AIMS Big Data Course
Universal, unsupervised and understandable representations

Dr Andrea Vedaldi
Dr Andrew Zisserman

For lecture notes and updates see
http://www.robots.ox.ac.uk/~vedaldi/teach.html

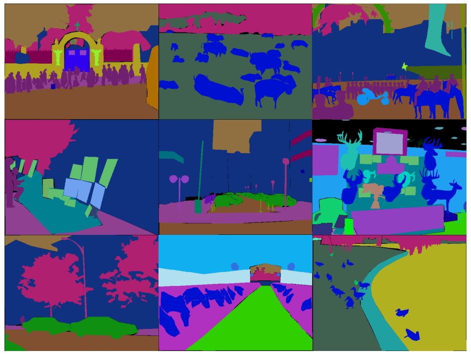
AIMS Big Data Course
Introduction to deep learning

Part 1: Data efficiency

Supervised learning

Images

Labelled concepts

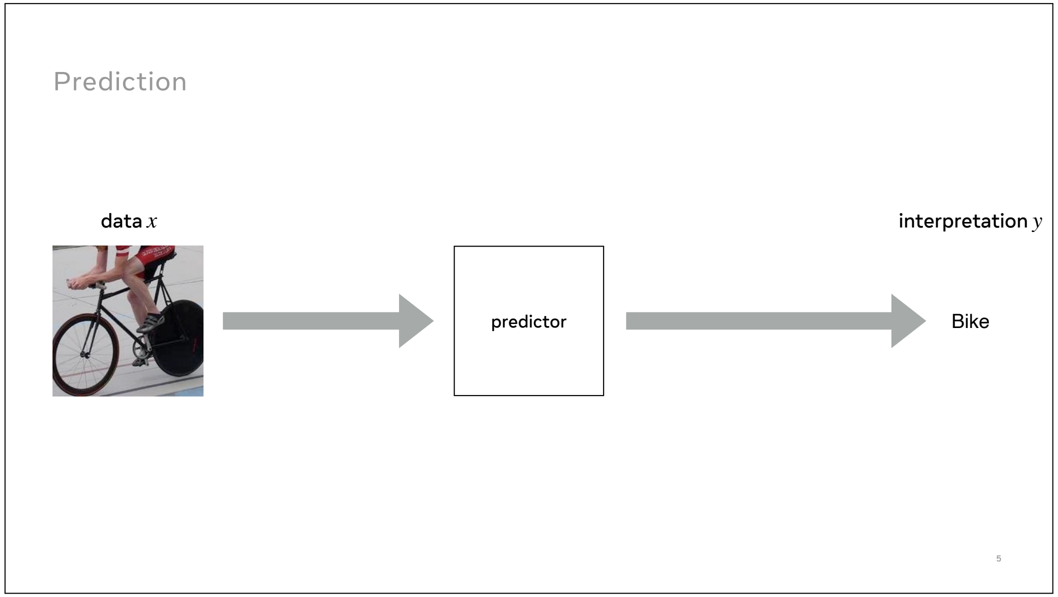


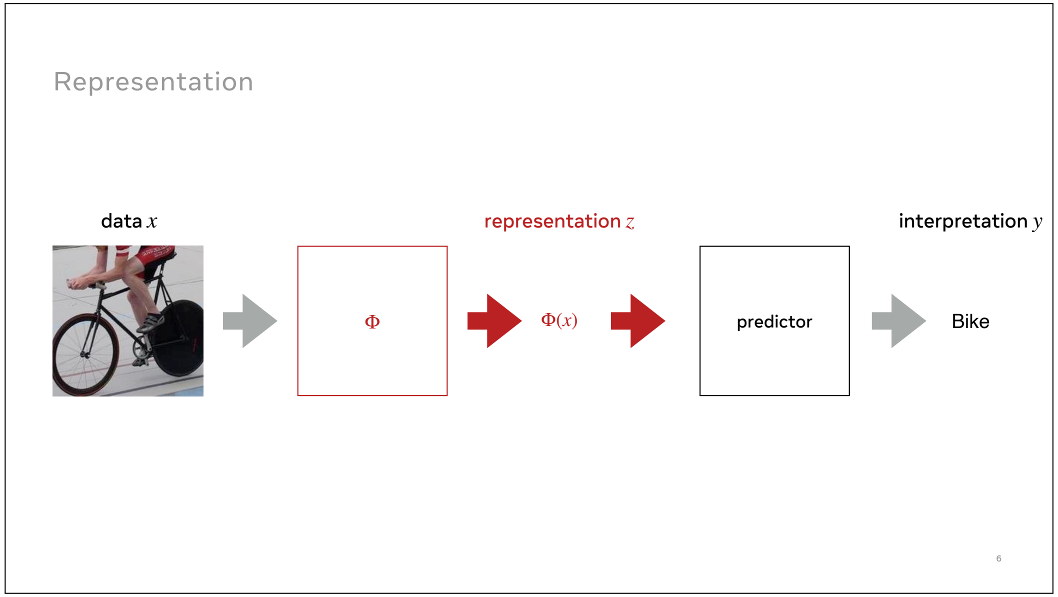
Scene parsing through ADE20K dataset. Zhou, Zhao, Puig, Fidler, Barriuso, Torralba. CVPR, 2017.

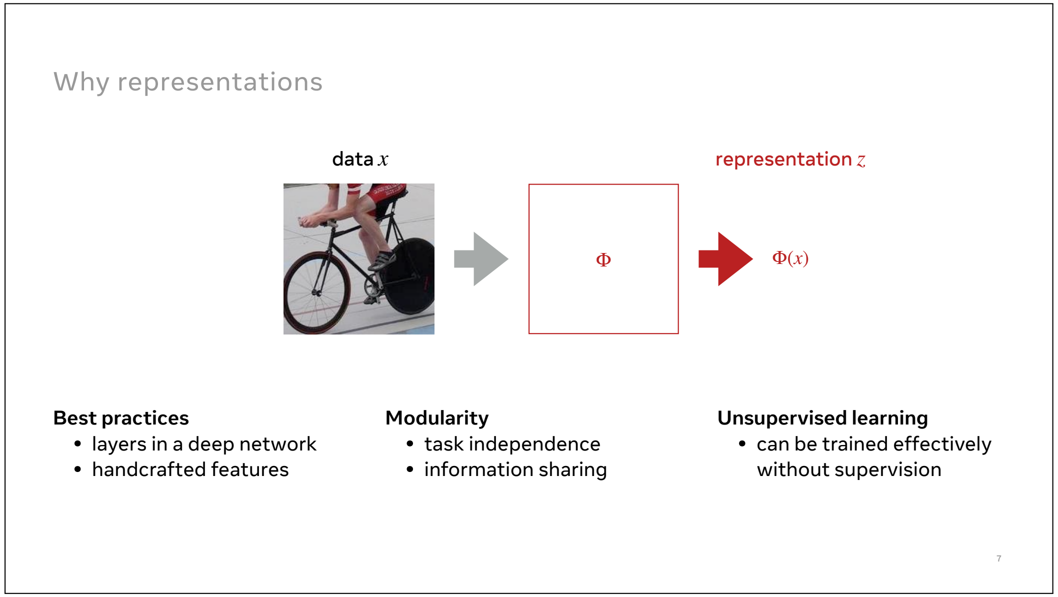
Learning without supervision

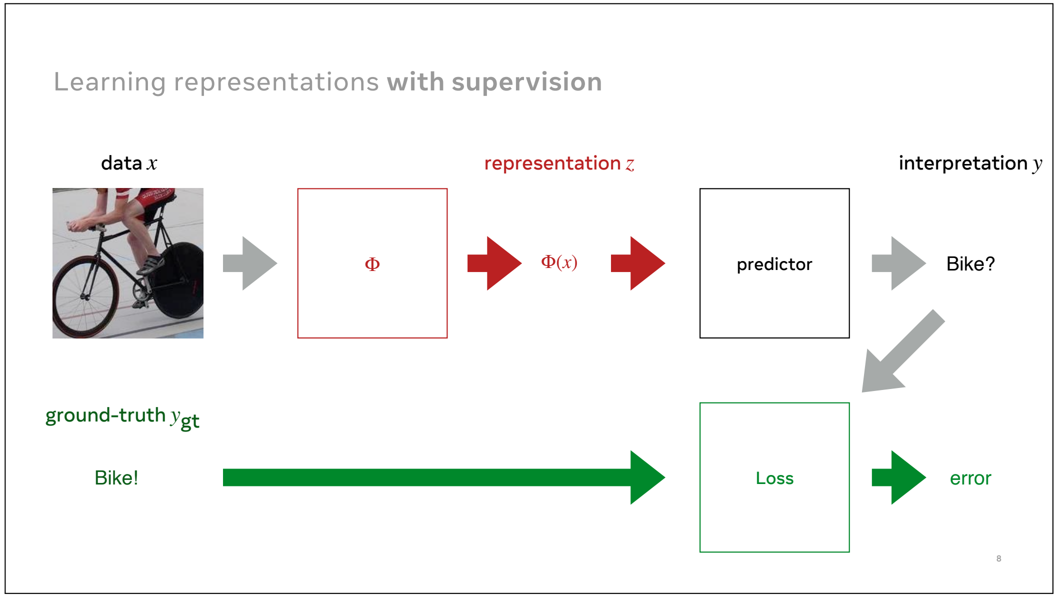
Images

Scene parsing through ADE20K dataset. Zhou, Zhao, Puig, Fidler, Barriuso, Torralba. CVPR, 2017.


Prediction
data x
interpretation y
predictor
Bike
5

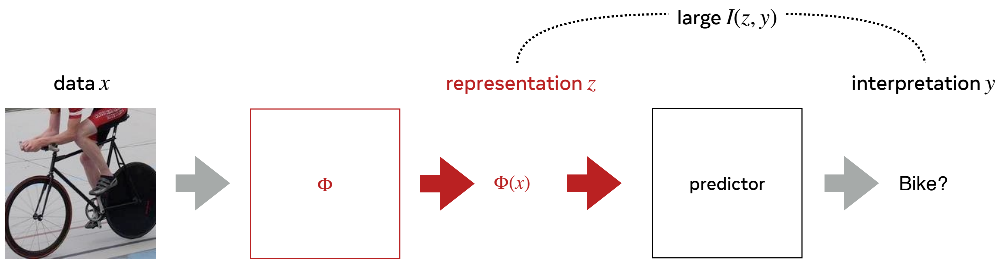

Representation
data x
representation z
interpretation y
Φ
$\Phi(x)$
predictor
Bike
6


Why representations
data x
representation z
Φ
$\Phi(x)$
Best practices
• layers in a deep network
• handcrafted features
Modularity
• task independence
• information sharing
Unsupervised learning
• can be trained effectively without supervision
7


Learning representations with supervision
data x
representation z
interpretation y
Φ
$\Phi(x)$
predictor
Bike?
ground-truth y_{gt}
Bike!
Loss
error
8

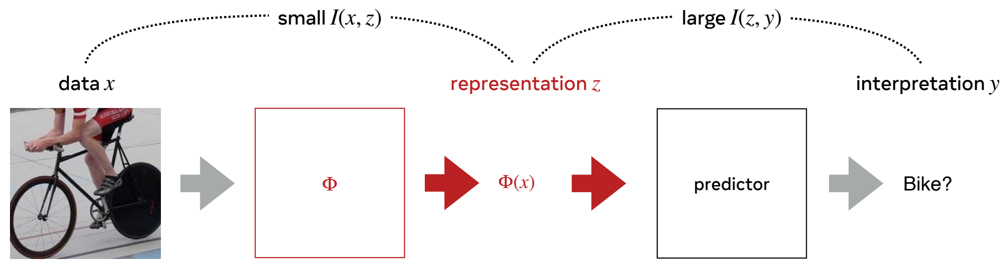
Using information to characterise representations

A representation should contain **information relevant** to the **prediction task**



The **information bottleneck** principle

$$\max_{\Phi} I(z, y) - \beta I(x, z)$$



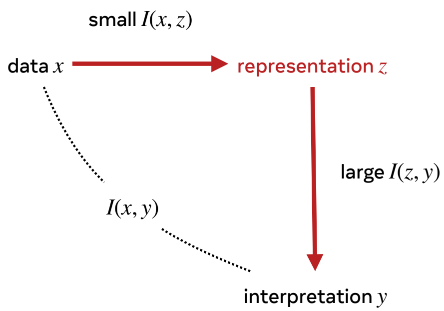
The information bottleneck method. Tishby, Pereira, Bialek. Allerton Conf. on Communication, Control and Computing, 1999
Deep learning and the information bottleneck principle. Tishby Zaslavsky. Information Theory Workshop, 2015

The **data processing inequality**

Representations cannot create information

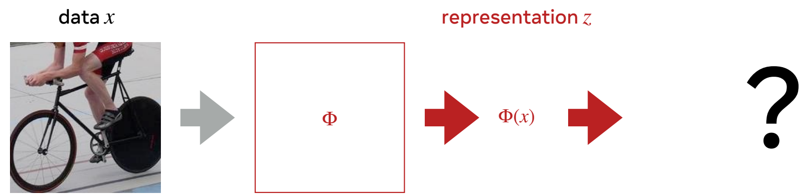
$$I(x, y) \geq I(z, y)$$

The representation should preserve the **relevant information** and discard the rest



Elements of Information Theory. Cover Thomas. Wiley, 2006

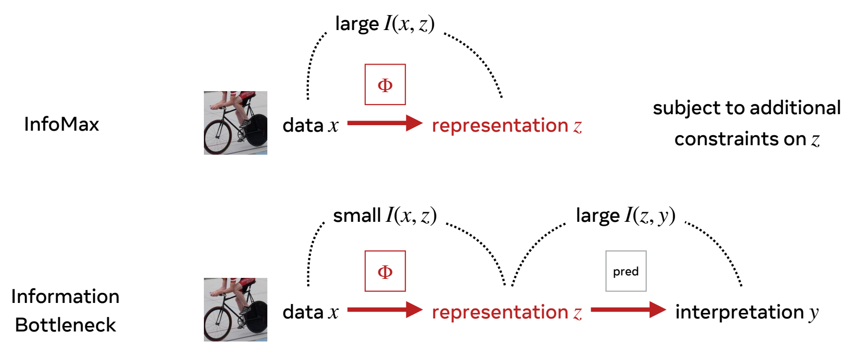
Learning representations **without supervision**



Lacking supervision, it is unclear what should be the aim of a representation

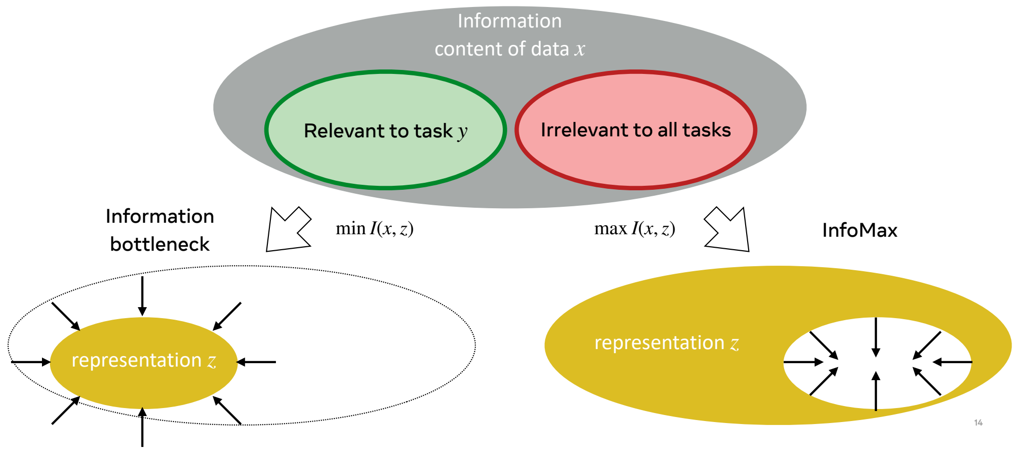
We require task-agnostic principles for learning them

The **Information Maximisation (InfoMax)** principle



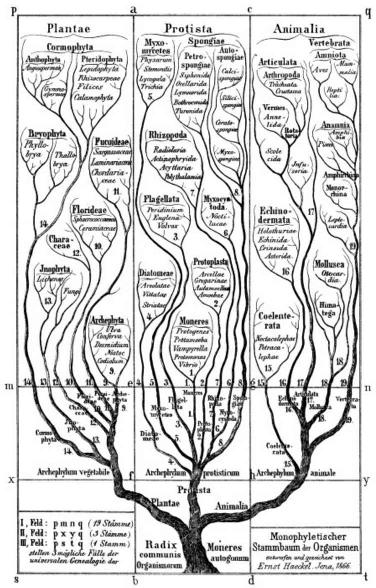
Self-organization in a perceptual network. Linsker. Computer, 21(3), 1988.

