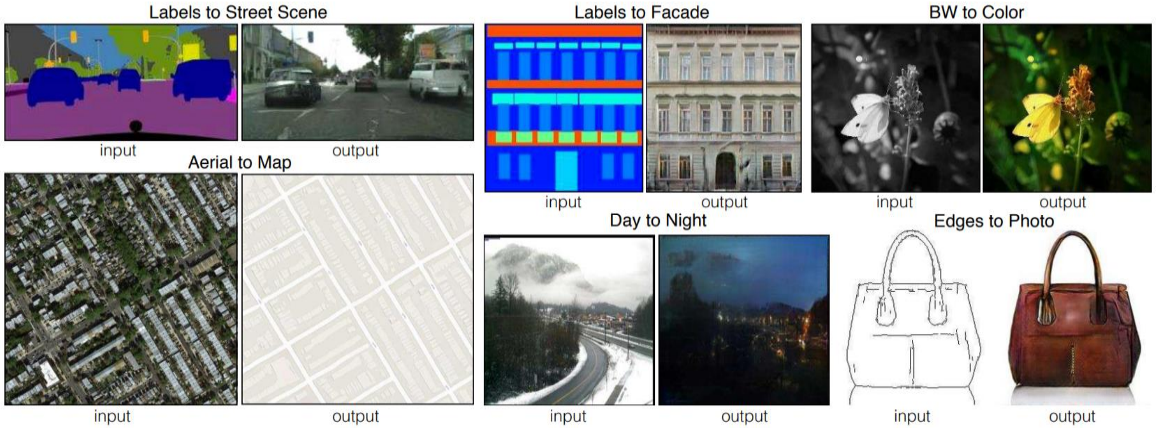
Image-to-image Translation

Standard GAN: Random noise vector to output image

$$\boldsymbol{E}_{\boldsymbol{x}}[\log D(\boldsymbol{x})] + \boldsymbol{E}_{\boldsymbol{z}}[\log(1 - D(G((z)))]$$

Conditional GAN: Random noise vector + observed image to output image

$$\boldsymbol{E}_{\boldsymbol{x},\boldsymbol{y}}[\log D(\boldsymbol{x}, \boldsymbol{y})] + \boldsymbol{E}_{\boldsymbol{x},\boldsymbol{z}}[\log(1 - D(\boldsymbol{x}, G(\boldsymbol{x}, \boldsymbol{z})))]$$



Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. CVPR (pp. 1125-1134).

Interpretability, Generalization and Domain Gaps

Visualization and Saliency

Visualization

Problem

- ConvNets are a bit of a black-box
- Input sample in, prediction out

→ Can we understand the prediction process?
→ Explainability, interpretability, reasoning, …
→ Important for discovering associations that were unknown a priori

Q: How to do this?

Visualization

Approach 1: Creating images that excite certain features

dumbbell cup dalmatian

Approach 2: Highlighting image regions that excite neurons

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, *1341*(3), 1.
Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Visualization

Approach 1: Creating images that excite certain features

dumbbell cup dalmatian

Approach 2: Highlighting image regions that excite neurons

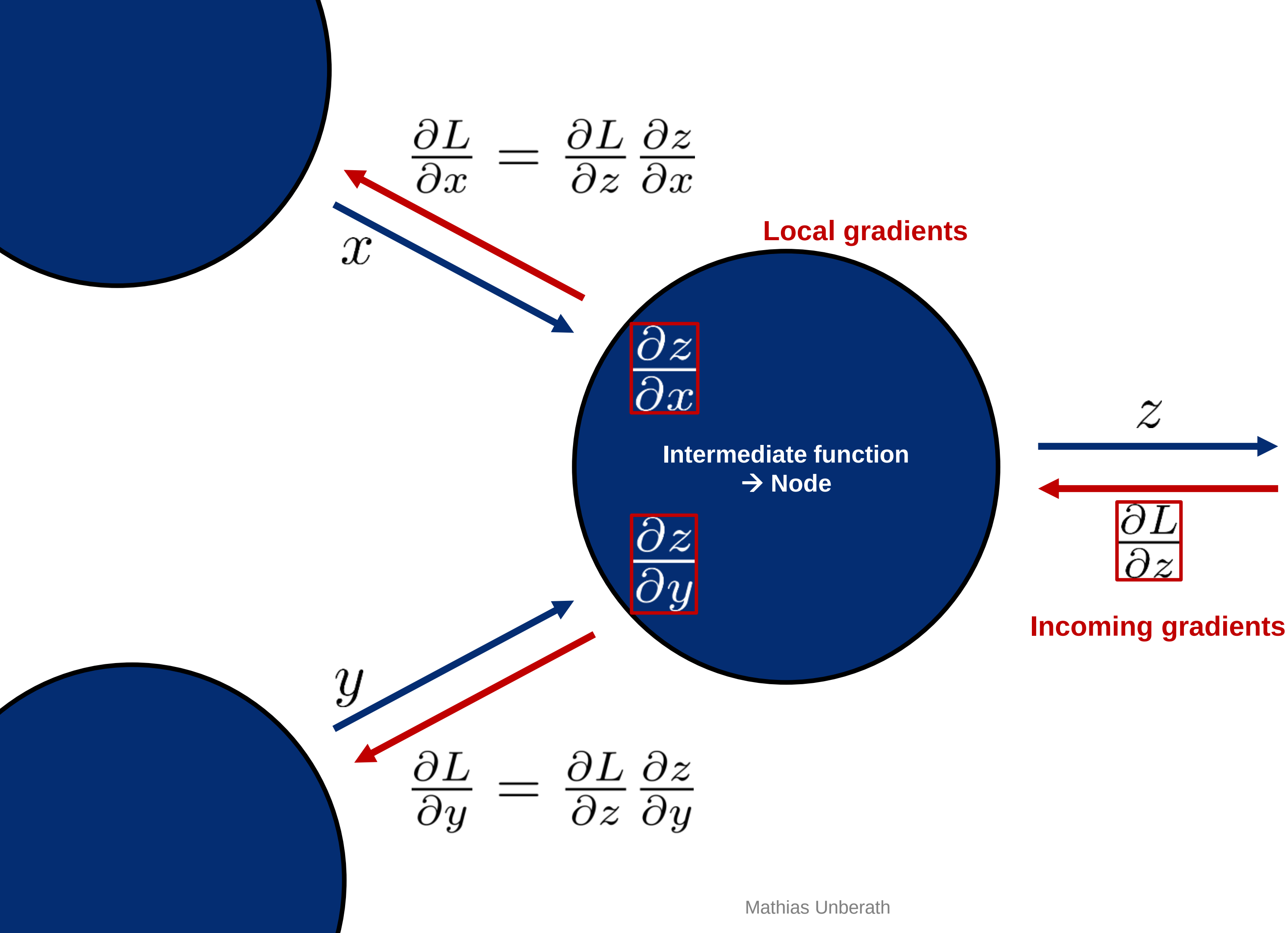
[Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, *1341*(3), 1.]
[Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.]

Creating Images that Excite Certain Features

Given a ConvNet that classifies images

What input image will result in the highest class activation?

Q: Do we know anything that can do this?



$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial x}$$
x
Local gradients
$\boxed{\frac{\partial z}{\partial x}}$
Intermediate function
→ Node
$\boxed{\frac{\partial z}{\partial y}}$
z
$\boxed{\frac{\partial L}{\partial z}}$
Incoming gradients
y
$$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial y}$$

Creating Images that Excite Certain Features

Given a ConvNet that classifies images

What input image will result in the highest class activation?

Backpropagation!

- Gradients w.r.t. weights
- But also: Upstream gradient! This is the gradient w.r.t. the input (the image!)

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Creating Images that Excite Certain Features

Backpropagation!

- Gradients w.r.t. weights
- But also: Upstream gradient! This is the gradient w.r.t. the input (the image!)

Then, via backprop solve $\arg\max_I S_c(I) + \lambda \|I\|_2^2$, where

- S_c is the class score of the class to be visualized,
- λ is a hyper parameter regularizing the L2 norm of the generated image

Q: What exactly is Sc? Posterior probabilities or unnormalized class scores?

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Creating Images that Excite Certain Features

Backpropagation!

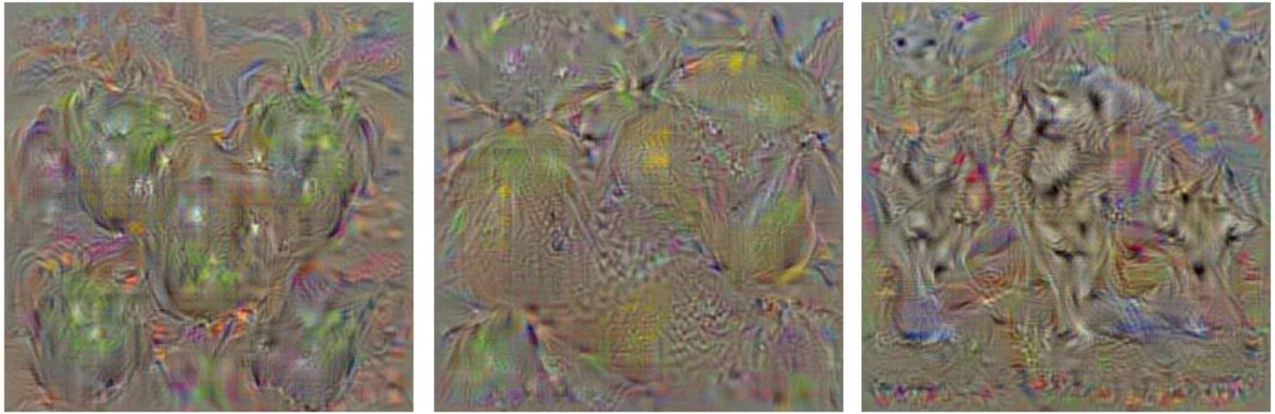
- Gradients w.r.t. weights
- But also: Upstream gradient! This is the gradient w.r.t. the input (the image!)

Then, via backprop solve $\arg\max_I S_c(I) + \lambda \|I\|_2^2$, where

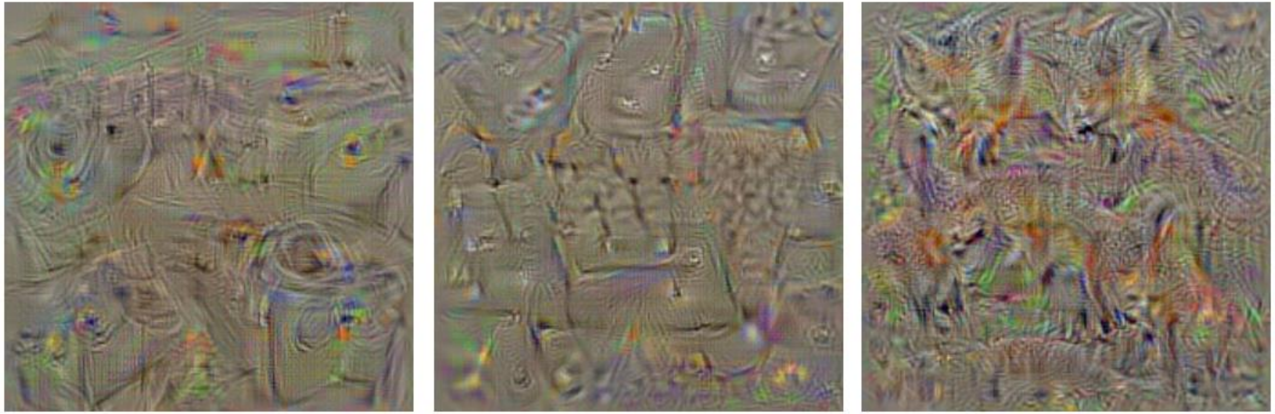
- S_c is the class score of the class to be visualized,
- λ is a hyper parameter regularizing the L2 norm of the generated image

Unnormalized class scores! Because $\arg\max_I P_c(I) + \lambda \|I\|_2^2 = \arg\max_I \frac{\exp S_c(I)}{\sum_c \exp S_c(I)} + \lambda \|I\|_2^2$ could be maximized by minimizing other class scores!