Information Bottleneck vs Maximisation



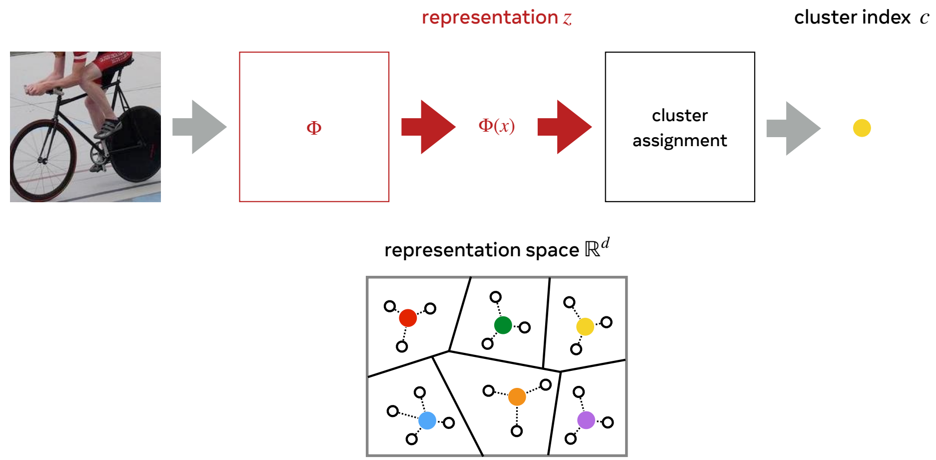
Learning interpretations

Can we make representations **easily interpretable**?

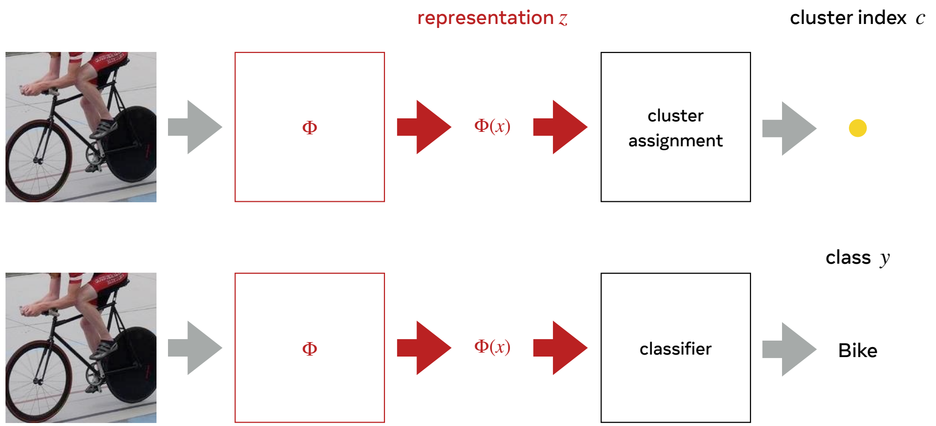


Ernst Haeckel [Wikimedia Commons]

Clustering representations



Clustering vs classification



Learning a classifier via cross-entropy minimisation

Standard **softmax classifier**

$$p(c\,|\,x) = \frac{\exp w_c^\top \Phi(x)}{\sum_k \exp w_k^\top \Phi(x)}$$

Learning **objective**:

$$\min_{\Phi,w} H(q,p)$$

Cross-entropy loss

$$H(q,p) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c} q(c\,|\,x_i)\log p(c\,|\,x_i)$$

where

$$q(c\,|\,x) = \delta(c, c_i)$$

is the **empirical distribution** of the **ground-truth labels**

Self-labelling via conditional entropy minimisation

Assume no g.t. is available

Replace the g.t. distribution with the predicted one

$$q(c\,|\,x_i) \longrightarrow p(c\,|\,x_i)$$

Learning **objective**:

$$\min_{\Phi,w} H(p,p)$$

I.e., replace the cross-entropy

$$H(q,p) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c} q(c\,|\,x_i)\log p(c\,|\,x_i)$$

with the **conditional entropy**

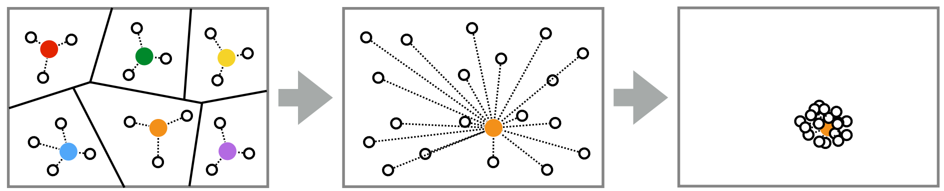
$$H(p,p) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c} p(c\,|\,x_i)\log p(c\,|\,x_i) = H(c\,|\,x)$$

Deep clustering for unsupervised learning of visual features. Caron, Bojanowski, Joulin, Douze. Proc. ECCV, 2018.

Conditional entropy minimisation

$$\min_{\Phi,w} H(c\,|\,x), \quad \text{where} \quad H(c\,|\,x) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c} p(c\,|\,x_i)\log p(c\,|\,x_i)$$

Mode collapse



InfoMax fixes mode collapse

Information vs. conditional entropy

$I(x, c) = H(c) - H(c \mid x)$

Maximising information maximises the **change in label entropy** before and after observing the input image

The largest possible value of entropy is:

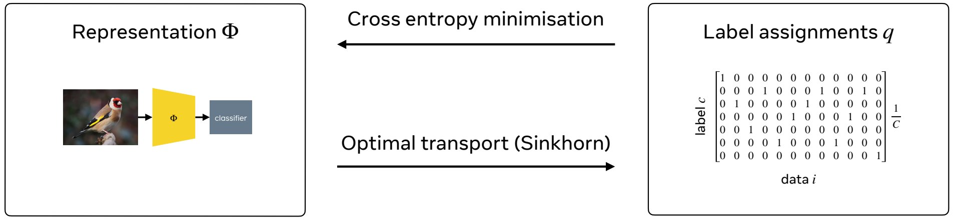
$$H(c) = \frac{1}{N}\sum_{i=1}^{N}\sum_{c} p(c \mid x_i) \ln \left(\frac{1}{N}\sum_{j=1}^{N} p(c \mid x_j) \right) \leq \ln C$$

In practice, we can use **constrained conditional entropy** minimization:

$$\min_{\Phi,w} H(c \mid x), \quad \text{subject to} \quad H(c) = \ln C$$

A practical implementation: Self-Labelling (SeLa)

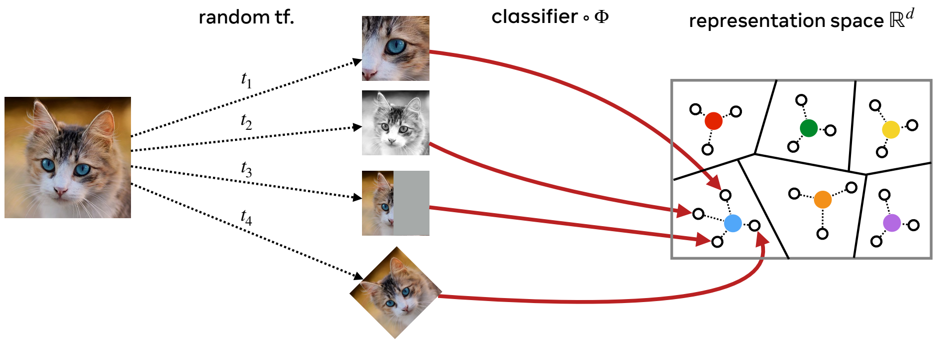
$$\min_{q,p} H(q,p) \quad \text{subject to} \quad \mathbb{E}[q(c \mid x)] = \frac{1}{C}$$



Self-labelling via simultaneous clustering and representation learning. Asano, Rupprecht, Vedaldi. ICLR, 2020.
Labelling unlabelled videos from scratch with multi-modal self-supervision. Asano, Patrick, Rupprecht, Vedaldi. NeurIPS, 2020.

Transformation-invariant clustering

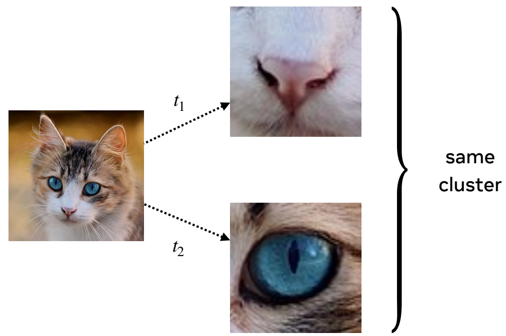
Replace predictor by $p(c \mid t(x))$ where t is a **random transformation** (aka **augmentation**)



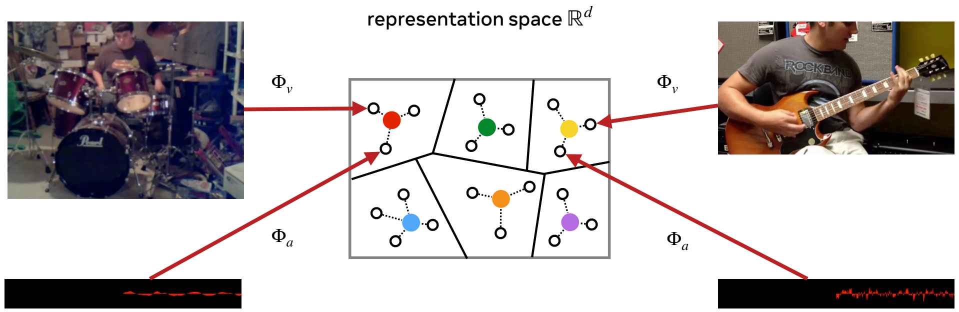
Strong augmentations

Strong augmentations remove "uninteresting correlations" between samples (e.g., partial overlap, matching colour, continuity)

Induce clustering based on **more abstract latent factors**, such as class identity

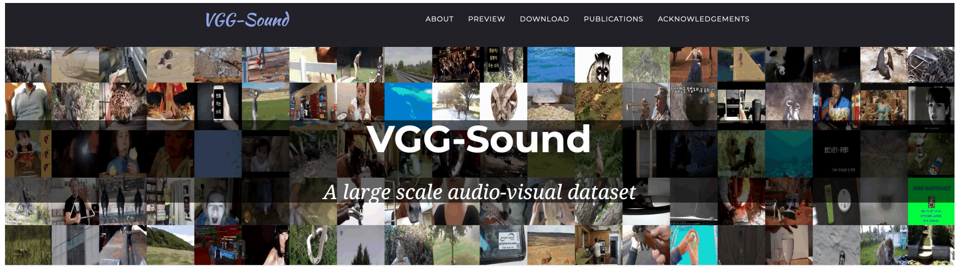


Cross-modal clustering: SeLaVi



Labelling unlabelled videos from scratch with multi-modal self-supervision. Asano, Patrick, Rupprecht, Vedaldi. NeurIPS, 2020

Example: VGG-Sound



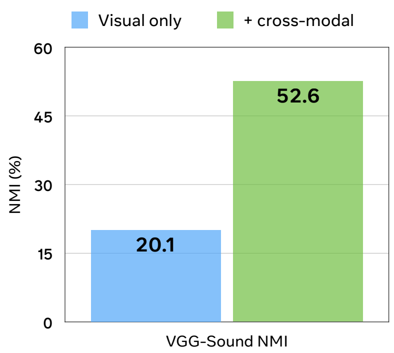
A large dataset of manually-annotated audio-visual sound classes

VGG-Sound: A large-scale audio-visual dataset. Chen, Xie, Vedaldi, Zisserman. ICASS, 2020.

Classes discovered in VGG-Sound

Video classes discovered using audio-visual clustering vs clustering using only the visual modality

Multi-modal clustering more than doubles the correspondence between human labels and automated clusters



Demo

https://www.robots.ox.ac.uk/~vgg/research/selavi/#demo

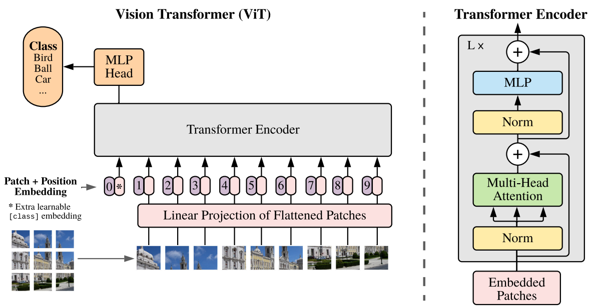
OXFORD

Scaling to large data and models

By avoiding the annotation cost, unsupervised learning allow to scale to much larder training datasets

High-capacity models can take advantage of larger datasets

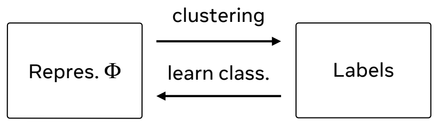
For instance, Vision Transformers (ViT)



An image is worth 16×16 words: Transformers for image recognition at scale. Dosovitskiy et al. Proc. ICLR, 2021.

Generating training targets on the fly

SeLa alternates learning the representation and recomputing the labels (clusters)

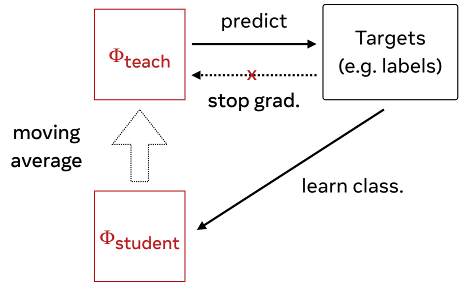


Unsupervised learning of visual features by contrasting cluster assignments. Caron, Misra, Mairal, Goyal, Bojanowski, Joulin. Proc. NeurIPS, 2020.

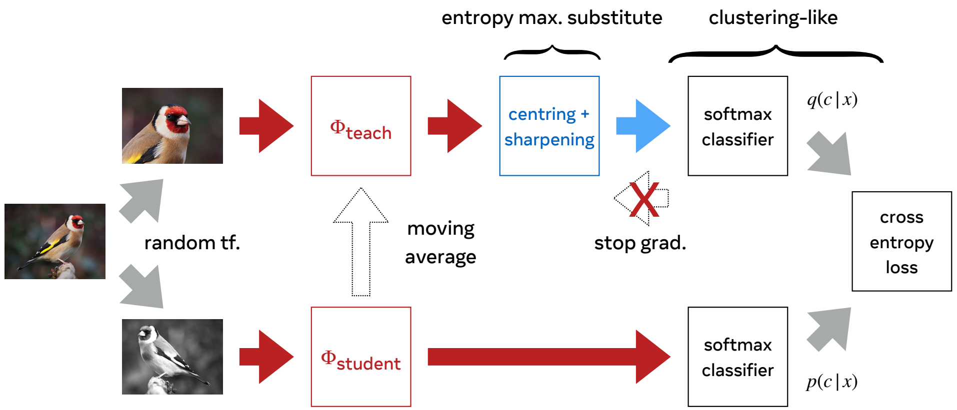
Bootstrap your own latent: A new approach to self-supervised learning. Grill et al., Proc. NeurIPS, 2020.

Momentum contrast for unsupervised visual representation learning. He, Fan, Wu, Xie, Girshick. Proc. CVPR, 2020

A **mean-teacher** allows to update the (self) labels or other training targets online



DINO: Self-distillation with no labels



Emerging properties in self-supervised vision transformers. Caron, Touvron, Misra, Jégou, Mairal, Bojanowski, Joulin. Proc. ICCV, 2021.

Noise contrastive learning

[InfoMax] **Self-organization in a perceptual network.** Linsker. Computer, 21(3), 1988.

[InstanceDiscr] **Unsupervised feature learning via non-parametric instance discrimination.** Wu, Xiong, Yu, Lin. Proc. CVPR, 2018

[DeepInfoMax] **Learning deep representations by mutual information estimation and maximization.** Hjelm, Fedorov, Lavoie-Marchildon, Grewal, Bachman, Trischler, Bengio. Proc. ICLR, 2019

[CPC] **Representation learning with contrastive predictive coding.** Oord, Li, Vinyals. Proc. NeurIPS, 2019.

[CMC] **Contrastive multiview coding.** Tian, Krishnan, Isola. Proc. ECCV, 2020.

[SimCLR] **A simple framework for contrastive learning of visual representations.** Chen, Kornblith, Norouzi, Hinton. Proc. ICML, 2020

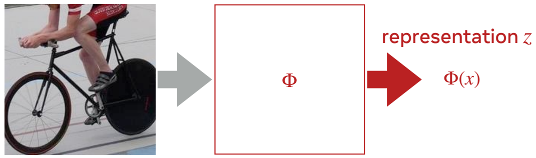
[MoCo] **Momentum contrast for unsupervised visual representation learning.** He, Fan, Wu, Xie, Girshick. Proc. CVPR, 2020

[SwAV] **Unsupervised learning of visual features by contrasting cluster assignments.** Caron, Misra, Mairal, Goyal, Bojanowski, Joulin. Proc. NeurIPS, 2020.

[BYOL] **Bootstrap your own latent: A new approach to self-supervised learning**. Grill, Strub, Altché, Tallec, Richemond, Buchatskaya, Doersch, Pires, Guo, Azar, Piot, Kavukcuoglu, Munos, Valko. Proc. NeurIPS, 2020.

[Review] **On mutual information maximization for representation learning.** Tschannen, Djolonga, Rubenstein, Gelly, Lucic. Proc. ICLR, 2020

The **InfoNCE estimator**



Information bound

$$I(x,z) \geq \max_f \mathop{\mathbb{E}}_{\text{batch}} \left[\frac{1}{K} \sum_{i=1}^{K} \log \frac{\exp f(x_i, z_i)}{\sum_{j=1}^{K} \exp f(x_i, z_j)} \right]$$

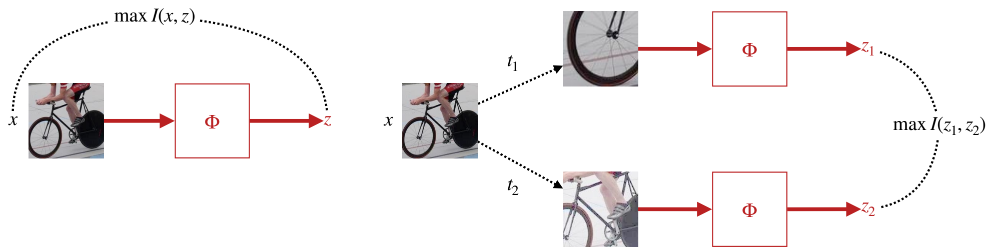
The estimator works by drawing and **contrasting** K pairs (x_i, z_i) from the joint distribution $p(x, z)$

Contrasting means learning a **critic** $f : (x, z) \mapsto \mathbb{R}$ that tells if x and z "go together" or not, similar to $p(x, z)$

Multi-view InfoNCE

It is rare to see InfoNCE used to implement standard InfoMax

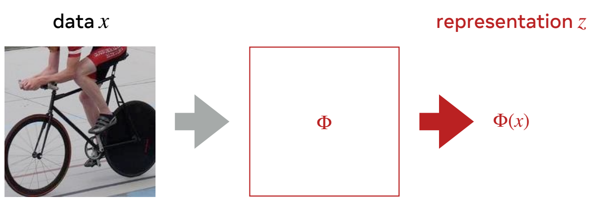
Instead, one almost always look at **multi-view learning**



Limitations of the information perspective

A representation's primary goal must be to simplify data analysis tasks

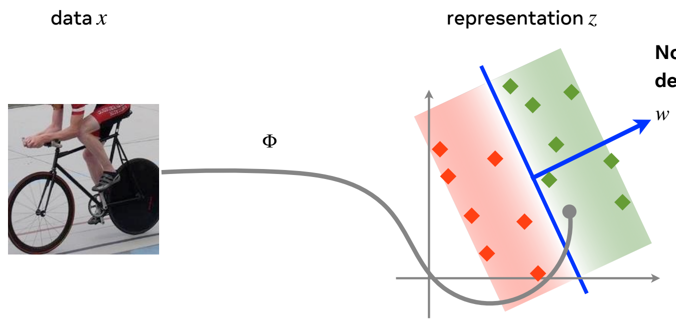
Information poorly quantifies this aspect



Recall that information is maximised by doing nothing!

$$I(x, z) \leq I(x, x)$$

Vector representations



interpretation y

bicycle?

$\langle w, \Phi(x) \rangle$

Vector representations

A vector representation maps semantic similarity to vector similarity

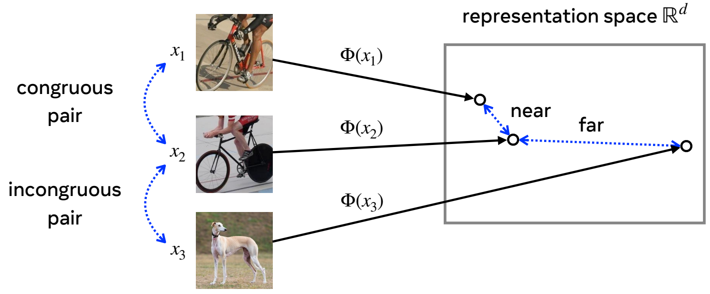
representation

Semantic similarity → Vector similarity (distance)

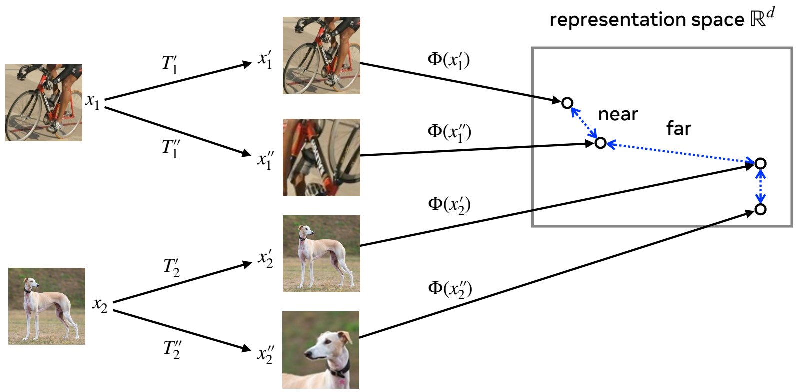
Invariant and distinctive representations

Representation vectors should be:

- **invariant** to nuisance factors
- **distinctive** for semantic factors



Contrastive learning: invariance and distinctiveness



Simple decision rules in clustering and InfoNCE

Self-labelling

The clustering layer is a **linear classifier**

$$p(c\,|\,x) = \frac{\exp w_c^\top z}{\sum_{k\neq c} \exp w_k^\top z}$$

where the **representation** is $z = \Phi(t(x))$

Contrastive learning

The critic function is the **dot product**

$$\mathop{\mathbb{E}}_{\text{batch}} \left[\frac{1}{K} \sum_{i=1}^{K} \log \frac{e^{z_i^\top z_i'}}{\sum_{j\neq i} e^{z_i^\top z_j}} \right]$$

where the **representation** $z_i = \Phi(t_i(x_i))$

The learned representations must support "simple" (linear) processing

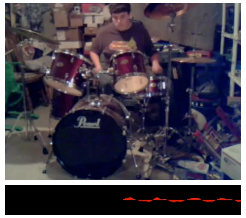
Note — for normalised features: $\|f(u) - f(v)\|^2 = 2 - 2f(u)^\top f(v)$

Should representations be invariant or distinctive?

Standard augmentations
- crop
- flip
- rotate
- …

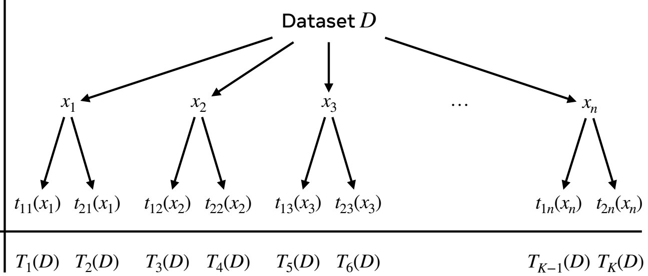
Time
- shift
- reversal

Modality slicing



Generalized Data Transformations (GDT)

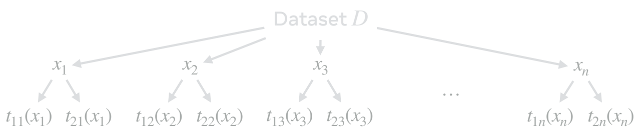
Transformation	Representation
sample an image	distinctive
standard augment	invariant
composition	



On compositions of transformations in contrastive self-supervised learning. Patrick, Asano, Kuznetsova, Fong, Henriques, Zweig, Vedaldi. ICCV, 2021.

Contrast matrix

The **contrast matrix** tells which transformation pairs should be invariant (+1), distinctive (-1) or ignored (0)



	$T_1(D)$	$T_2(D)$	$T_3(D)$	$T_4(D)$	$T_5(D)$	$T_6(D)$		$T_{K-1}(D)$	$T_K(D)$
$T_1(D)$	0	1	-1	-1	-1	-1		-1	-1
$T_2(D)$	1	0	-1	-1	-1	-1		-1	-1
$T_3(D)$	-1	-1	0	1	-1	-1		-1	-1
$T_4(D)$	-1	-1	1	0	-1	-1		-1	-1
$T_5(D)$	-1	-1	-1	-1	0	1		-1	-1
$T_6(D)$	-1	-1	-1	-1	1	0		-1	-1
…									
$T_{K-1}(D)$	-1	-1	-1	-1	-1	-1		0	1
$T_K(D)$	-1	-1	-1	-1	-1	-1		1	0

On compositions of transformations in contrastive self-supervised learning. Patrick, Asano, Kuznetsova, Fong, Henriques, Zweig, Vedaldi. ICCV, 2021.

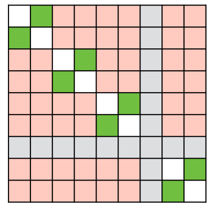
Contrastive loss using generalised transformations

Contrast positive pairs of GDTs against all the other pairs, except the ignored ones

$$\sum_{C(T,T')=1} \log \frac{e^{f(T(D),T'(D))}}{\sum_{C(T,T'')\neq 0} e^{f(T(D),T''(D))}}$$

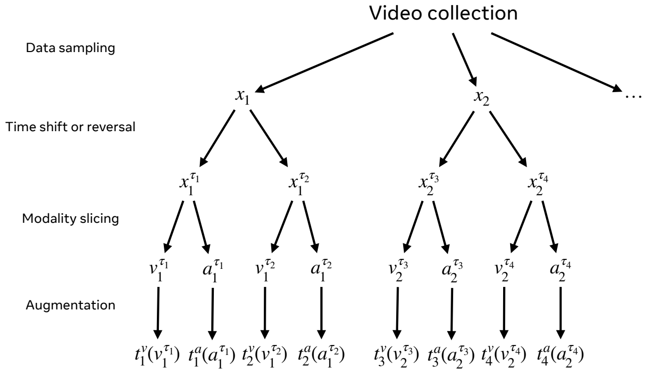
Overall contrast value can be deduced from the desired effects on individual transformations

Contrast matrix

$C =$ 

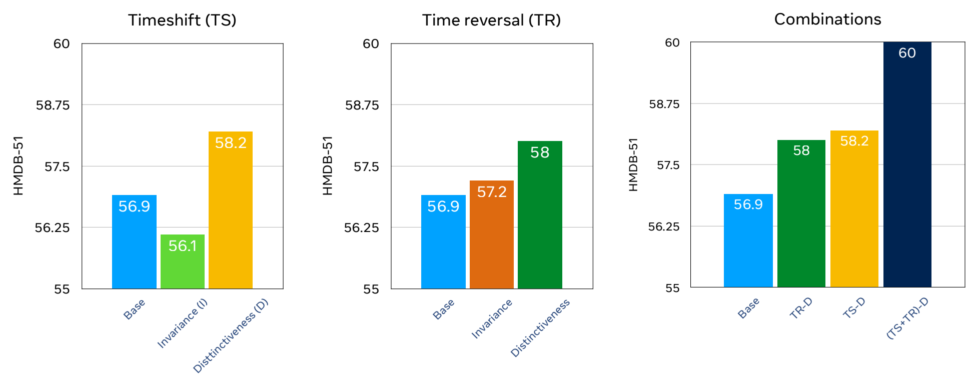
Multimodal GDTs

GDTs can be used to combine several multi-modal transformations in a single learning formulation



45

GDT example results

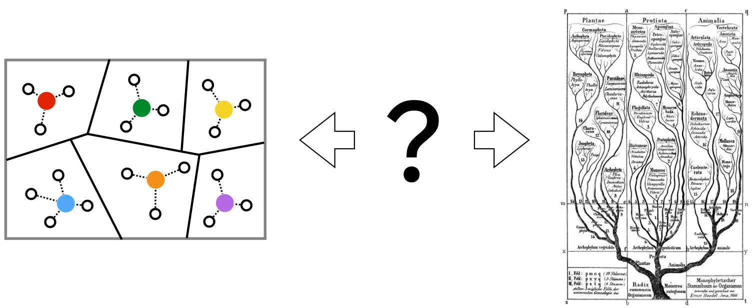


On compositions of transformations in contrastive self-supervised learning. Patrick, Asano, Kuznetsova, Fong, Henriques, Zweig, Vedaldi. ICCV, 2021.