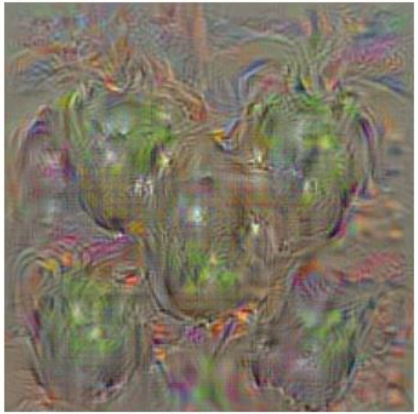
Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.



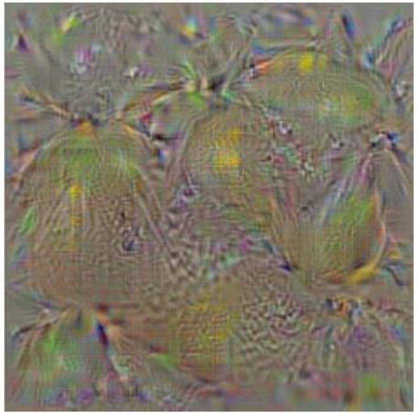
Which classes are maximally activated?



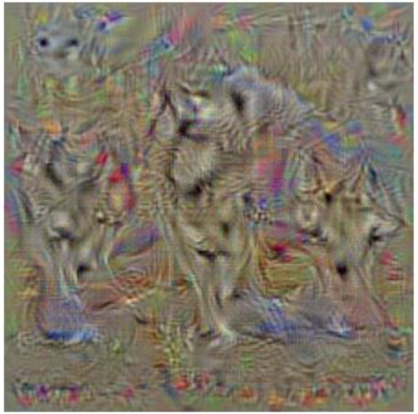
Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.



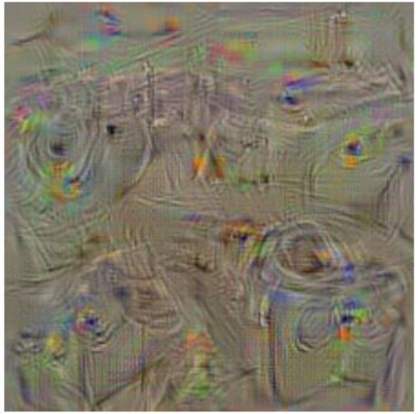
bell pepper



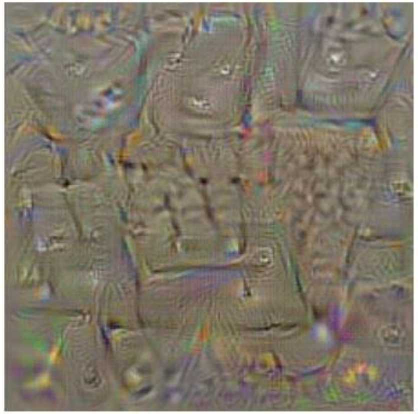
lemon



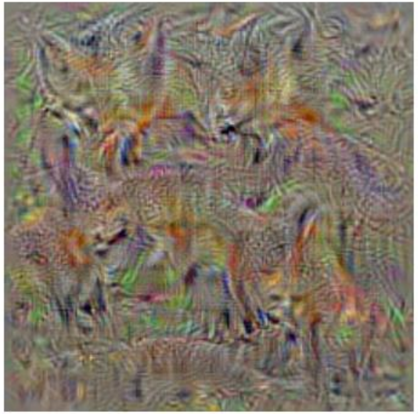
husky



washing machine



computer keyboard



kit fox

[Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.]

Creating Images that Excite Certain Features

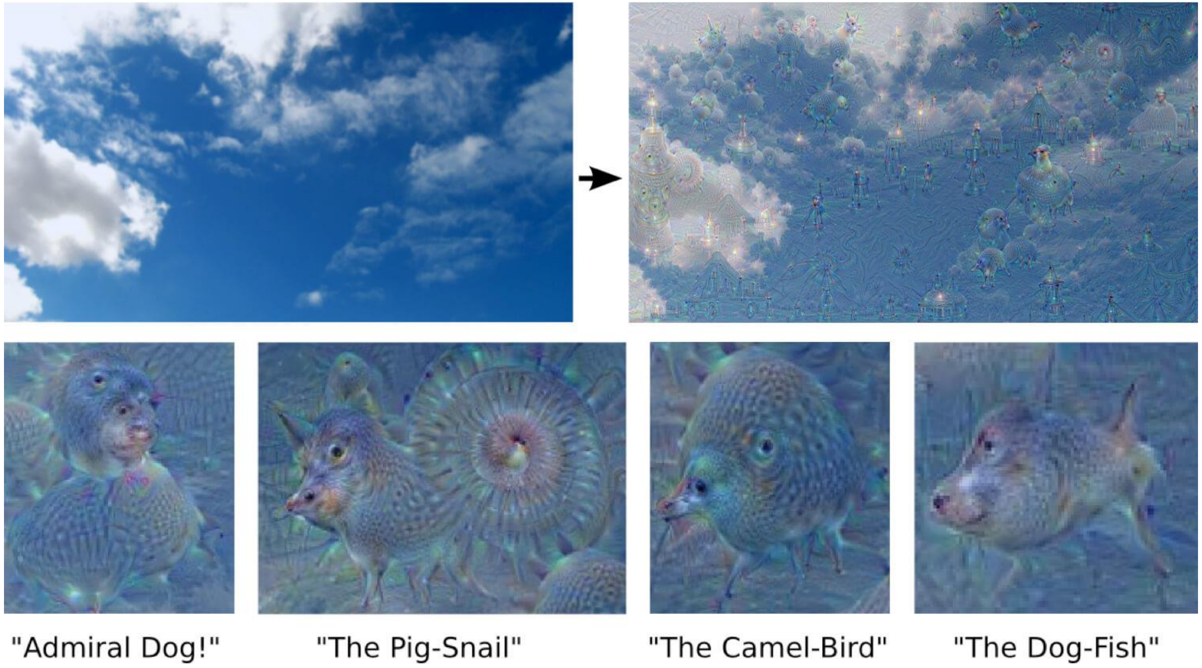
Backpropagation! $$\arg\max_{I} S_c(I) + \lambda \|I\|_2^2$$

The same concept of maximizing response w.r.t. the input image works for any feature map in a ConvNet (images generated with VGG16)

From a Blogpost written by Fabio Graetz.

Creating Images that Excite Certain Features

DeepDream: More or less the same thing but this time with hype!
→ Do not start from an all-zero image, but from any image



Images from a blogpost.

Visualization

Approach 1: Creating images that excite certain features

dumbbell cup dalmatian

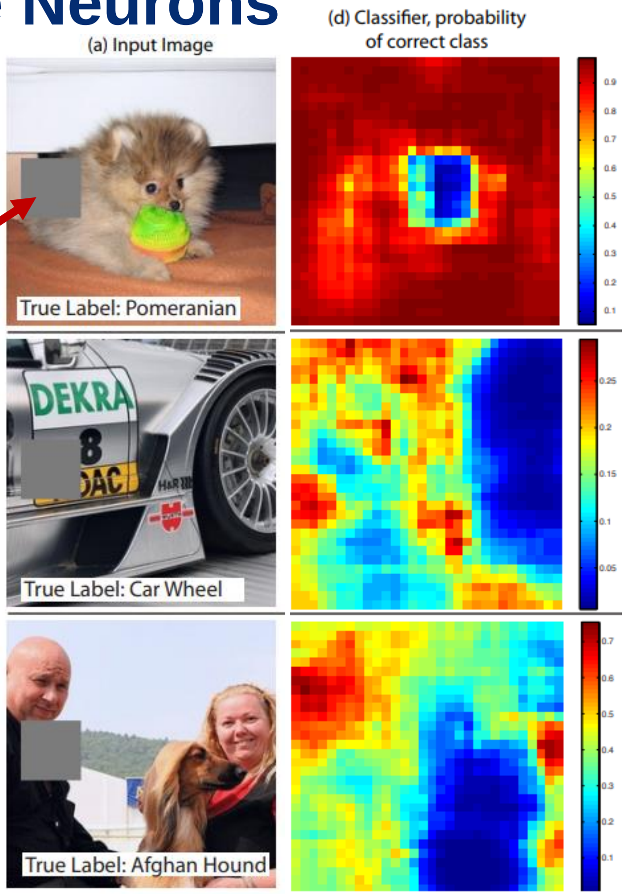
Approach 2: Highlighting image regions that excite neurons

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, *1341*(3), 1.
Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Highlighting Image Regions that Excite Neurons

Which regions contribute the most to the resulting prediction?

- Successively occlude regions in the image using a gray patch
- Monitor change in activation of correct class label



Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). springer, Cham.

Highlighting Image Regions that Excite Neurons

Which regions contribute the most to the resulting prediction?

Saliency

ConvNet forward pass

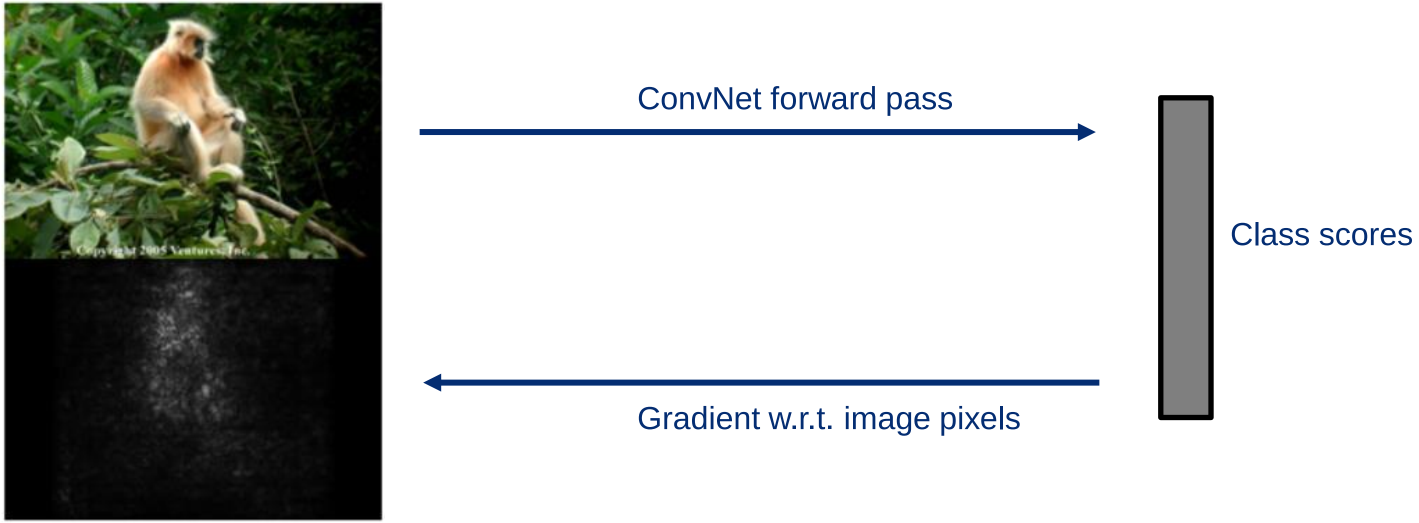
Class scores

[Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.]