46

Measuring representation interpretability

Is there a simple relation between a representation and concepts?



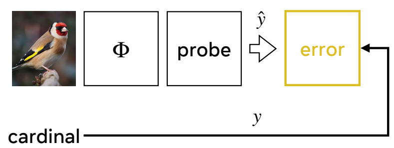
47

Direct vs reverse probing

Direct (standard) probing

Pick a dataset labelled for certain concepts (e.g. 1000 classes in ImageNet)

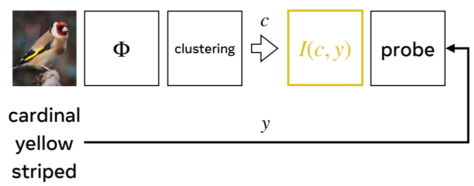
Try to **linearly** map representation vectors to concepts



Inverse probing

Cluster the representation vectors (via K-means)

Try to **linearly** map combinations of concepts to clusters



Network dissection: Quantifying interpretability of deep visual representations. Bau, Zhou, Khosla, Oliva, Torralba. Proc. CVPR, 2017

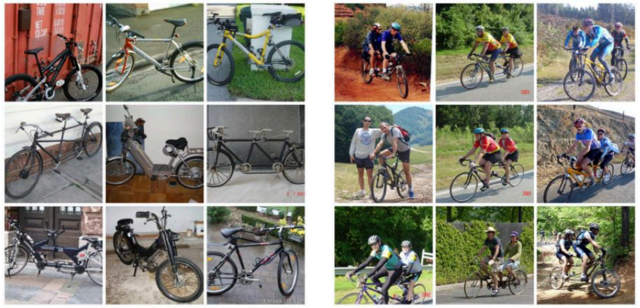
Measuring the interpretability of unsupervised representations via quantized reversed probing. Laina, Asano, Vedaldi. Proc. ICLR, 2021.

48

MoCo V2 clusters explained

ImageNet Bikes

Additional explanatory factors



MoCo(v2)

Measuring the interpretability of unsupervised representations via quantized reversed probing. Laina, Asano, Vedaldi. Proc. ICLR, 2021.

SimCLR clusters explained

ImageNet “Hockey”

Additional explanatory factors

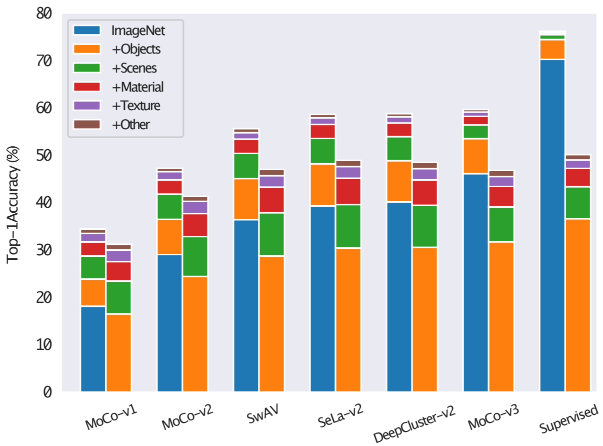
SimCLR

Measuring the interpretability of unsupervised representations via quantized reversed probing. Laina, Asano, Vedaldi. Proc. ICLR, 2021.

Interpretability of representations quantified

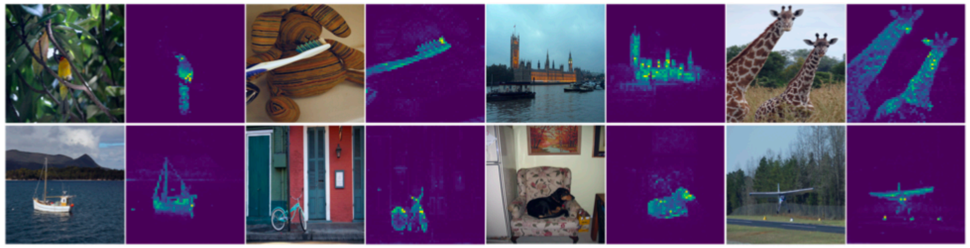
Explainability of unsupervised representations by concept families

Clustering-based methods are generally more interpretable



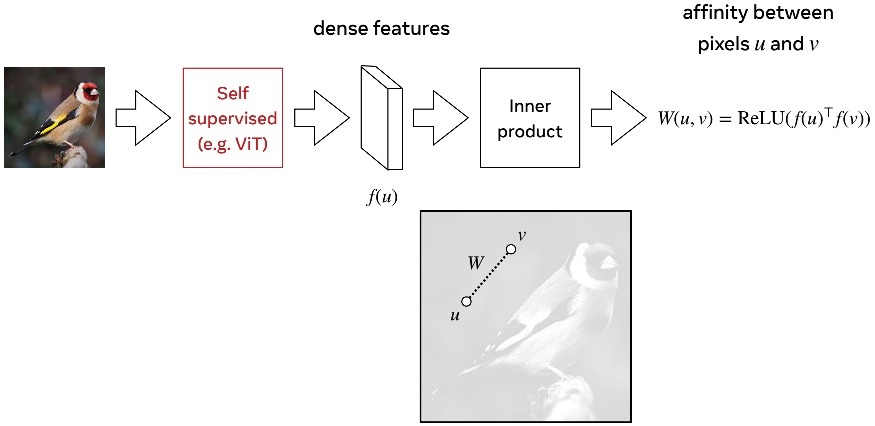
Measuring the interpretability of unsupervised representations via quantized reversed probing. Laina, Asano, Vedaldi. Proc. ICLR, 2021.

Emergent properties of self-supervised ViTs



Emerging properties in self-supervised vision transformers. Caron, Touvron, Misra, Jégou, Mairal, Bojanowski, Joulin. Proc. ICCV, 2021.

Forming spatial clusters of dense features



Extracting segments from affinities

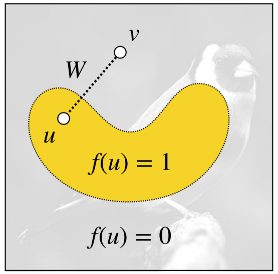
Let $f(u) \in [0,1]$ be the **indicator function** of a segment

The **segment's smoothens** according to the affinity W is

$$E(f) = \sum_{uv} W(u,v) \cdot (f(u) - f(v))^2$$

Rewritten in matrix form:

$E(f) = f^\top L f$ where $L = D - W$ (Laplacian matrix)



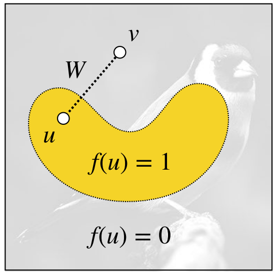
Extracting segments from affinities

The **eigenvectors** h_k of L form an **orthonormal basis** for f

This means that:

$$f(u) = a_0 h_0(u) + a_1 h_1(u) + \ldots + a_{n-1} h_{n-1}(u)$$

for some coefficients a_k



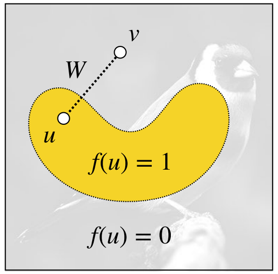
Extracting segments from affinities

By expanding f using the eigenvectors:

$$f(u) = a_0 h_0(u) + a_1 h_1(u) + \ldots + a_{n-1} h_{n-1}(u)$$

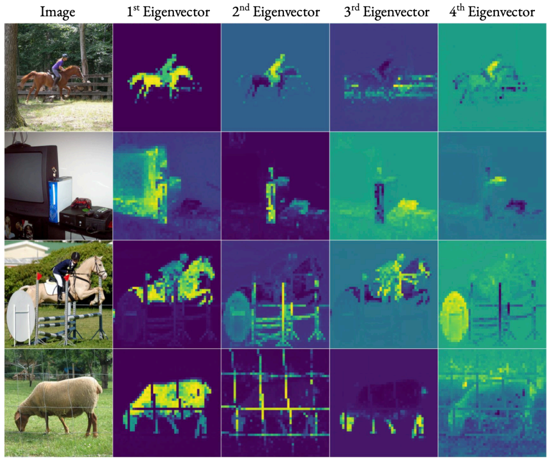
The segment's smoothness is given by the eigenvalues:

$$E(f) = \lambda_0 a_0^2 + \lambda_1 a_1^2 + \ldots + \lambda_{n-1} a_{n-1}^2$$



Eigenvectors of self-supervised affinities

Eigenvectors for DINO ViT

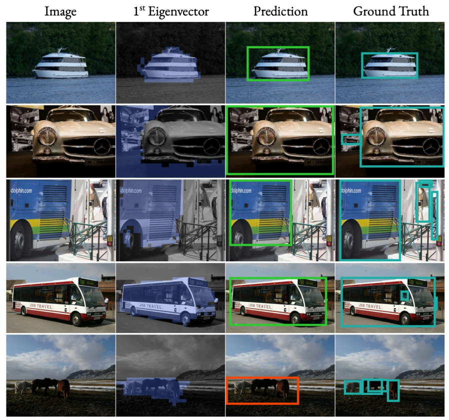


Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. Melas-Kyriazi, Rupprecht, Laina, Vedaldi. CVPR, 2022.

Unsupervised object localisation

CorLoc

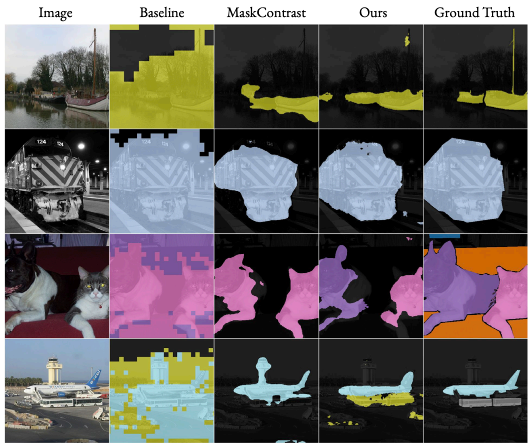
Method	VOC-07	VOC-12	COCO-20k
Selective Search [78]	18.8	20.9	16.0
EdgeBoxes [73]	31.1	31.6	28.8
Kim et al. [48]	43.9	46.4	35.1
Zhang et al. [94]	46.2	50.5	34.8
DDT+ [84]	50.2	53.1	38.2
rOSD [99]	54.5	55.3	48.5
LOD [79]	53.6	55.1	48.5
DINO- `[CLS]` [8]	45.8	46.2	42.1
LOST [67]	61.9	64.0	50.7
Ours	**62.7**	**66.4**	**52.2**



Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. Melas-Kyriazi, Rupprecht, Laina, Vedaldi. CVPR, 2022.

Unsupervised semantic segmentation

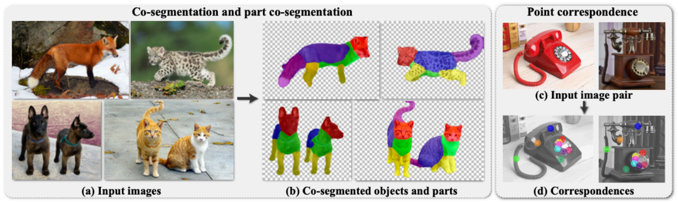
Method	mIoU
Pretext task methods	
Co-Occurrence [40]	4.0
CMP [92]	4.3
Colorization [95]	4.9
Clustering/Contrastive methods	
IIC [41]	9.8
MaskContrast† [74]	35.0
Additional baselines	
Cluster-Patch	5.3
Cluster-Seg	12.1
Saliency-DINO-ViT-B†	30.1
MaskContrast-DINO-ViT-B†	31.2
Ours w/o self-training	30.8 ± 2.7
Ours	**37.2** ± 3.8



Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. Melas-Kyriazi, Rupprecht, Laina, Vedaldi. CVPR, 2022.

More…

Unsupervised part discovery from contrastive reconstruction. Choudhury, Laina, Rupprecht, Vedaldi. NeurIPS, 2021.



Deep ViT features as dense visual descriptors. Amir, Gandelsman, Bagon, Dekel. CoRR, abs/2112.05814, 2021.

Conclusions for part I

Principles of self-supervised representations
- Informative representations (no collapse)
- Information vs metric view
- Transformation/modality invariance

Tricks of the trade
- Strong augmentations
- Distillation and mean-teacher
- High-capacity models (ViTs)

Measuring interpretability
- Direct and inverse probing
- Clustering parts

Applications
- data clusterings
- object/part segmentation
- many more…

Not covered but important: generative modelling
- Masked autoencoders (e.g., SiT, MAE)
- Inspired by ultra-large language models

SiT: Self-supervised vision transformer. Ahmed, Awais, Kittler. arXiv.cs, abs/2104.03602, 2021.

Masked autoencoders are scalable vision learners. He, Chen, Xie, Li, Dollár, Girshick. Proc. CVPR, volume abs/2111.06377, 2021.

AIMS Big Data Course
Introduction to deep learning

Part 2: Interpretation

Kind of explanations

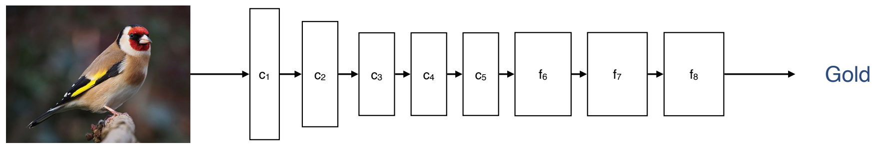
Analysis
Given an off-the-shelf networks, explain what it knowns, how it works, and how it learns

Win an argument
The network explains its decision to a user, with the goal of **convincing** her

Communicating a skill
Explain to a human or machine how to solve a certain class of problems, in general

facebook Artificial Intelligence — University of Oxford

Analysing deep neural networks



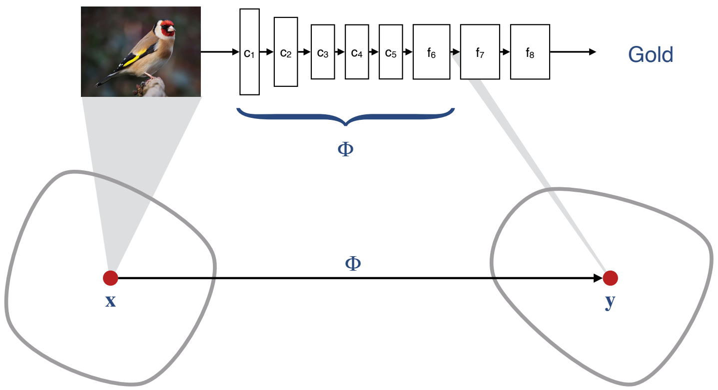
What does a net do?
- What concepts can it recognise?
- Spurious correlations?
- Limitations?

How does it do it?
- Template matching?
- Compositionality?
- Spatial reasoning?

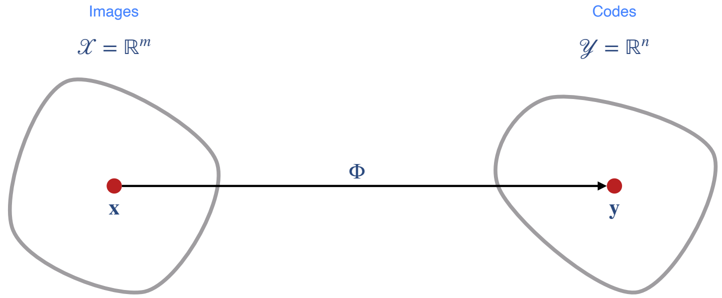
How does it learn it?
- Generalization?
- Optimisation?

facebook Artificial Intelligence — University of Oxford

Deep networks as encoders



Deep networks as encoders



Generating iconic examples

Attribution

Generating iconic examples

Attribution

How much information about $\mathbf{x}$ does $\mathbf{y}$ contain?

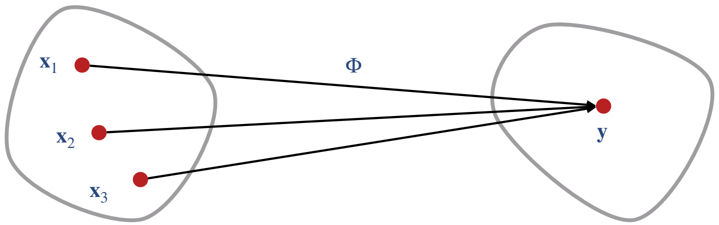
Multiple images map to the same code

Images

$\mathcal{X} = \mathbb{R}^m$

Codes

$\mathcal{Y} = \mathbb{R}^n$



Pre-image

Reconstructions form an **equivalence class** of images, called a pre-image

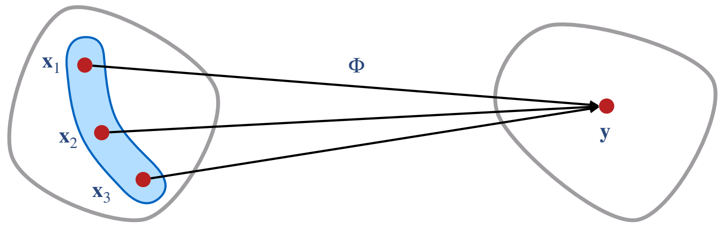
All pre-images hat are indistinguishable for the network

Images

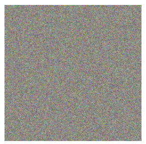
$\mathcal{X} = \mathbb{R}^m$

Codes

$\mathcal{Y} = \mathbb{R}^n$



Finding pre-images via optimisation



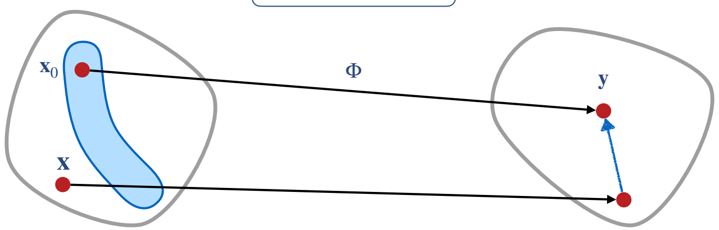
Images

$\mathcal{X} = \mathbb{R}^m$

$$\boxed{\min_{\mathbf{x}} \|\Phi(\mathbf{x}) - \Phi(\mathbf{x}_0)\|^2}$$

Codes

$\mathcal{Y} = \mathbb{R}^n$

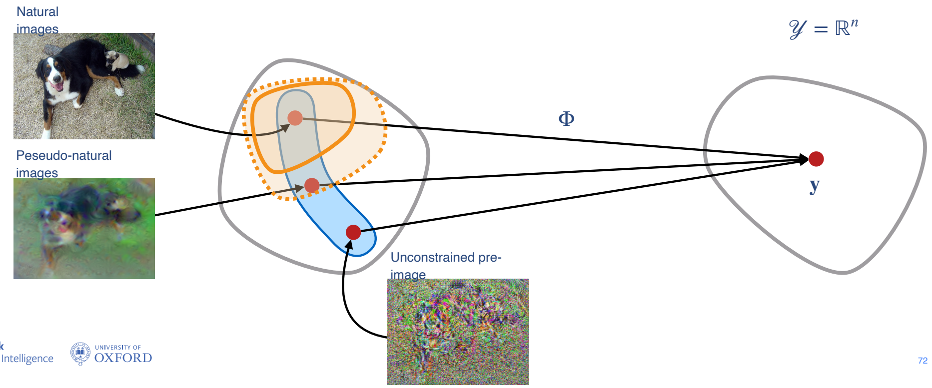


Natural pre-images

We are interested in pre-images that can realistically be network inputs

Codes

$\mathcal{Y} = \mathbb{R}^n$



Pseudo-natural pre-images

Regularised energy

$$\min_{\mathbf{x}} \|\Phi(\mathbf{x}) - \Phi(\mathbf{x}_0)\|^2 + \mathscr{R}(\mathbf{x})$$

For example TV-norm

Understanding deep image representations by inverting them
Mahendran Vedaldi, CVPR, 2015

Constrained optimisation

$$\min_{\mathbf{x} \in \mathscr{X}_{pn}} \|\Phi(\mathbf{x}) - \Phi(\mathbf{x}_0)\|^2$$

For example Deep Image Prior

Deep image prior
Ulyanov Vedaldi Lempistky, CVPR, 2018

Posterior probability

$$p(\mathbf{x} \mid \mathbf{y}) \sim \delta(\Phi(\mathbf{x}) - \mathbf{y}) \cdot p(\mathbf{x})$$

For example Plug & Play gen. nets

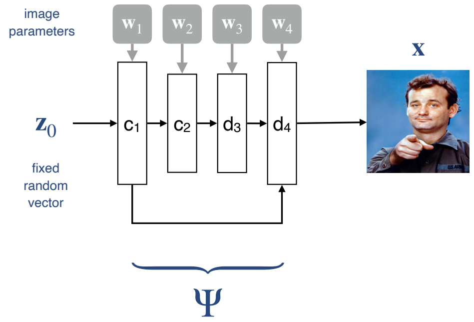
Plug & play generative networks: Conditional iterative generation of images in latent space
Nguyen, Yosinski, Bengio, Dosovitskiy, Clune, CVPR, 2017

Generator nets as image parameterisations

Consider a **generator network** Ψ with a fixed input $\mathbf{z}_0$

The network parameters $\mathbf{w}$ can be thought as **image parameters**

$$\mathbf{w} \longmapsto \mathbf{x} = \Psi(\mathbf{z}_0; \mathbf{w})$$



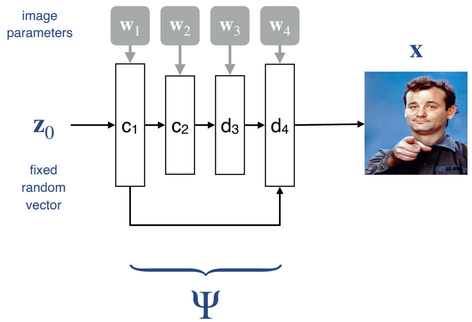
Fit a network to a single example

Start **randomly-initialised** network

Given an image $\mathbf{x}$, its parameter $\mathbf{w}$ is recovered by solving the optimisation problem

$$\min_{\mathbf{w}} \|\mathbf{x} - \Psi(\mathbf{z}_0; \mathbf{w})\|^2$$

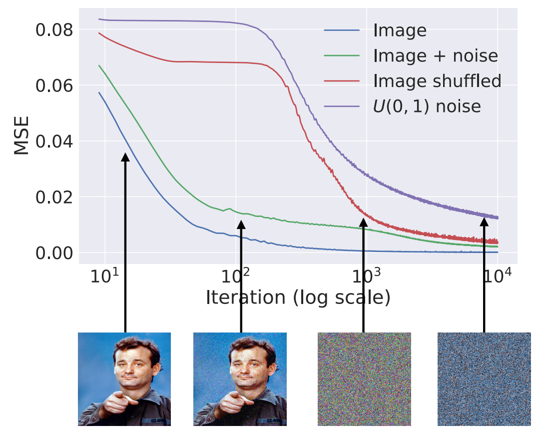
This is similar to learning the network from a single image



Deep image prior

For most generator networks fitting naturally-looking images is easier/faster than fitting others

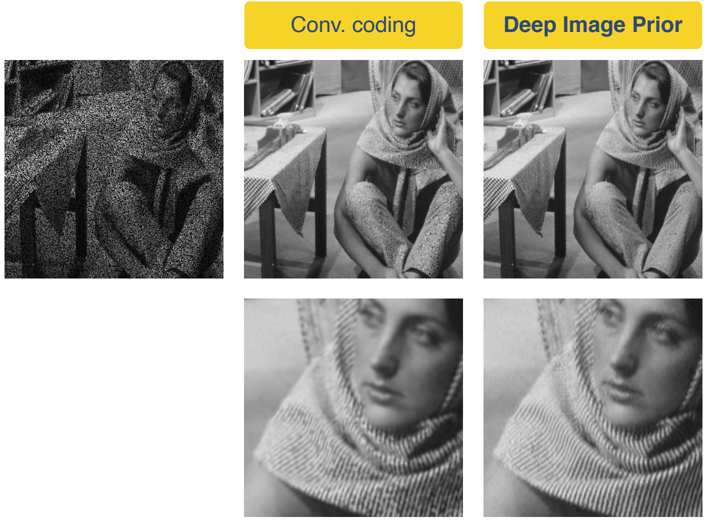
Deep image prior
Ulyanov Vedaldi Lempistky, CVPR, 2018



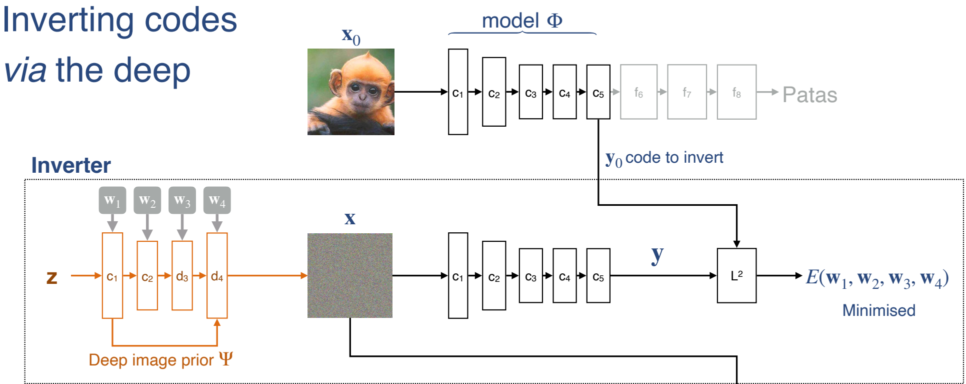
Deep image

For **inpainting** we only reconstruct the visible pixels, implicitly infer the others

$$\min_{\mathbf{w}} \|\mathbf{m} \odot (\mathbf{x} - \Phi(\mathbf{w}))\|^2$$

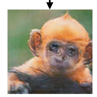


Inverting codes *via* the deep



The inverter is only given the **code**;
it is **not** learned from data in any way

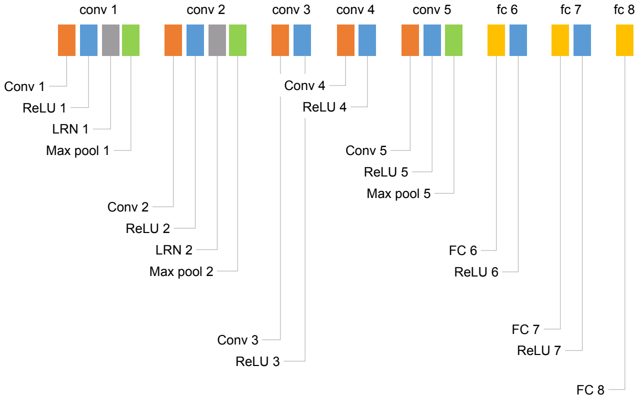
$$\min_{\mathbf{w}} \|\Phi(\Psi(\mathbf{w})) - \Phi(\mathbf{x}_0)\|^2$$



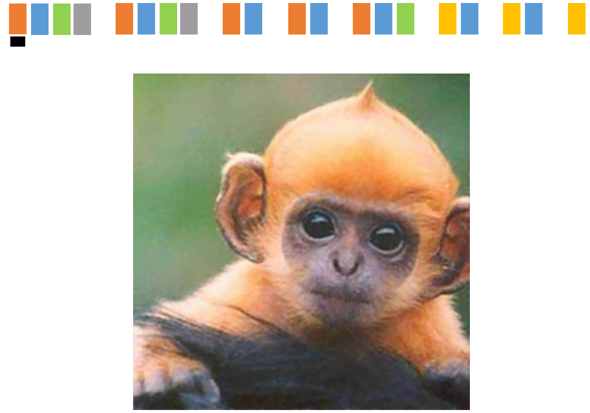
Inversion result

Inverting AlexNet

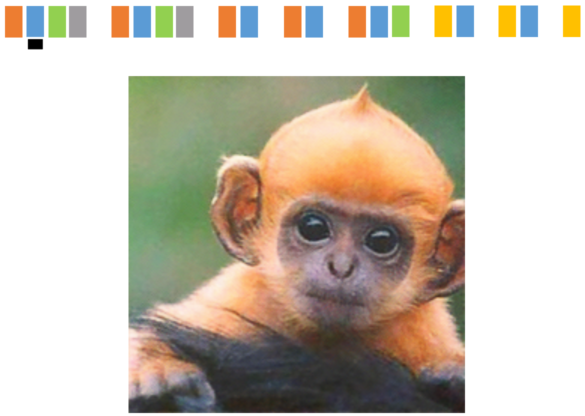
[Krizhevsky et al. 2012]