Highlighting Image Regions that Excite Neurons

Which regions contribute the most to the resulting prediction?

Saliency: Extremely fast and no additional annotations!

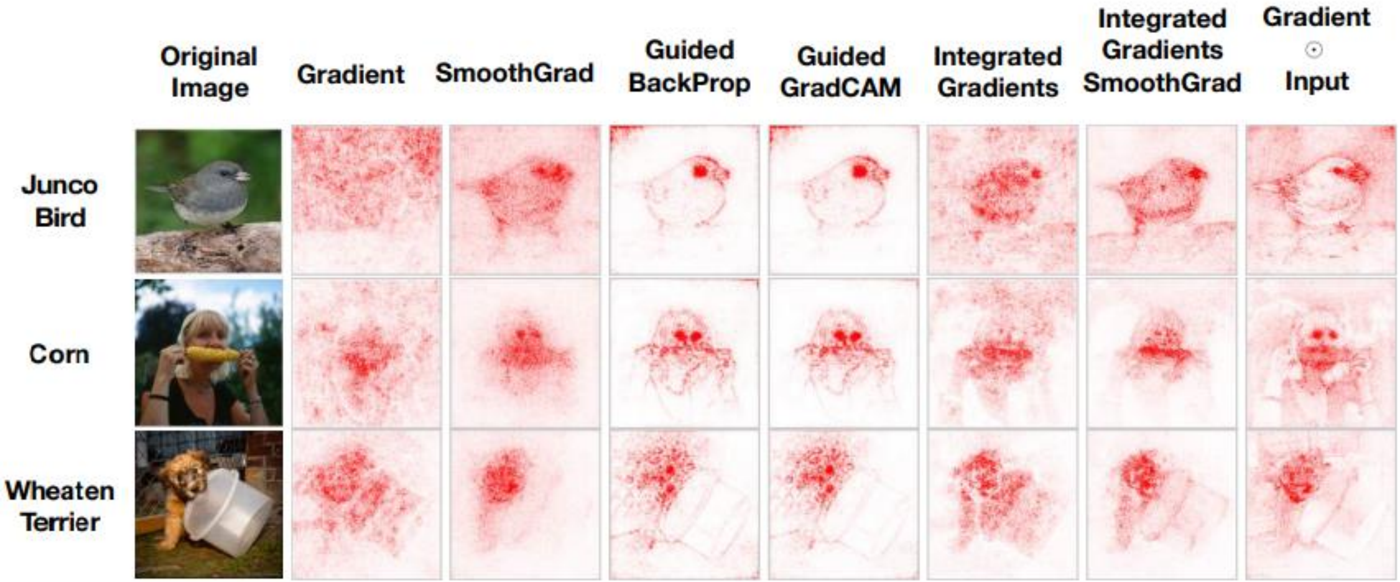


[Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.]

Saliency: Some Criticism

Saliency is popular!

Many different / advanced ways to compute saliency responses



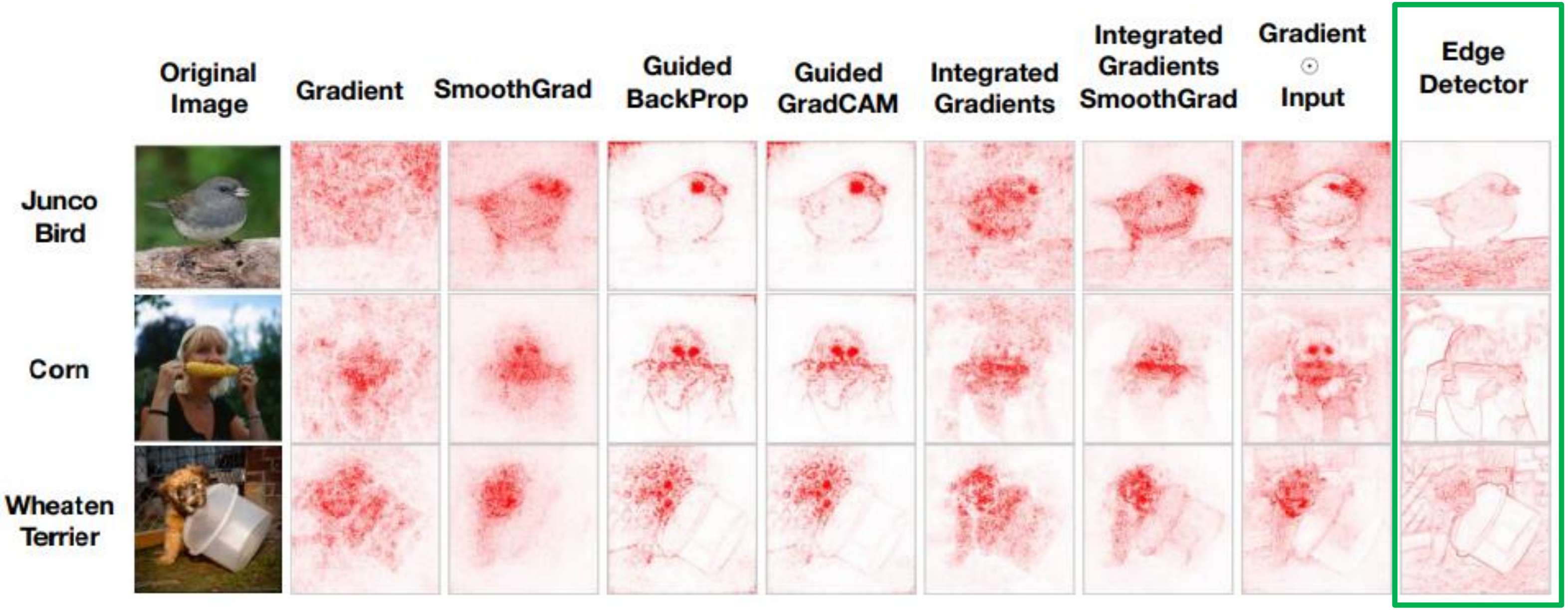
Q: Do you notice anything interesting?

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. NeurIPS (pp. 9505-9515).

Saliency: Some Criticism

Saliency is popular!

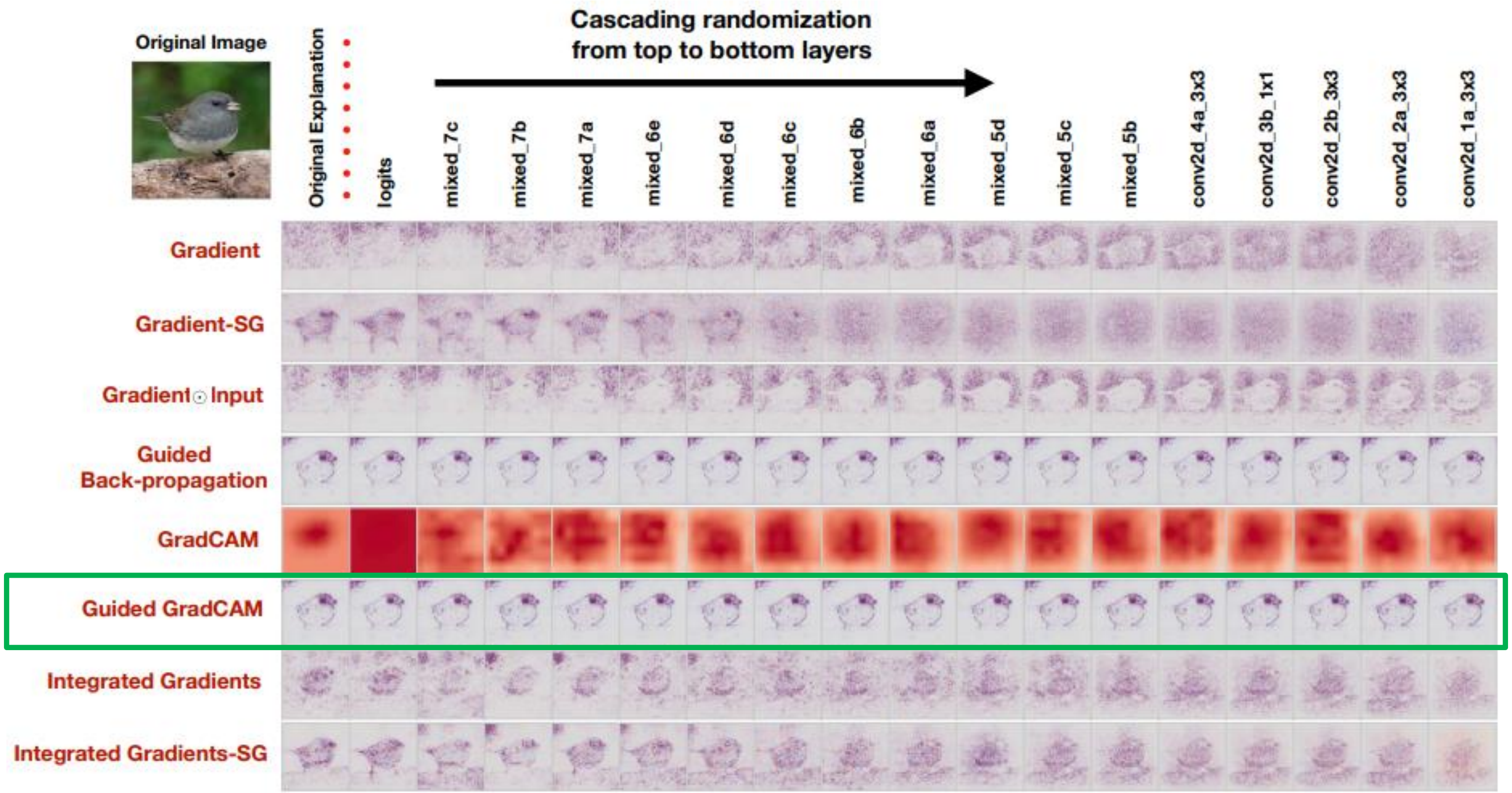
Many different / advanced ways to compute saliency responses



These results look surprisingly similar to simple edge detection…

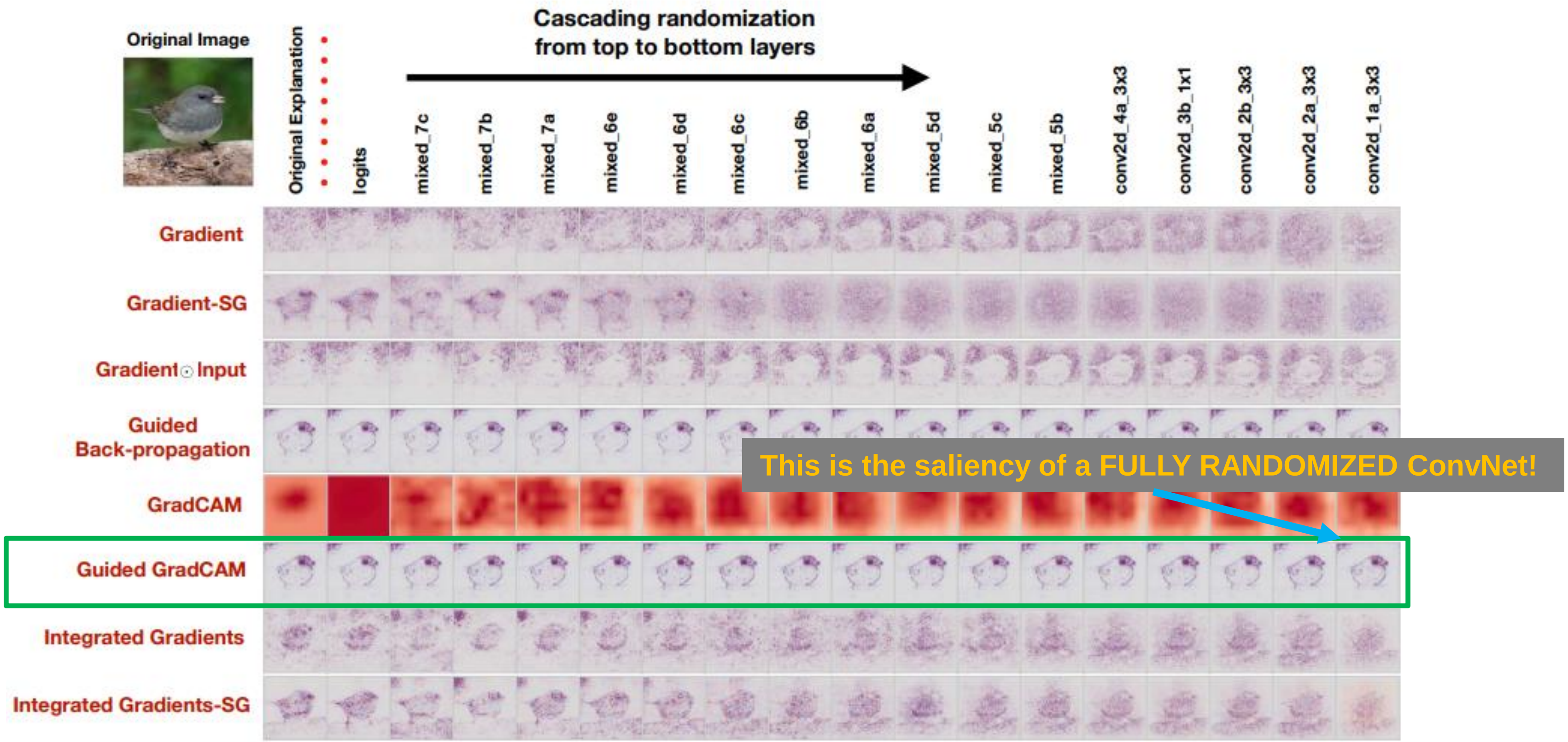
Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. NeurIPS (pp. 9505-9515).

Saliency: Some Criticism



Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. NeurIPS (pp. 9505-9515).

Saliency: Some Criticism



Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. NeurIPS (pp. 9505-9515).

Saliency: Some Criticism

Saliency is popular!

Many different / advanced ways to compute saliency responses

Why is this interesting / problematic?

- Interesting, because some image features seem to propagate very easily through ConvNets
- Problematic, because the general public (including medical field) increasingly relies on such visualizations
  → **Correlation vs. causation!**

We need to further investigate explanation methods

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. NeurIPS (pp. 9505-9515).

Explanations vs Interpretability

- Many (if not most) recent ML models are black boxes
  - Such ML models do not (and cannot) explain their predictions
  - Lack of transparency, accountability
  - Humans cannot understand
  - → Poor, high impact decisions in sensitive areas (criminal justice, medicine, …)

- Explanations – Additional *post hoc* model to explain the black box
  Problems:
  - Explanations are not faithful. If they were, the explanation itself would be sufficient
  - Incorrect explanations reduce trust
  - Many more challenges

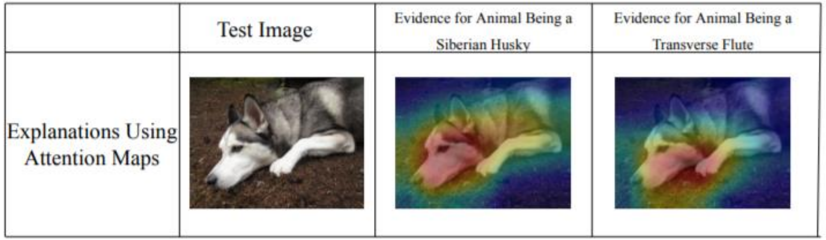


Figure 2: Saliency does not explain anything except where the network is looking. We have no idea why this image is labeled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Figure credit: Chaofan Chen and [28].

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Explanations vs Interpretability

- Interpretability
  - Domain specific notion
  - Constrained in model form (e.g. following structural knowledge and constraints)
  - Often, sparse models (few parameters): Allow to observe variable interaction

COMPAS	CORELS
black box 130+ factors might include socio-economic info expensive (software license), within software used in U.S. Justice System	full model is in Figure 3 only age, priors, (optional) gender no other information free, transparent

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

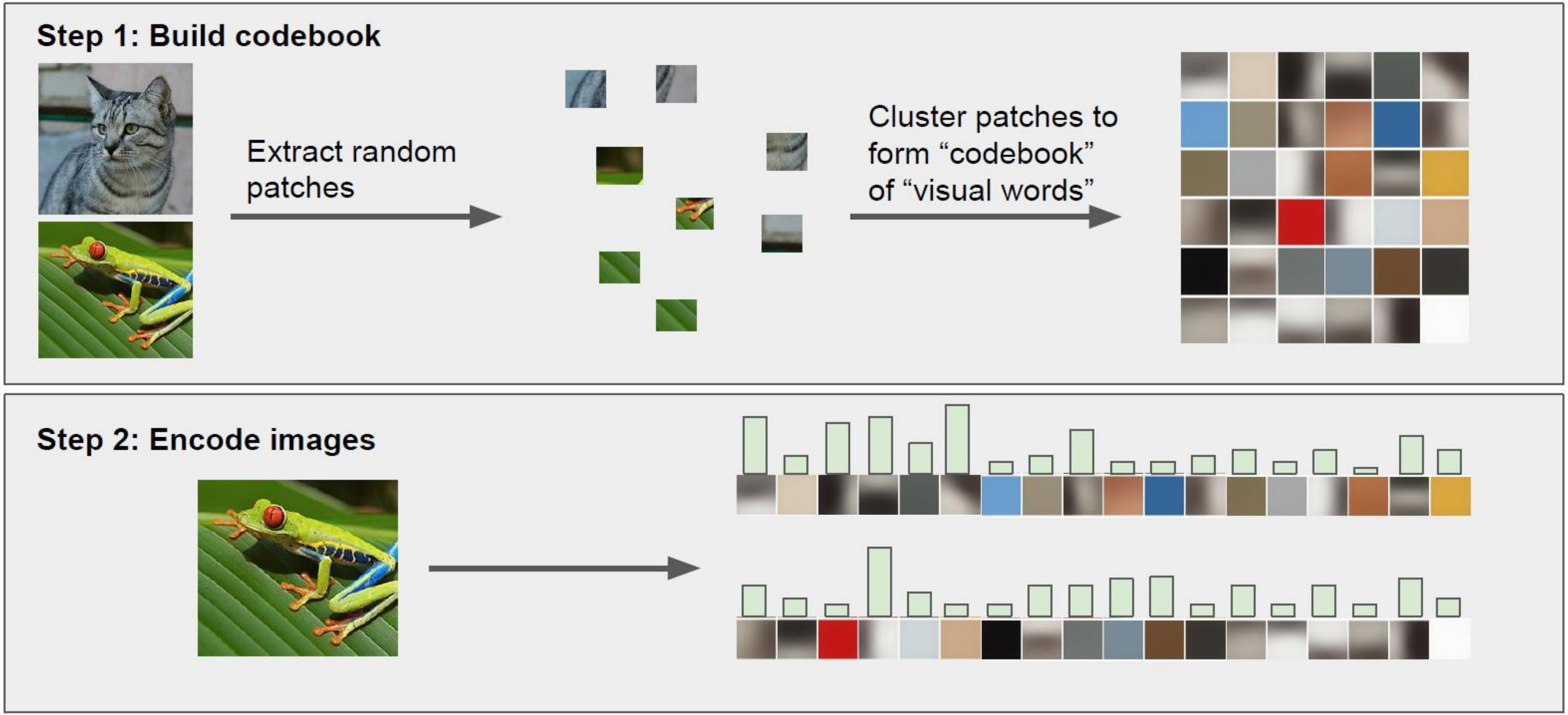
- Why not more interpretability?
  - Perceived trade-off between interpretability and performance (this is most often incorrect)
  - Corporations can capitalize on black box models
  - Hard to construct (domain knowledge, constrained problems, …)

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Interpretability, Generalization and Domain Gaps

Bag of Features - BagNet

Bag-of-Features?



[Fei-Fei, L., & Perona, P. (2005) A bayesian hierarchical model for learning natural scene categories. CVPR]

Images from Stanford cs231n.

Bag-of-Features?

Deep Features

- Input Q*Q px image patch
- BagNet generates 2048-dim feature vector
- Linear classifier on this 2048-dim vector
- After linear classification: #classes heatmaps of “class evidence”
- Average class evidence and pass into softmax → Probability

[Brendel, W., & Bethge, M. (2018). Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. ICLR.]