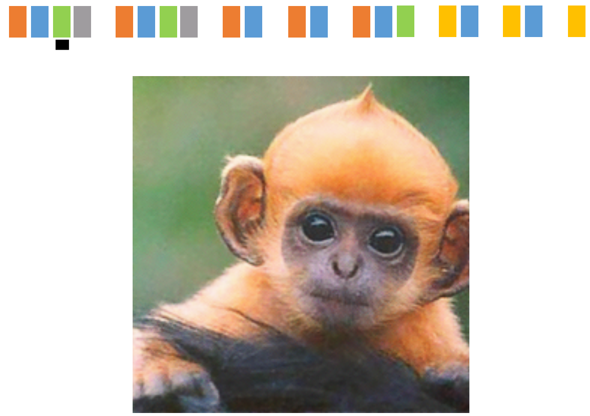
Inverting AlexNet



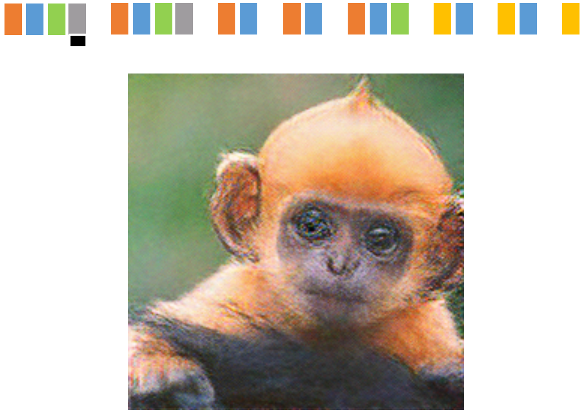
Inverting AlexNet



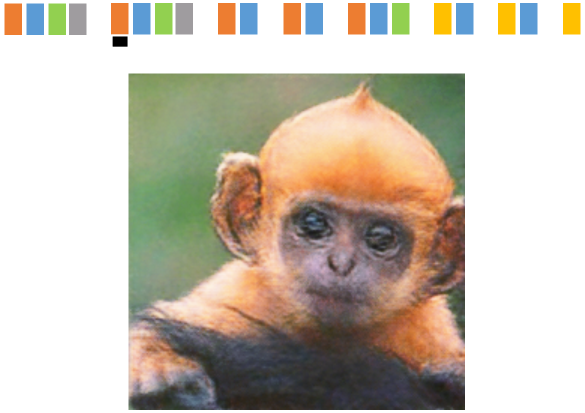
Inverting AlexNet



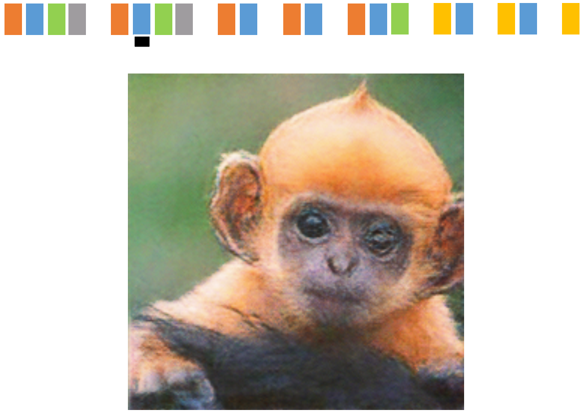
Inverting AlexNet



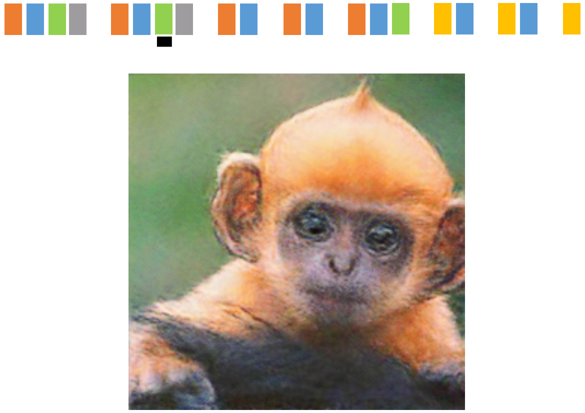
Inverting AlexNet



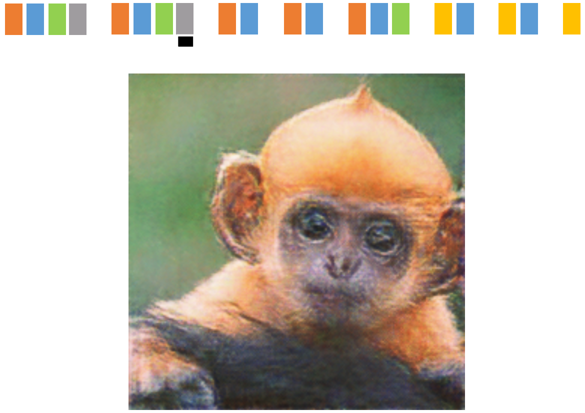
Inverting AlexNet



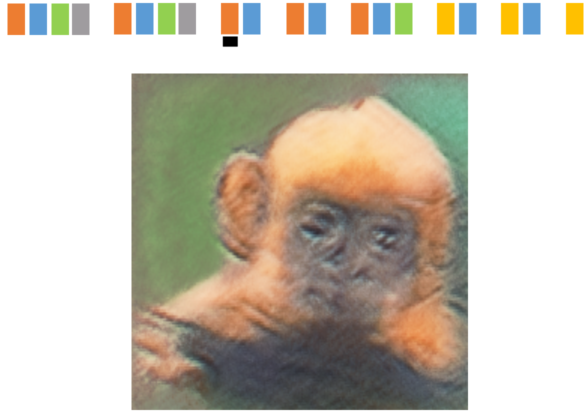
Inverting AlexNet



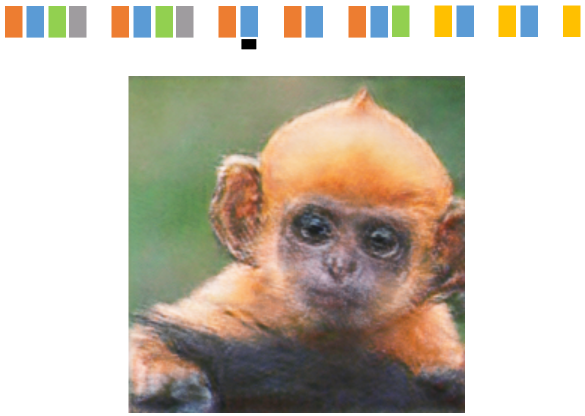
Inverting AlexNet



Inverting AlexNet



Inverting AlexNet



Inverting AlexNet

Inverting AlexNet

Inverting AlexNet

Inverting AlexNet

Inverting AlexNet

Inverting AlexNet

Artificial Intelligence

Inverting AlexNet

Artificial Intelligence

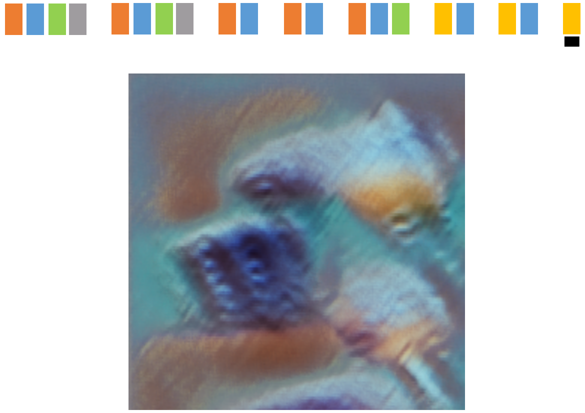
Inverting AlexNet

Artificial Intelligence

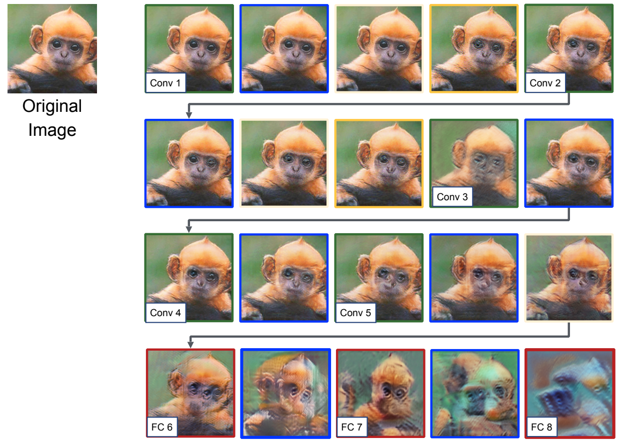
Inverting AlexNet

Artificial Intelligence

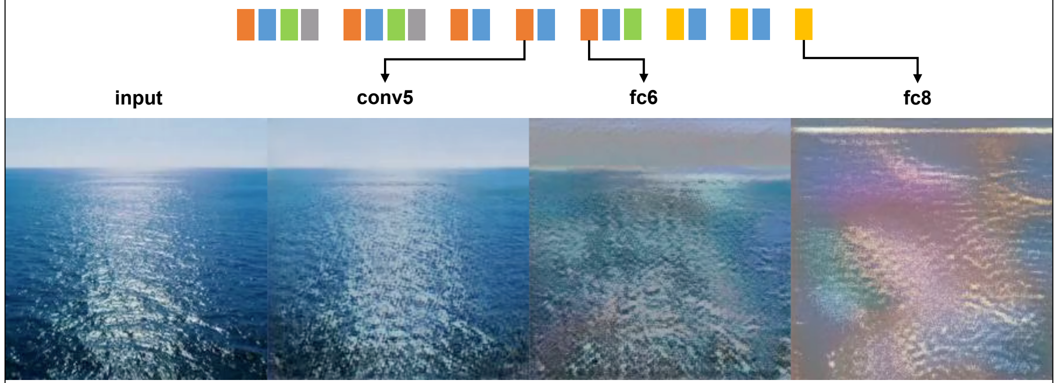
Inverting AlexNet



Inverting AlexNet



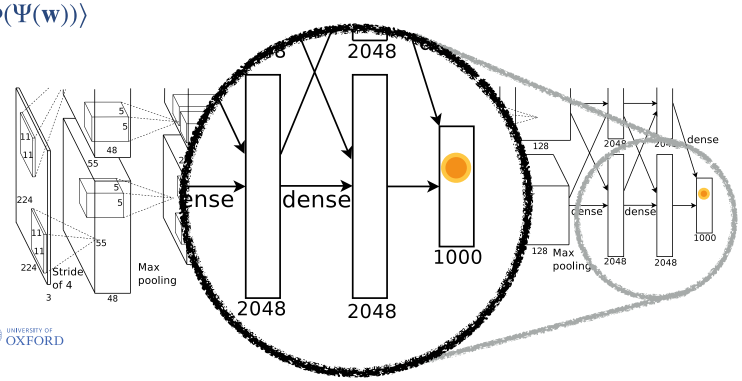
Is the code semantic or visual?



fc8 is a 1000-dimensional **class score vector**…

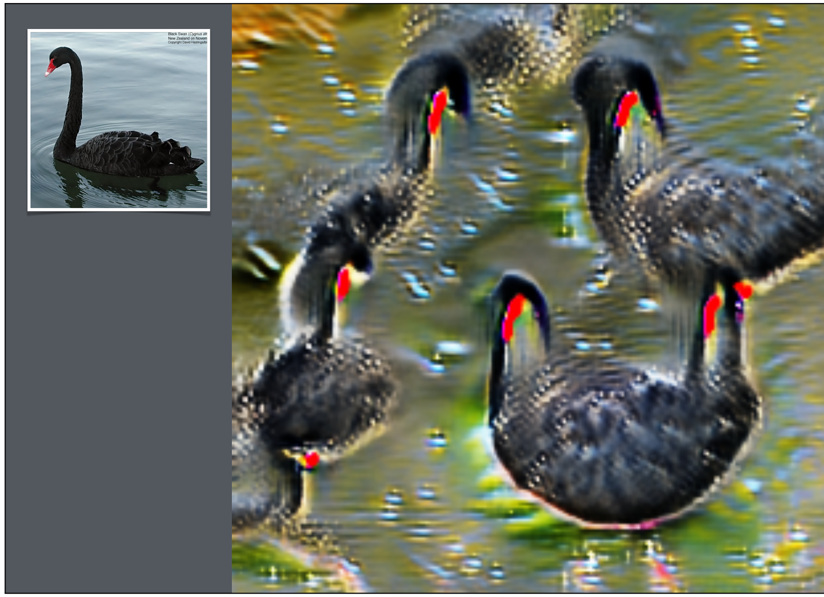
Activation maximization

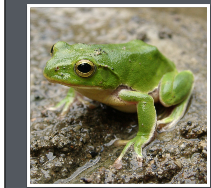
$$\min_{\mathbf{w}} -\langle \mathbf{e}_k, \Phi(\Psi(\mathbf{w}))\rangle$$

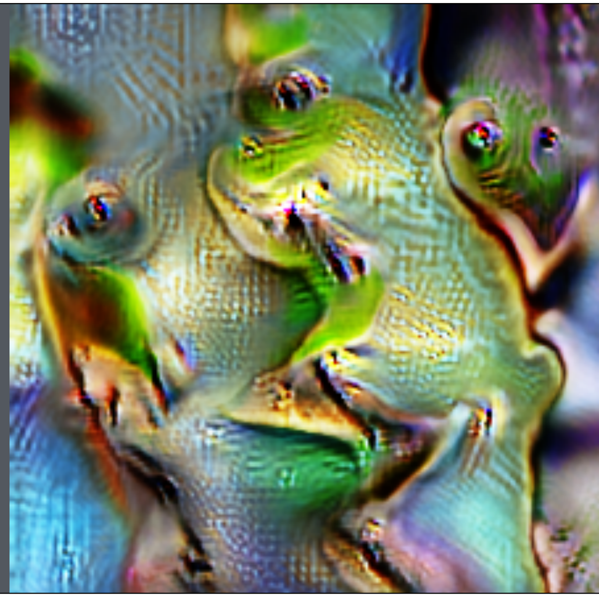


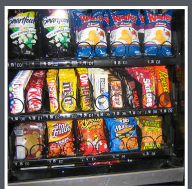
Deep Quiz

https://goo.gl/jURsCP









References

Visualizing higher-layer features of a deep network.
Erhan, Bengio, Courville, U Montreal, 2009

Activation maximisation for class neurons

Visualizing and understanding convolutional networks
Zeiler Fergus. Proc. ECCV, 2014.