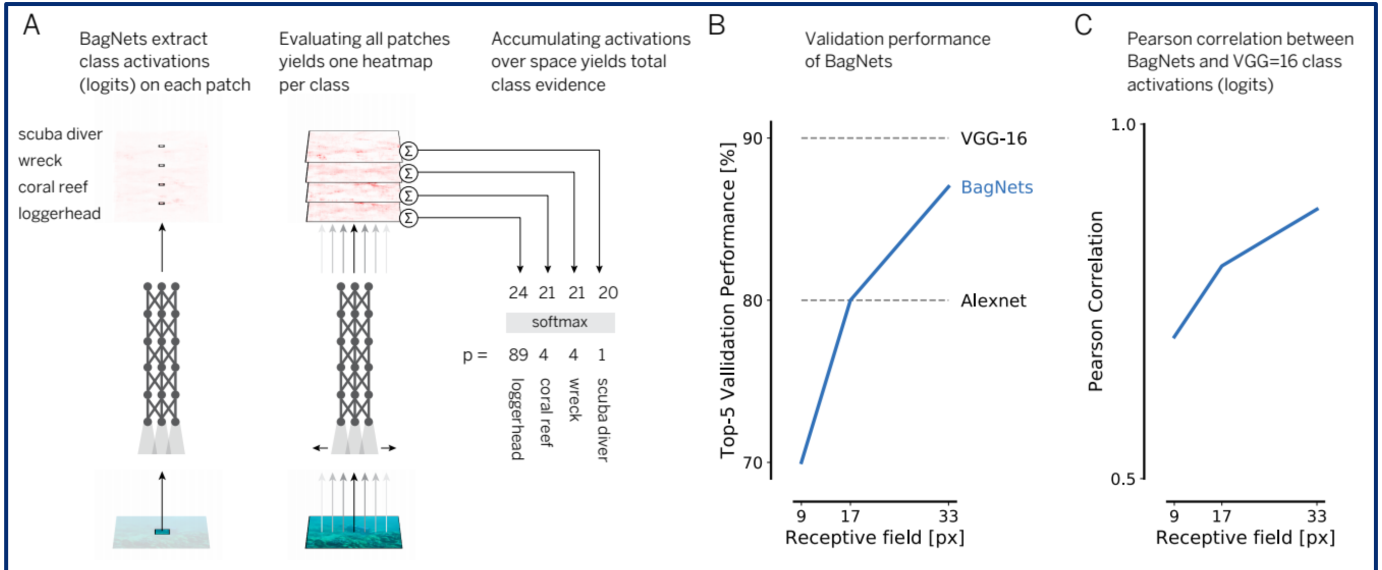
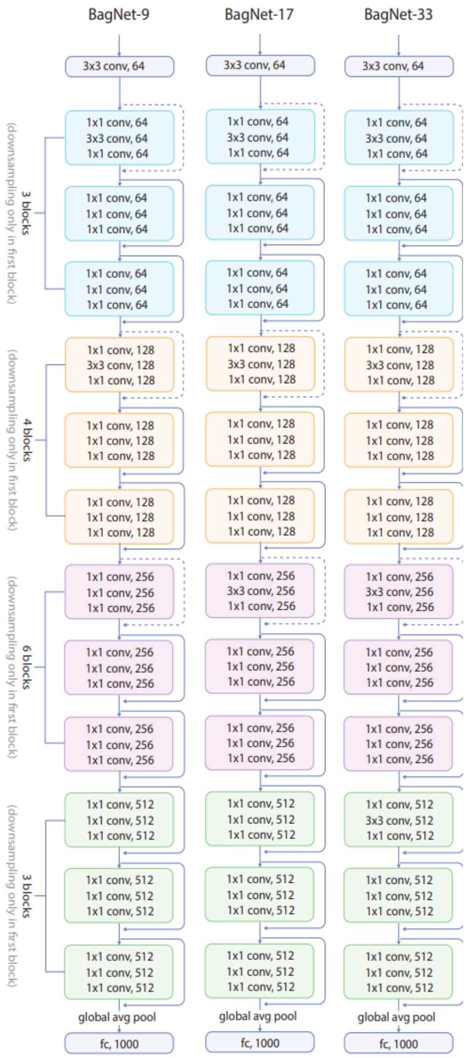
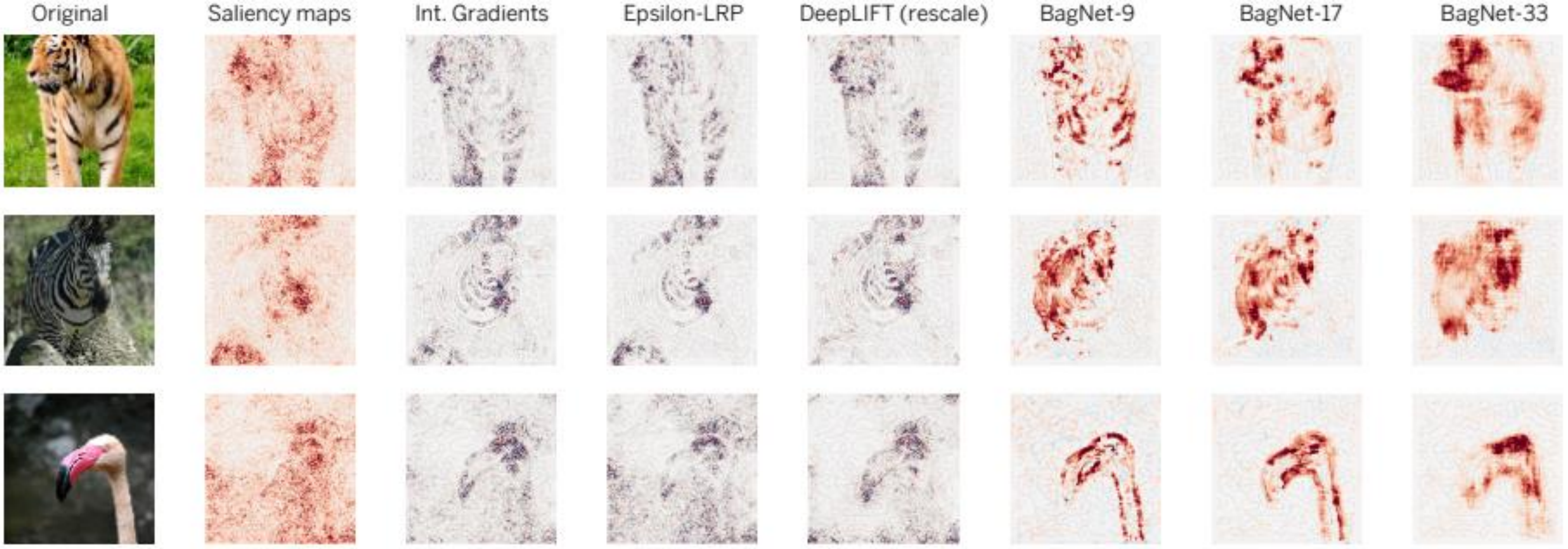


Figure 1: Deep bag-of-features models (BagNets). (A) The models extract features from small image patches which are each fed into a linear classifier yielding one logit heatmap per class. These heatmaps are averaged across space and passed through a softmax to get the final class probabilities. (B) Top-5 ImageNet performance over patch size. (C) Correlation with logits of VGG-16.



- BagNet architecture is ResNet like with 1x1 conv to limit receptive field
- Surprisingly good performance!

Brendel, W., & Bethge, M. (2018). Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. ICLR.



- Similarities in “decision making” (saliency) to SOTA models suggest that current network architectures base their decisions on a large number of weak and local statistical regularities

- One way forward is to define novel tasks that cannot be solved using local statistical regularities

Brendel, W., & Bethge, M. (2018). Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. ICLR.

Interpretability, Generalization and Domain Gaps

Texture vs. Shape

Navigating the Texture-shape Cue Conflict

(a) Texture image		(b) Content image		(c) Texture-shape cue conflict
81.4%	**Indian elephant**	71.1%	**tabby cat**	
10.3%	indri	17.3%	grey fox	
8.2%	black swan	3.3%	Siamese cat	

Q: What’s this?

Figure 1: Classification of a standard ResNet-50 of **(a)** a texture image (elephant skin: only texture cues); **(b)** a normal image of a cat (with both shape and texture cues), and **(c)** an image with a texture-shape cue conflict, generated by style transfer between the first two images.

[Geirhos, R., et al. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. ICLR.](#)

Navigating the Texture-shape Cue Conflict

(a) Texture image		(b) Content image		(c) Texture-shape cue conflict	
81.4%	**Indian elephant**	71.1%	**tabby cat**	63.9%	**Indian elephant**
10.3%	indri	17.3%	grey fox	26.4%	indri
8.2%	black swan	3.3%	Siamese cat	9.6%	black swan

Figure 1: Classification of a standard ResNet-50 of **(a)** a texture image (elephant skin: only texture cues); **(b)** a normal image of a cat (with both shape and texture cues), and **(c)** an image with a texture-shape cue conflict, generated by style transfer between the first two images.

[Geirhos, R., et al. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. ICLR.]

Bias Towards Texture

Observations (CNN = ResNet-50)

- CNN trained on Sylized-ImageNet (SIN): 79% top-5
- CNN trained on regular ImageNet (rIN): 93% top-5

→ SIN much harder than rIN: Textures are no longer predictive!

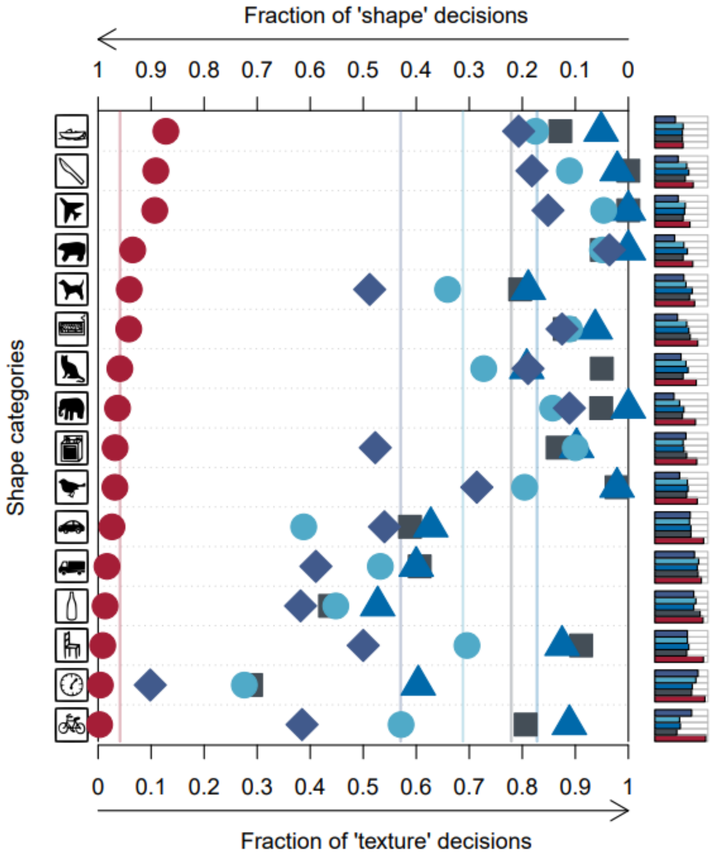
Figure 3: Visualisation of Stylized-ImageNet (SIN), created by applying AdaIN style transfer to ImageNet images. Left: randomly selected ImageNet image of class `ring-tailed lemur`. Right: ten examples of images with content/shape of left image and style/texture from different paintings. After applying 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.

Generalization

- rIN-trained on SIN: 16% top-5
- SIN-trained on rIN: 83% top-5

[Geirhos, R., et al. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. ICLR.]

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.



→ More plausible models of human visual object recognition

Geirhos, R., et al. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. ICLR.

Interpretability, Generalization and Domain Gaps

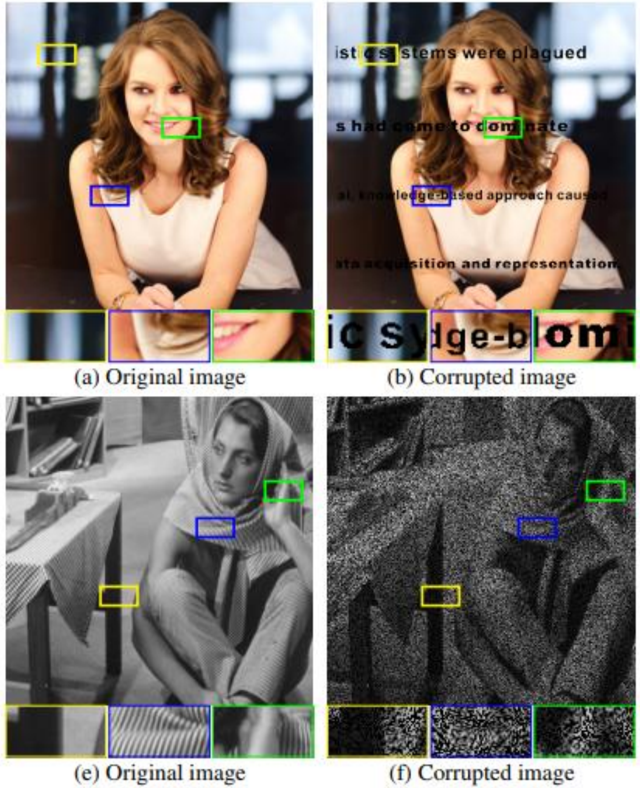
Architecture as Prior?

Deep Image Prior

- Image restoration $x^{\star} = \min_{x} E(x, x_0) + R(x)$, where

  E is a task-dependent data term (similarity),
  $\mathbf{x_0}$ is a corrupted observation (noise, occlusion, etc), and
  R is a regularizer (e.g. total variation, L2, wavelets,…)

- Obvious choice: Train a network to restore images
  - Training on large datasets will yield a good **image prior**
  - For generalization, network structure must “resonate” with data structure



(a) Original image (b) Corrupted image

(e) Original image (f) Corrupted image

Interestingly, learning is NOT required for building good image priors!

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2018). Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9446-9454).

Deep Image Prior

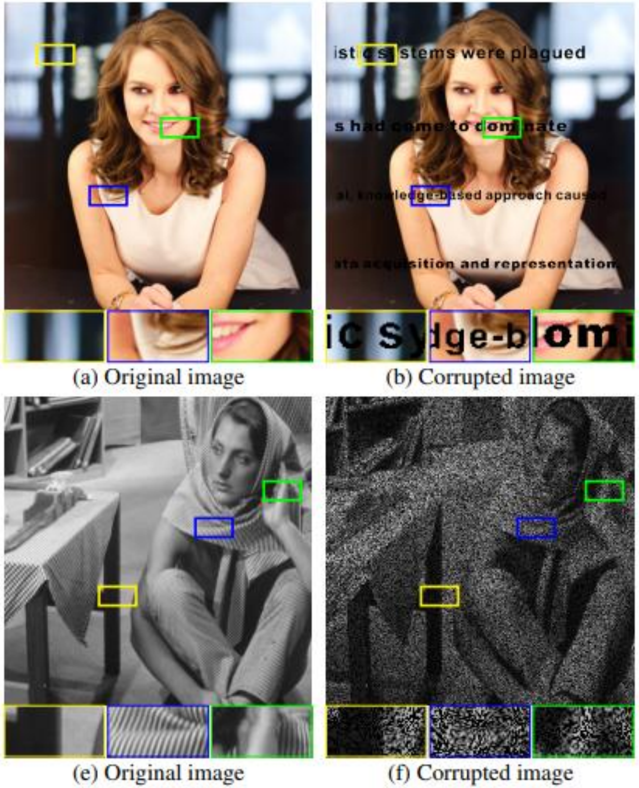
- Image restoration $x^{\star} = \min_{x} E(x, x_0) + R(x)$, where

  E is a task-dependent data term (similarity),
  $\mathbf{x_0}$ is a corrupted observation (noise, occlusion, etc), and
  R is a regularizer (e.g. total variation, L2, wavelets,…)

- Regularizer is replaced by a CNN

$$\theta^{\star} = \min_{\theta} E(f_{\theta}(z), x_0), \quad x^{\star} = f_{\theta^{\star}}(z)$$

  Minimizer is optimized using a random z, starting from random initialization of θ



(a) Original image
(b) Corrupted image
(e) Original image
(f) Corrupted image

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2018). Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9446-9454).

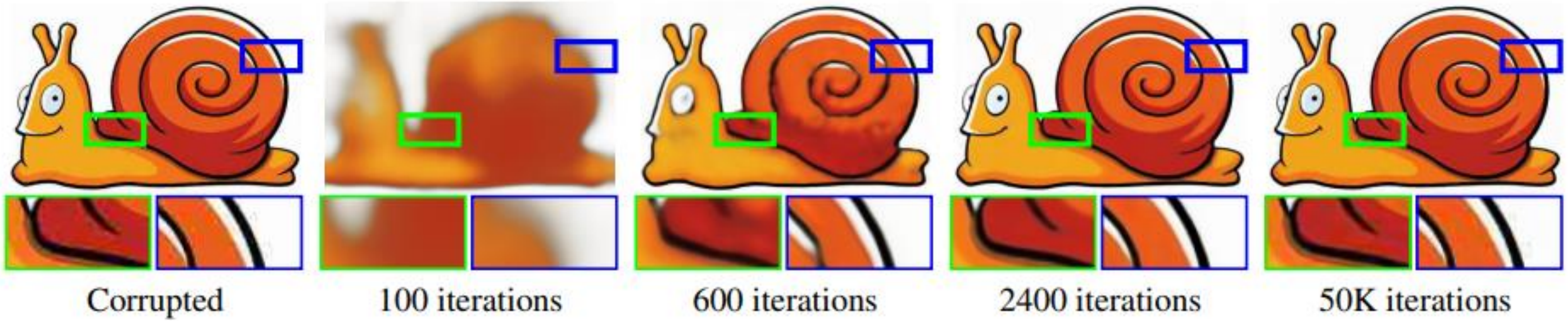


Figure 3: **Blind restoration of a JPEG-compressed image.** (*electronic zoom-in recommended*) Our approach can restore an image with a complex degradation (JPEG compression in this case). As the optimization process progresses, the deep image prior allows to recover most of the signal while getting rid of halos and blockiness (after 2400 iterations) before eventually overfitting to the input (at 50K iterations).

- Optimization is fast for “natural images” without corruption
- CNN can fit corruption, but very reluctantly

- Parametrization (CNN architecture) has high impedance to noise and low impedance to signal!

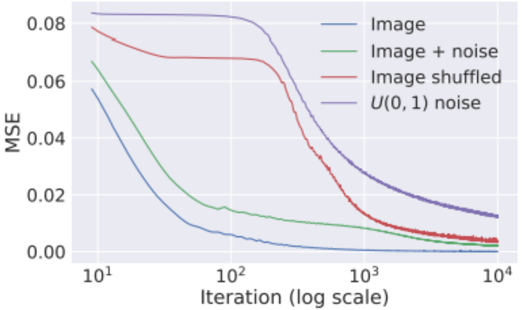


Figure 2: Learning curves for the reconstruction task using: a natural image, the same plus i.i.d. noise, the same randomly scrambled, and white noise. Naturally-looking images result in much faster convergence, whereas noise is rejected.

[Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2018). Deep image prior. CVPR (pp. 9446-9454).]

(a) Original image (b) Corrupted image (c) Shepard networks [27] (d) Deep Image Prior

(e) Original image (f) Corrupted image (g) [25], PSNR = 28.1 (h) Deep Img. Prior, PSNR = 32.22

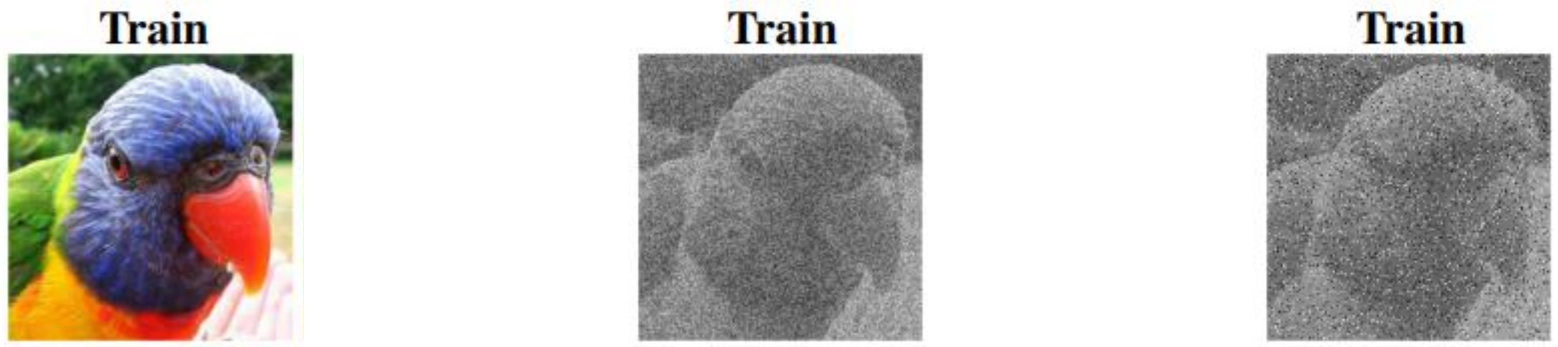
Figure 7: **Comparison with two recent inpainting approaches**. Top – comparison with Shepard networks [27] on text inpainting example. Bottom – comparison with convolutional sparse coding [25] on inpainting 50% of missing pixels. In both cases, our approach performs better on the images used in the respective papers.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2018). Deep image prior. CVPR (pp. 9446-9454).

Interpretability, Generalization and Domain Gaps

Generalization

Generalization

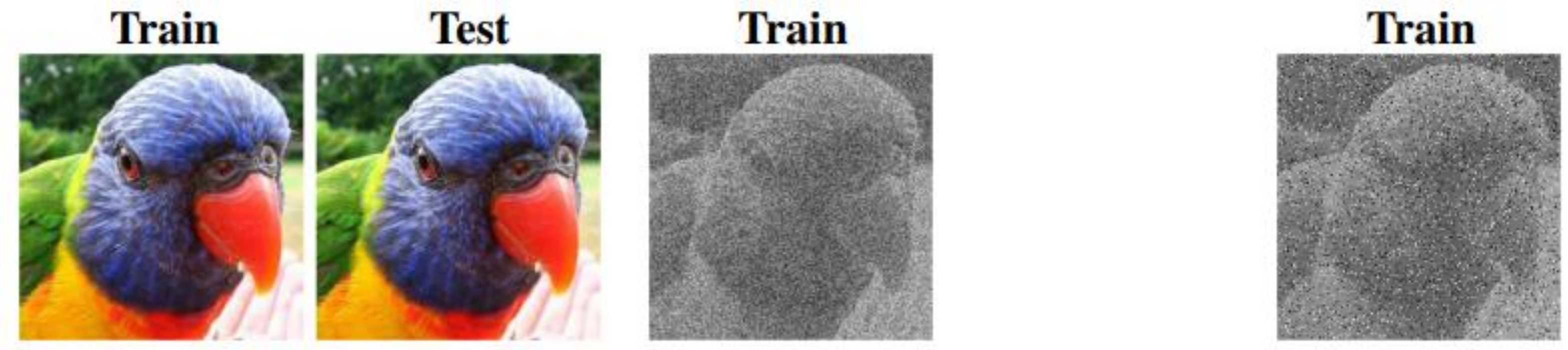


An easy test: This is an example for “Bird”

Consider this your *training stage*. **This is a “Bird”.**

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).

Generalization

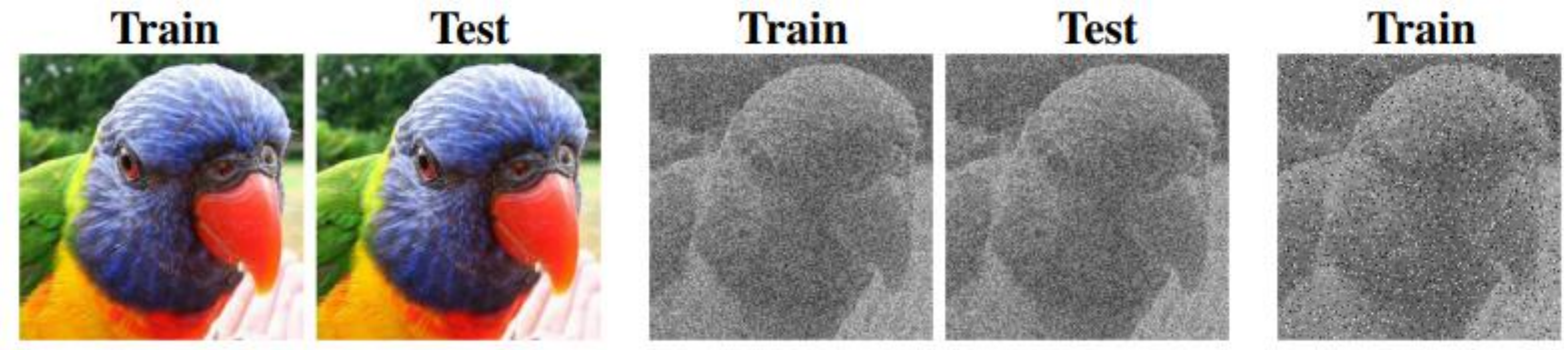


An easy test: This is an example for “Bird”

Consider this your ***testing*** *stage.* **What is this?**

[Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).]