Activation maximization using **empirical prior**, **deconvnet**

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps
Simonyan Zisserman Vedaldi, ICLR, 2104

Activation maximization and **saliency**

Understanding deep image representations by inverting them
Mahendran Vedaldi, CVPR, 2015

Inversion at different depths, **natural image prior**

Google "inceptionsm"
Mordvintsev et al. 2015

Activation maximisation for **intermediate neurons**
Improved regularizers, artistic applications (deep dreams)

Understanding neural networks through deep visualisation
Yosinksi et al. ICMLW, 2015

Activation maximization using **empirical prior**, **deconvnet**
More regularizers, toolbox

Plug & play generative networks: Conditional iterative generation of images in latent space
Nguyen, Yosinksi, Bengio, Dosovitskiy, Clune, CVPR, 2017

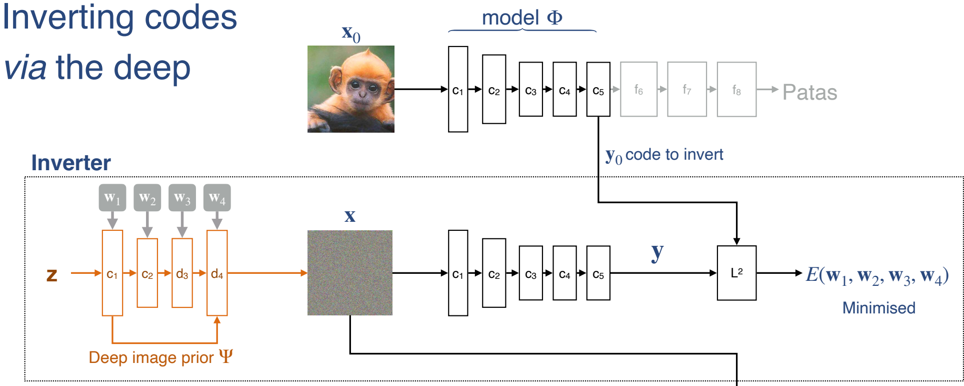
Strong learned regularizer, sample **diversity**

Deep image prior
Ulyanov Vedaldi Lempitsky, CVPR, 2018

Advanced "data agnostic" regularization

Effect of the prior

Inverting codes *via* the deep

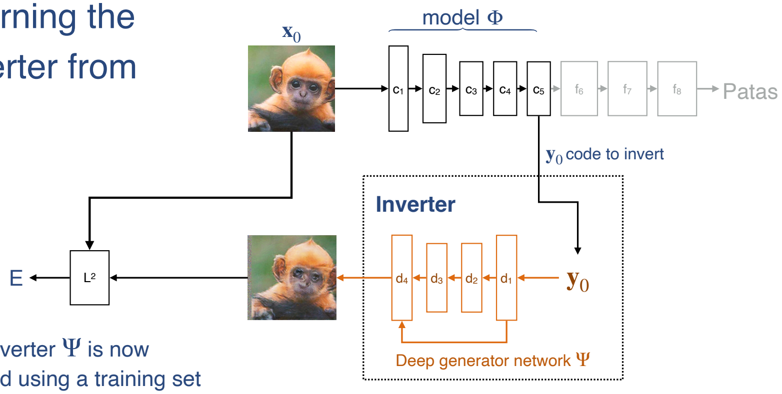


The inverter is only given the **code**;
it is **not** learned from data in any way

$$\min_{\mathbf{w}} \|\Phi(\Psi(\mathbf{w})) - \Phi(\mathbf{x}_0)\|^2$$

Inversion result

Learning the inverter from



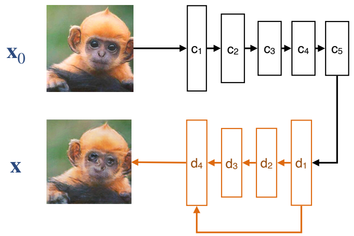
The inverter Ψ is now learned using a training set

$$\min_{\Psi} \frac{1}{N} \sum_{i=1}^{N} \|\Psi(\Phi(\mathbf{x}_i)) - \mathbf{x}_i\|^2 \; +$$ IMAGENET

Learning the inverter

Popular methods combine:

- perceptual loss $\mathbf{x}_0 \approx \mathbf{x}$
- feature rec. loss $\Phi(\mathbf{x}_0) \approx \Phi(\mathbf{x})$
- adversarial loss (GAN) $p(\mathbf{x}_0) \approx p(\mathbf{x})$
- IMAGENET



Inverting convolutional networks with convolutional networks
Dosovitskiy Brox, CVPR, 2016

Synthesizing the preferred inputs for neurons in neural networks via deep generator networks
Nguyen, Dosovitskiy, Yosinski, Brox, Clune, NIPS, 2016

Generating images with perceptual similarity metrics based on deep networks
Dosovitskiy Brox, NIPS, 2016

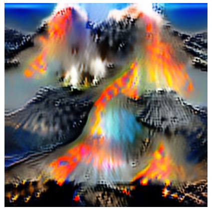
Plug & play generative networks: Conditional iterative generation of images in latent space
Nguyen, Yosinski, Bengio, Dosovitskiy, Clune, CVPR, 2017

Diagnostic *vs*

Our goal: diagnose a given **network** Φ

But inversions **also** reflect the chosen "natural image" **prior** $p(\mathbf{x})$

$p(\mathbf{x}) =$

Deep Image Prior	Plug & Play Gen. Net.	Empirical prior
		
only prior is the **structure** of the gener.	prior comes from training a **GAN** on **ImageNet**	**ImageNet empirical distribution**

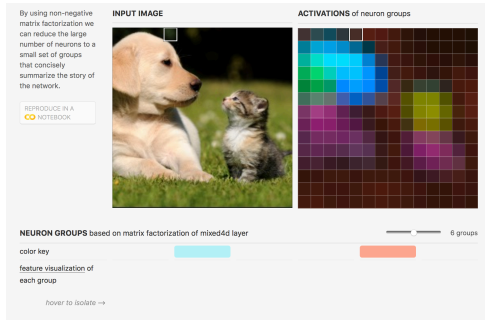
Illustrates the **model** Φ → Illustrates the **prior** $p(\mathbf{x})$

Reviews and interfaces

The building blocks of interpretability
Olah, Satyanarayan, Johnson, Carter, Schubert, Ye, Mordvintsev
Distill, 2018. https://distill.pub/2018/building-blocks

Understanding neural networks through deep visualisation
Yosinski et al. ICMLW, 2015

Definitely check out **Distill**!

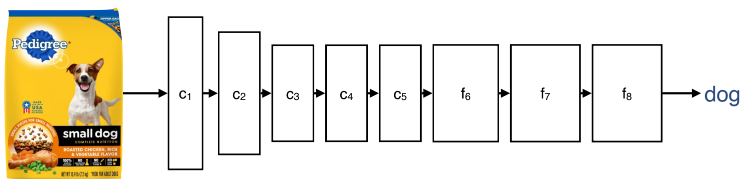


Generating iconic examples

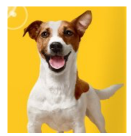
Attribution

Attribution

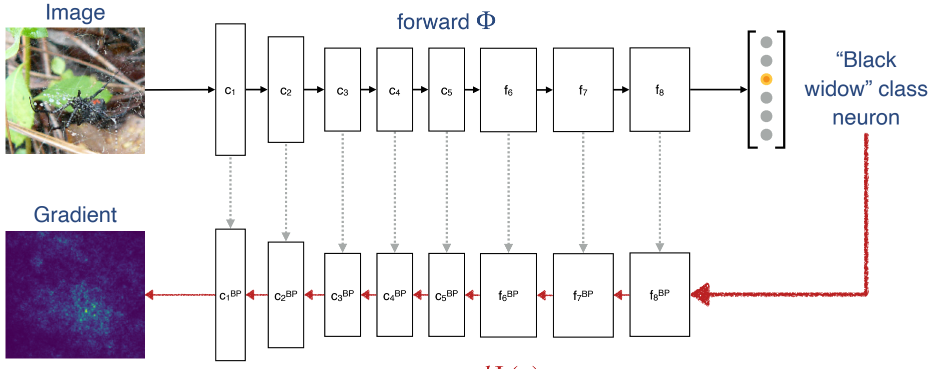
Where is the model **looking**?



?



Backprop methods: grad



backward $J = \dfrac{d\Phi(\mathbf{x})}{d\mathbf{x}}$ The "salient" pixels usually light up

Deep inside convolutional networks, Simonyan, Vedaldi, Zisserman, ICLR, 2014

Early backprop methods

Deconvolution

Visualizing and understanding convolutional networks
Zeiler Fergus, ECCV, 2014

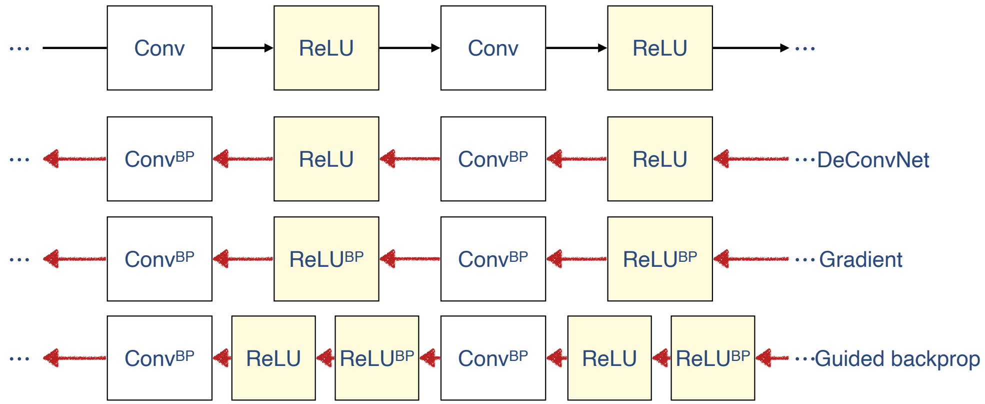
Gradient (backpropagation)

Deep inside convolutional networks: Visualising image classification models and saliency maps
Simonyan, Vedaldi, Zisserman, ICLR, 2014

Guided backpropagation

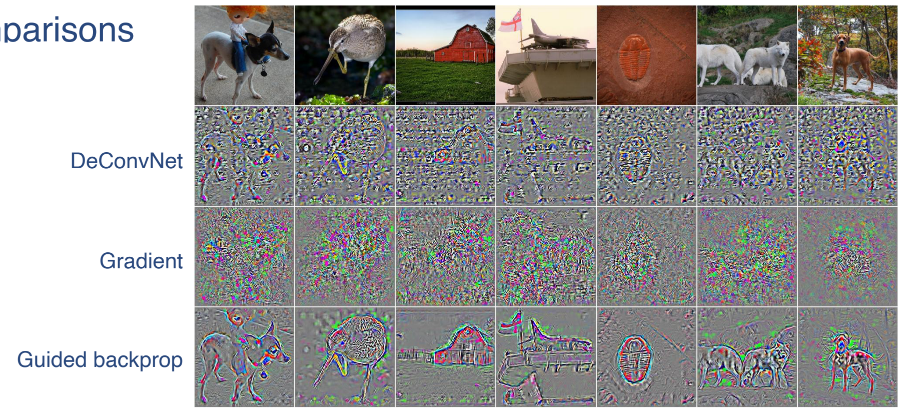
Striving for simplicity: The all convolutional net
Springenberg, Dosovitskiy, Brox, Riedmiller, ICLR, 2015

Backprop: deconv, grad, guided grad



Salient deconvolutional networks, Mahendran Vedaldi, ECCV, 2016

Comparisons



Salient deconvolutional networks. Mahendran Vedaldi, ECCV, 2016

Comparisons

Deconvolution

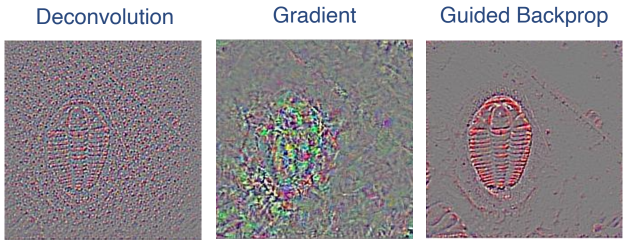
- Sharp
- Poor spatial selectivity

Gradient

- Blurry
- OK spatial selectivity

Guided Backprop

- Sharp
- OK spatial sensitivity



Warning: they all still have poor channel selectivity

Smoother grads

Gradient $\dfrac{d\Phi(x)}{d\mathbf{x}}$

Gradient × input $\mathbf{x} \odot \dfrac{d\Phi(x)}{d\mathbf{x}}$

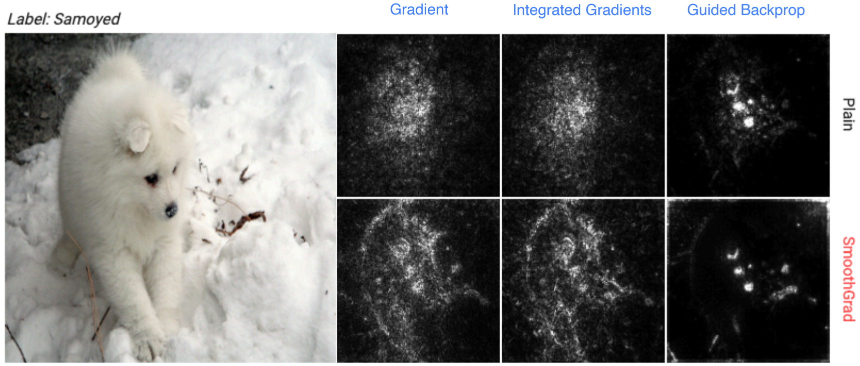
Integrated Gradients $(\mathbf{x} - \bar{\mathbf{x}}) \otimes \displaystyle\int_0^1 \frac{d\Phi(\bar{\mathbf{x}} - \alpha(\mathbf{x} - \bar{\mathbf{x}}))}{d\mathbf{x}}\, d\alpha$

Axiomatic attribution for deep networks. Sundararajan, Taly, Yan. Proc. ICML, 2017.

SmoothGrads $E\left[\dfrac{d\Phi(\mathbf{x}+\epsilon)}{d\mathbf{x}}\right], \quad \epsilon \sim \mathcal{N}$

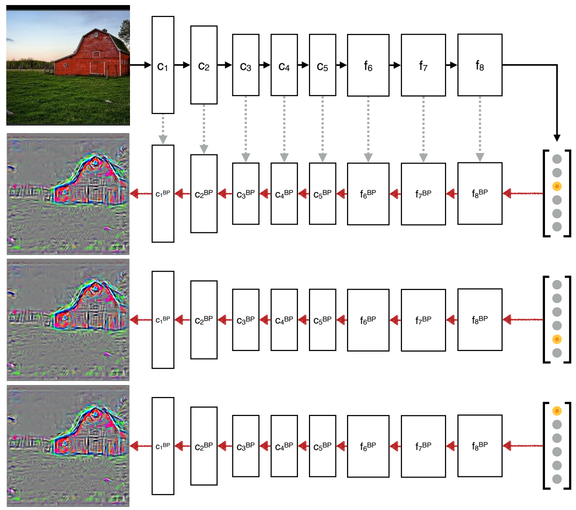
Smoothgrad: removing noise by adding noise. Smilkov, Thorat, Viegas, Wattenberg.