Generalization

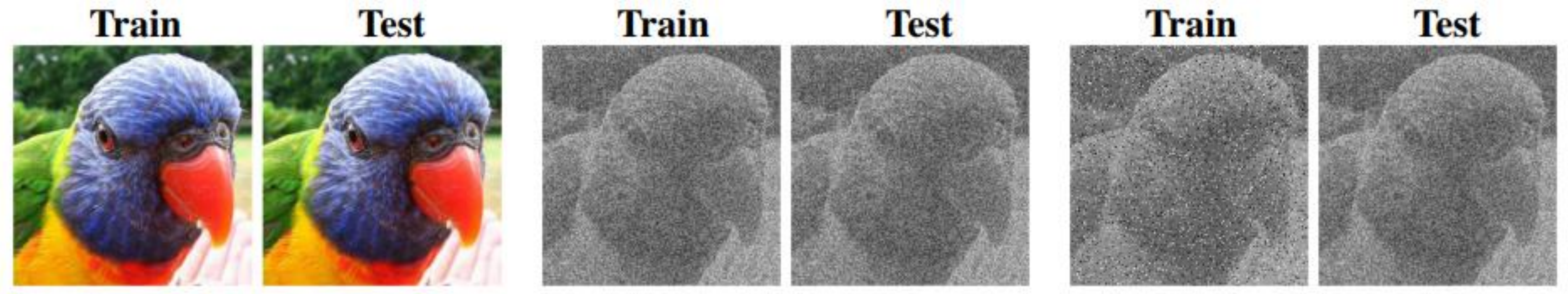


An easy test: This is an example for "Bird"

Consider this your ***testing*** *stage.* **What is this?**

[Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).]

Generalization



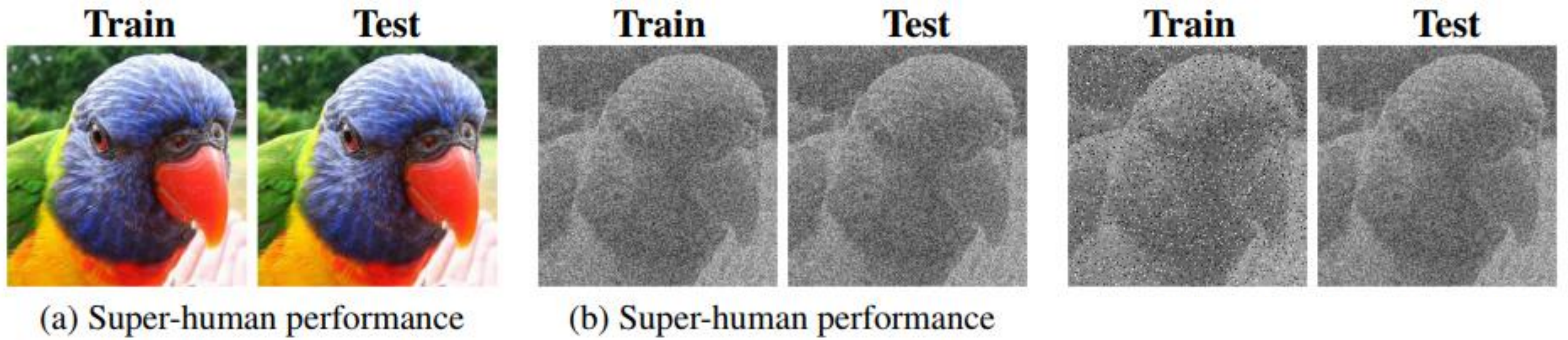
An easy test: This is an example for “Bird”

Consider this your ***testing*** *stage.* **What is this?**

“Bird”, too! Easy right? **So how would a CNN do?**

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).

Generalization



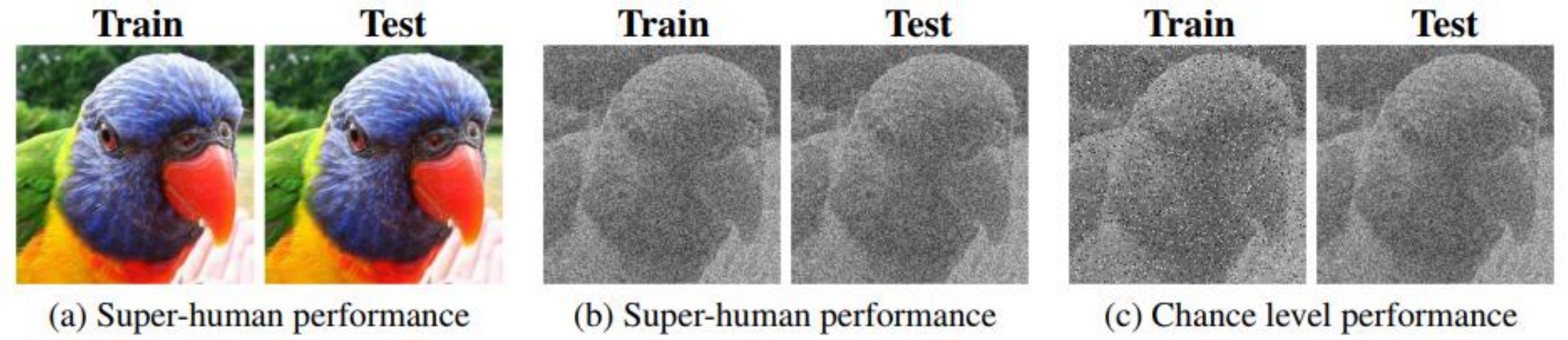
An easy test: This is an example for “Bird”

Consider this your ***testing*** *stage.* **What is this?**

“Bird”, too! Easy right? **So how would a CNN do?**

[Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).]

Generalization



An easy test: This is an example for “Bird”

Consider this your ***testing*** *stage.* **What is this?**

“Bird”, too! Easy right? **So how would a CNN do?**

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).

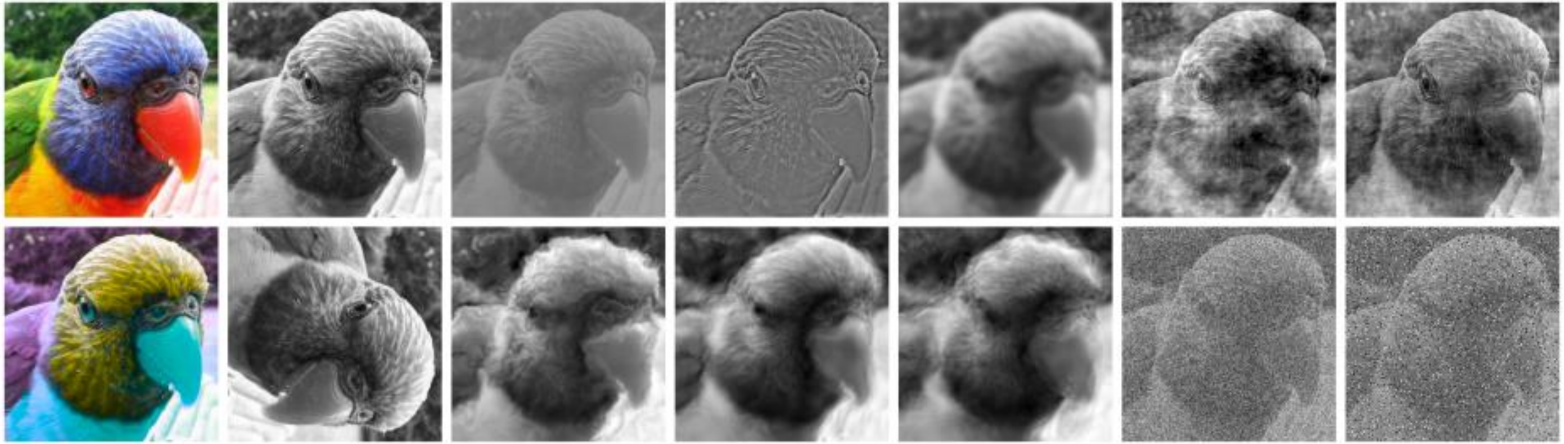
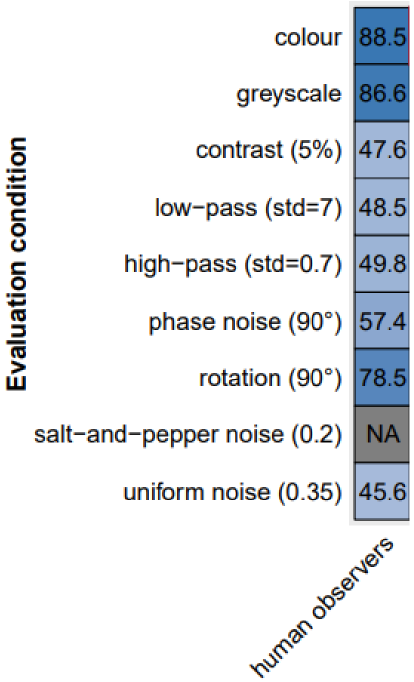


Figure 2: Example stimulus image of class bird across all distortion types. From left to right, image manipulations are: colour (undistorted), greyscale, low contrast, high-pass, low-pass (blurring), phase noise, power equalisation. Bottom row: opponent colour, rotation, Eidolon I, II and III, additive uniform noise, salt-and-pepper noise. Example stimulus images across all used distortion levels are

- Used distortions: See above
- Reduced ImageNet by condensing 1000 classes into 16 entry-level categories

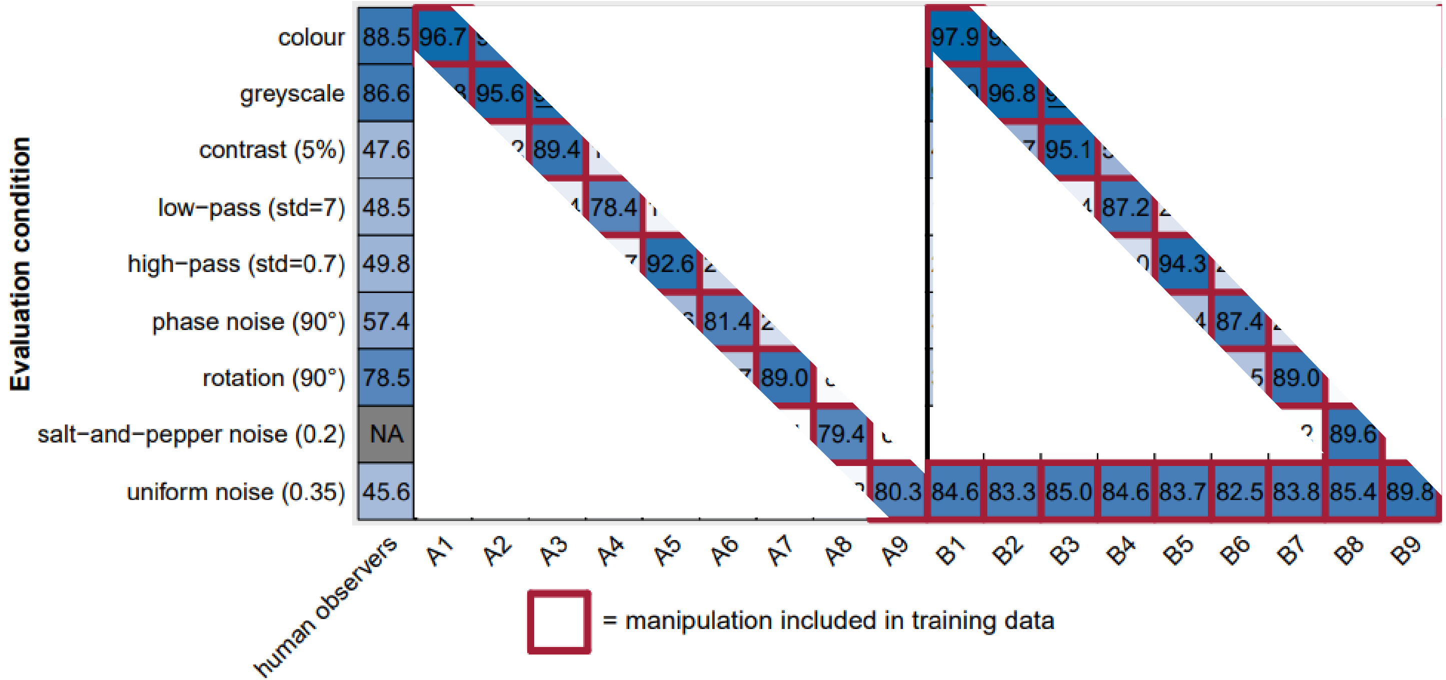
- 5 – 6 human raters “vs.” ResNet-50 architecture

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).



- 16 classes: Chance is at 6.25%
- Humans were presented the images for 200 ms

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).



- 16 classes: Chance is at 6.25%
- Humans were presented the images for 200 ms
- CNNs on the same data they were trained on: **Super-human performance!**
  Q: Generalization→ imperfect knowledge of testing conditions. How do CNNs perform there?

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).

Evaluation condition	human observers	A1	A2	A3	A4	A5	A6	A7	A8	A9	B1	B2	B3	B4	B5	B6	B7	B8	B9
colour	88.5	96.7	90.6	50.0	83.1	86.1	84.2	90.8	10.4	8.1	97.9	95.4	72.3	93.0	91.1	92.4	94.9	10.2	11.2
greyscale	86.6	87.8	95.6	94.1	86.2	93.2	87.8	90.5	10.3	9.8	94.0	96.8	96.2	93.3	95.7	94.3	90.9	11.4	12.8
contrast (5%)	47.6	13.1	14.2	89.4	19.6	39.8	17.1	10.2	28.6	29.0	46.3	51.7	95.1	50.5	79.1	59.4	45.2	34.6	37.9
low−pass (std=7)	48.5	18.9	16.1	16.4	78.4	11.9	16.0	9.8	6.9	6.6	16.0	18.6	14.4	87.2	20.5	13.8	13.5	7.1	9.3
high−pass (std=0.7)	49.8	21.1	24.7	29.9	11.7	92.6	27.7	8.3	10.4	20.6	25.1	22.8	29.2	25.0	94.3	27.5	28.3	18.9	19.8
phase noise (90°)	57.4	23.3	28.3	31.2	27.0	46.6	81.4	24.4	7.4	8.9	30.8	31.4	30.6	31.4	43.4	87.4	24.1	7.8	7.6
rotation (90°)	78.5	36.5	43.3	39.9	31.8	40.4	37.7	89.0	8.5	8.0	38.5	41.9	40.3	35.2	40.1	40.5	89.0	8.3	8.8
salt−and−pepper noise (0.2)	NA	6.1	6.4	5.8	7.9	6.2	6.2	6.4	79.4	6.2	6.2	6.1	6.3	5.4	5.8	5.7	6.2	89.6	6.2
uniform noise (0.35)	45.6	6.2	7.3	6.9	9.0	7.3	6.2	6.0	10.2	80.3	84.6	83.3	85.0	84.6	83.7	82.5	83.8	85.4	89.8

□ = manipulation included in training data

- 16 classes: Chance is at 6.25%
- Humans were presented the images for 200 ms
- CNNs on the same data they were trained on: **Super-human performance!**

→ **As bad as chance level on unseen distortions!**
Q: What happens if we train on ALL BUT ONE distortion?

[Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).](#)

Evaluation condition	human observers	A1	A2	A3	A4	A5	A6	A7	A8	A9	B1	B2	B3	B4	B5	B6	B7	B8	B9	C1	C2
colour	88.5	96.7	90.6	50.0	83.1	86.1	84.2	90.8	10.4	8.1	97.9	95.4	72.3	93.0	91.1	92.4	94.9	10.2	11.2	95.5	95.9
greyscale	86.6	87.8	95.6	94.1	86.2	93.2	87.8	90.5	10.3	9.8	94.0	96.8	96.2	93.3	95.7	94.3	90.9	11.4	12.8	94.8	95.1
contrast (5%)	47.6	13.1	14.2	89.4	19.6	39.8	17.1	10.2	28.6	29.0	46.3	51.7	95.1	50.5	79.1	59.4	45.2	34.6	37.9	90.9	88.2
low−pass (std=7)	48.5	18.9	16.1	16.4	78.4	11.9	16.0	9.8	6.9	6.6	16.0	18.6	14.4	87.2	20.5	13.8	13.5	7.1	9.3	74.7	74.9
high−pass (std=0.7)	49.8	21.1	24.7	29.9	11.7	92.6	27.7	8.3	10.4	20.6	25.1	22.8	29.2	25.0	94.3	27.5	28.3	18.9	19.8	91.4	90.7
phase noise (90°)	57.4	23.3	28.3	31.2	27.0	46.6	81.4	24.4	7.4	8.9	30.8	31.4	30.6	31.4	43.4	87.4	24.1	7.8	7.6	82.9	82.6
rotation (90°)	78.5	36.5	43.3	39.9	31.8	40.4	37.7	89.0	8.5	8.0	38.5	41.9	40.3	35.2	40.1	40.5	89.0	8.3	8.8	80.1	80.5
salt−and−pepper noise (0.2)	NA	6.1	6.4	5.8	7.9	6.2	6.2	6.4	79.4	6.2	6.2	6.1	6.3	5.4	5.8	5.7	6.2	89.6	6.2	78.6	13.6
uniform noise (0.35)	45.6	6.2	7.3	6.9	9.0	7.3	6.2	6.0	10.2	80.3	84.6	83.3	85.0	84.6	83.7	82.5	83.8	85.4	89.8	11.0	71.5

= manipulation included in training data

- 16 classes: Chance is at 6.25%
- Humans were presented the images for 200 ms
- CNNs on the same data they were trained on: **Super-human performance!**

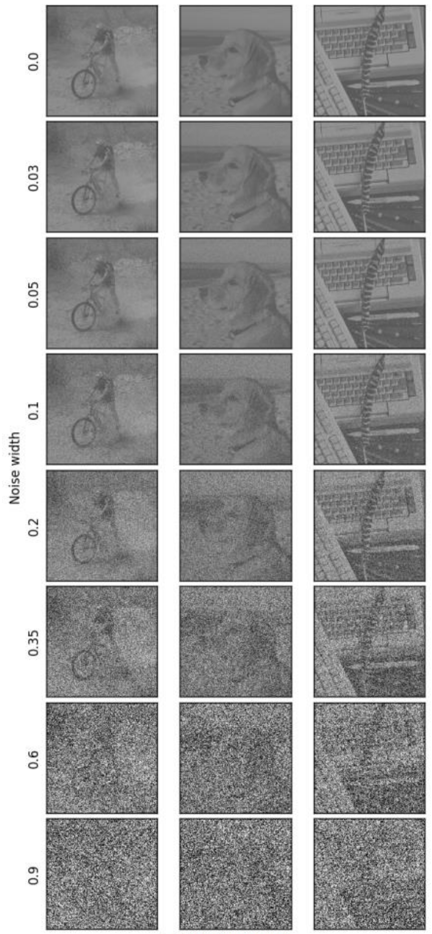
→ **As bad as chance level on unseen distortions!**

→ **Slightly better than chance!**

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).

Generalization

- Classification performance of SOTA CNNs declines rapidly with decreasing signal-to-noise ratio (not discussed)
- Progressively diverging patterns of classification errors (not discussed)
- **Strong generalization failure**
- **Solving this problem will be crucial for robust machine inference and better models of human object recognition**



(a) Uniform noise

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).

Interpretability, Generalization and Domain Gaps

Generalization: One More!

Generalization: One More!

Methods and experiments

- New test sets for CIFAR and ImageNet

Q: What do you think will happen?

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet?. arXiv preprint arXiv:1902.10811.

Generalization: One More!

Methods and experiments

- New test sets for CIFAR and ImageNet
- Decreased performance (3% – 14%)!

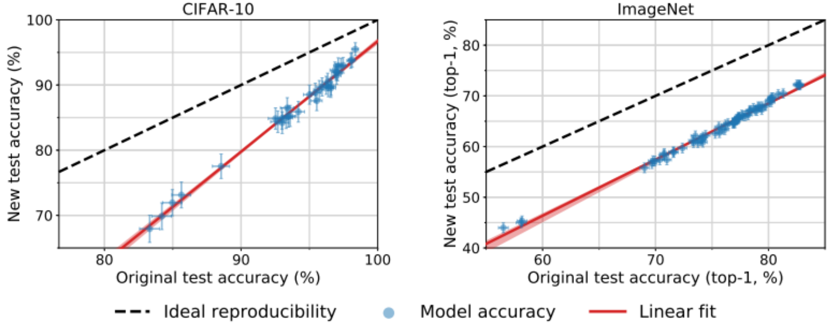


Figure 1: Model accuracy on the original test sets vs. our new test sets. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). The plots reveal two main phenomena: (i) There is a significant drop in accuracy from the original to the new test sets. (ii) The model accuracies closely follow a linear function with slope *greater* than 1 (1.7 for CIFAR-10 and 1.1 for ImageNet). This means that every percentage point of progress on the original test set translates into more than one percentage point on the new test set. The two plots

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet?. arXiv preprint arXiv:1902.10811.

Generalization: One More!

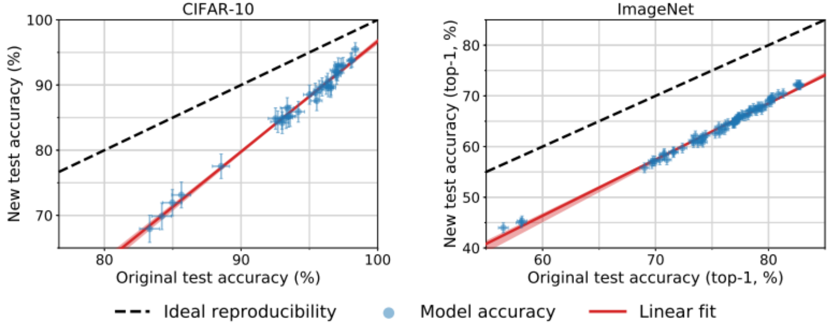


Figure 1: Model accuracy on the original test sets vs. our new test sets. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). The plots reveal two main phenomena: (i) There is a significant drop in accuracy from the original to the new test sets. (ii) The model accuracies closely follow a linear function with slope *greater* than 1 (1.7 for CIFAR-10 and 1.1 for ImageNet). This means that every percentage point of progress on the original test set translates into more than one percentage point on the new test set. The two plots

Methods and experiments

- New test sets for CIFAR and ImageNet
- Decreased performance (3% – 14%)!

Observations

- Models **better on original** test set also **better on new** test set
  → Non-diminishing returns (slope Fig. 1 greater than 1.0)!
  → Adaptivity (~overfitting) is not a likely explanation for accuracy drops

- **Leaving out “complicated” examples** in new test set: **Same accuracy**!
  → Even the best classifiers are sensitive to minutiae

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet?. arXiv preprint arXiv:1902.10811.

Generalization: One More!

Figure 1: Model accuracy on the original test sets vs. our new test sets. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). The plots reveal two main phenomena: (i) There is a significant drop in accuracy from the original to the new test sets. (ii) The model accuracies closely follow a linear function with slope *greater* than 1 (1.7 for CIFAR-10 and 1.1 for ImageNet). This means that every percentage point of progress on the ... test set. The two plots

Methods and experiments

- New test sets for CIFAR and ImageNet
- Decreased performance (3% – 14%)!

“Current classifiers still **do not generalize reliably** even in the benign environment of a carefully controlled reproducibility experiment.”

Observations

- Models **better on original** test set also **better on new** test set
  → Non-diminishing returns (slope Fig. 1 greater than 1.0)!
  → Adaptivity (~overfitting) is not a likely explanation for accuracy drops

- **Leaving out “complicated” examples** in new test set: **Same accuracy**!
  → Even the best classifiers are sensitive to minutiae

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet?. arXiv preprint arXiv:1902.10811.

Interpretability, Generalization and Domain Gaps

Questions?