Comparisons



Lack of channel

Visualising any output results in about the same result



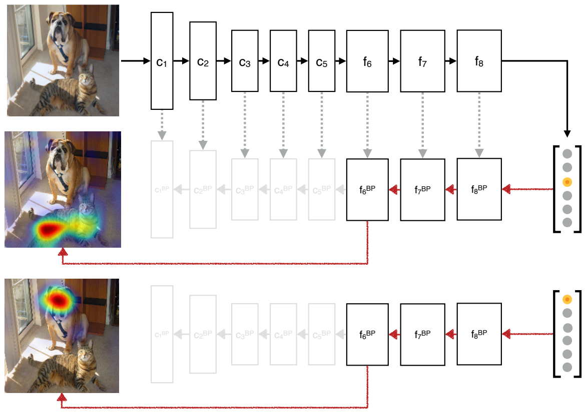
Attribution for:

maximally activated neuron

random neuron

minimally activated neuron

Backprop: CAM and



Attribution for:

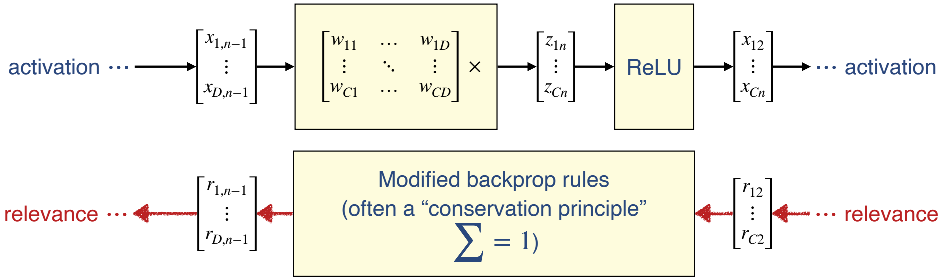
cat class neuron

dog class neuron

Learning deep features for discriminative localization
Zhou, Khosla, Lapedriza, Oliva, Torralba, CVPR, 2016

Grad-CAM: Visual explanations from deep networks via gradient-based localization
Selvaraju, Cogswell, Das, Vedantam, Parikh, Batra, ICCV, 2017

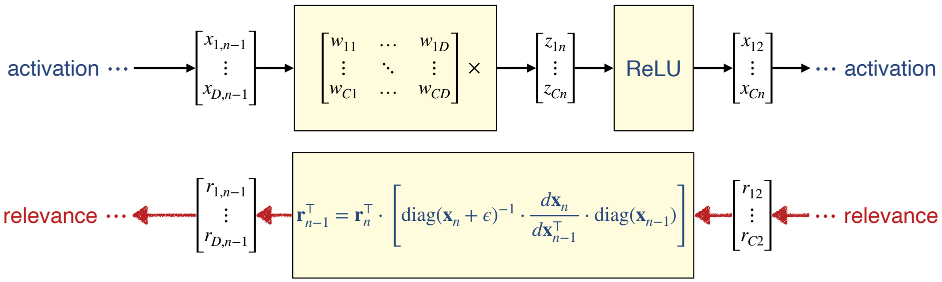
Relevance and excitation backprop



On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation
Bach, Binder, Montavon, Klauschen, Müller. PLOS one, 2015

Top-down neural attention by excitation backprop
Zhang, Lin, Brandt, Shen, Sclaroff, ECCV, 2016

Relevance and excitation backprop

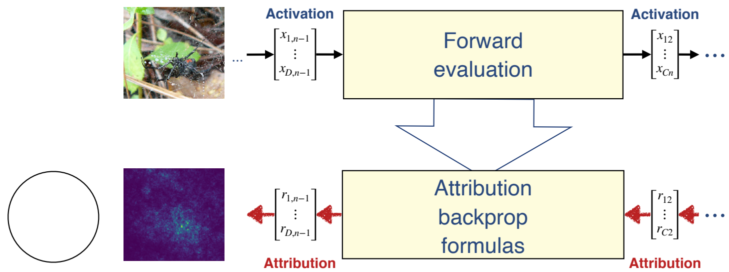


Actual rules are more sophisticated, please see references!

The meaning of attribution maps

For most methods, attribution is defined algorithmically

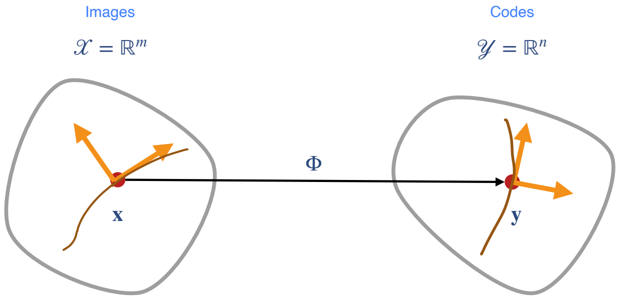
Hence, the **meaning** of the output is **not so clear**



Grad method = sensitivity analysis

The **gradient** can be <u>directly interpreted</u> as a **local linear approximation** of the model

$$\Phi(\mathbf{x}) \approx \left\langle \frac{d\Phi}{d\mathbf{x}}, \mathbf{x} - \mathbf{x}_0 \right\rangle + \Phi(\mathbf{x}_0)$$



Perturbation analysis

Study how $\Phi(\mathbf{x})$ changes up to perturbations $\pi(\mathbf{x})$ of the input $\mathbf{x}$

Perturbations should be meaningful (interpretable). E.g:

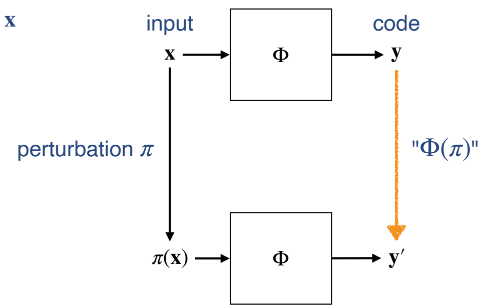
- Injecting noise
- Rotating or translating the image
- Erasing parts of the image

The representation may

- Be invariant (stay the same)
- Be equivariant (respond predictably)

The analysis may be

- Local around $\mathbf{x}$ and π
- For a distribution $p(\mathbf{x})$ and a fixed $p(\pi)$
- For a distribution $p(\pi)$ and a fixed $\mathbf{x}$



Perturbation analysis

Change the input and observe the effect on the output

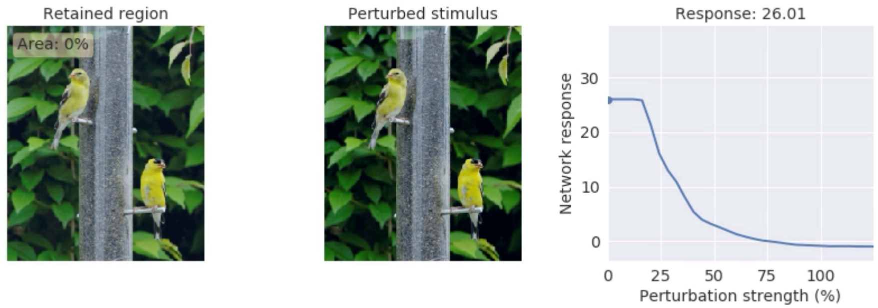
Input Occlusion RISE

Clear meaning, but can only test a small number of occlusion patterns

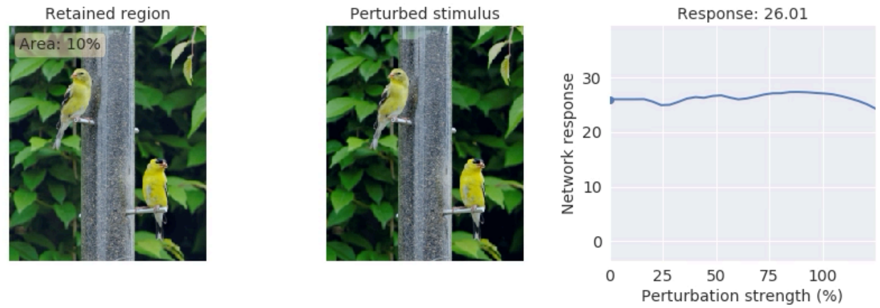
[Zeiler and Fergus, ECCV 2014; Petsiuk et al., BMVC 2018]

Extremal Perturbations

Blur everywhere ⇒ response suppressed



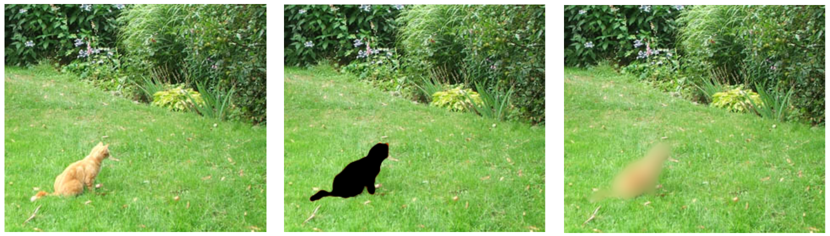
Preserve 10% ⇒ response preserved



Meaningful perturbations

We seek the "smallest elision" that maximally changes the neuron activation

Original	**Redact-out**	**Blur-out**



"cat" probability 1.00	"cat" probability 0.5 (ineffective)	"cat" probability 0.01 (more meaningful)

Adversarial perturbations

Neural networks are fragile to adversarial perturbations

Adversarial perturbations attract gradient descent

Intriguing properties of neural networks. Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus. CoRR 2013

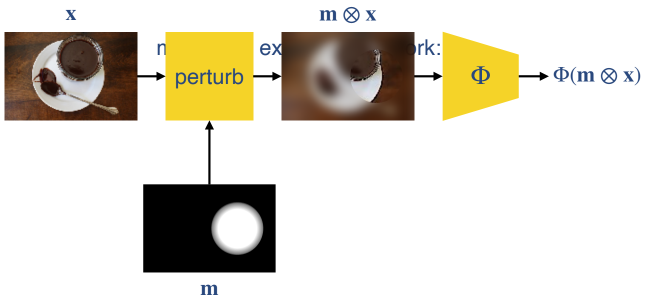
Original | Redacted | Mask

Extremal perturbations

A mask is optimized to

$$\operatorname*{argmax}_{\mathbf{m}} \Phi(\mathbf{m} \otimes \mathbf{x})$$

subject to $\text{area}(\mathbf{m}) = a$

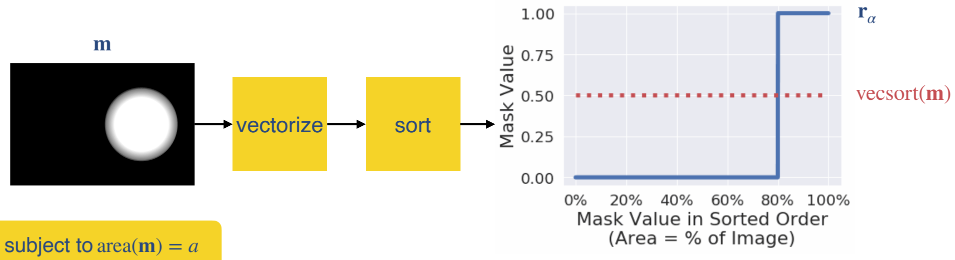


Area constraint

Optimizing w.r.t. to an area constraint is challenging

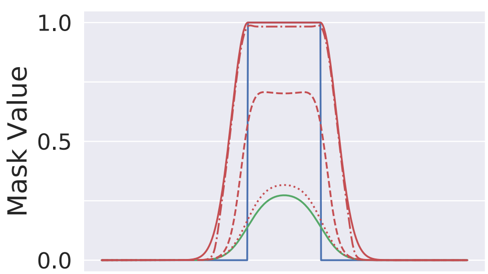
Here we re-formulate it as matching a **rank statistics**

$$\mathcal{L}_{area} = \|\text{vecsort}(\mathbf{m}) - \mathbf{r}_a\|^2$$



subject to $\text{area}(\mathbf{m}) = a$

Smooth masks



$m(v)$: mask

$$\text{conv}(u; m; k) = \frac{1}{Z} \sum_{v \in \Omega} k(u - v) m(v)$$

$$\text{maxconv}(u; m; k) = \max_{v \in \Omega} k(u - v) m(v)$$

$$\text{smoothconv}(u; m; k; T) = \text{smax}_{v \in \Omega; T}\, k(u - v) m(v)$$

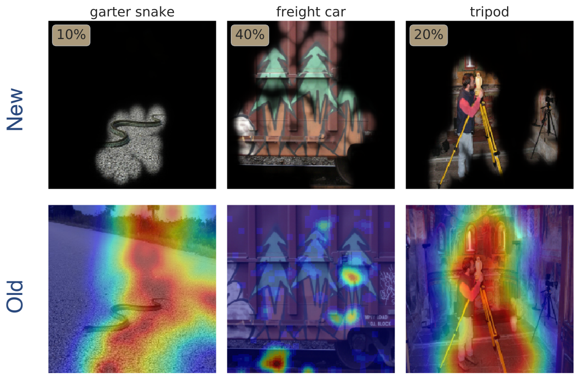
$$\text{smax}_{u \in \Omega; T} f(u) = \frac{\sum_u f(u) \exp(f(u)/T)}{\sum_u \exp(f(u)/T)}$$

Smooth masks

Comparison with prior work on “meaningful perturbations"

Compared to **Fong and Vedaldi, 2017**, we remove all regularization terms in the energy term.

Our innovations result in a method that's more **principled**, **stable**, and **sensitive**.



facebook
Artificial Intelligence

Algorithm

1. Pick an area a
2. Use SGD to solve the optimization problem for a large λ:

$$\underset{\mathbf{m}}{\text{argmax}}\ \Phi(\text{smooth}(\mathbf{m}) \otimes \mathbf{x}) - \lambda \| \operatorname{vecsort}(\text{smooth}(\mathbf{m})) - \mathbf{r}_a \|^2$$

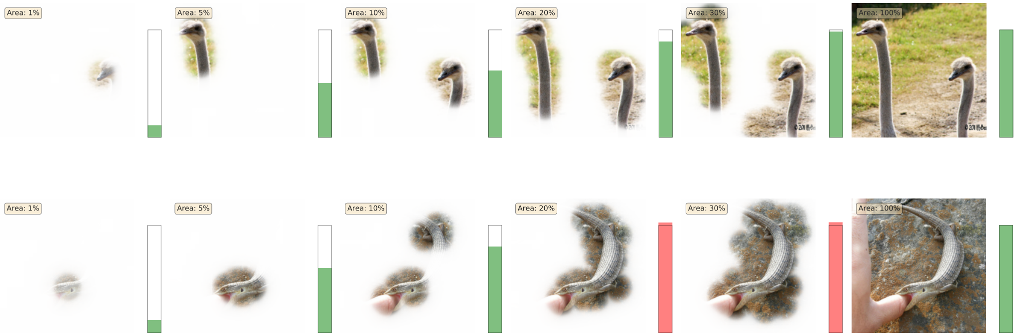
3. If needed, sweep a and repeat

facebook
Artificial Intelligence

Results

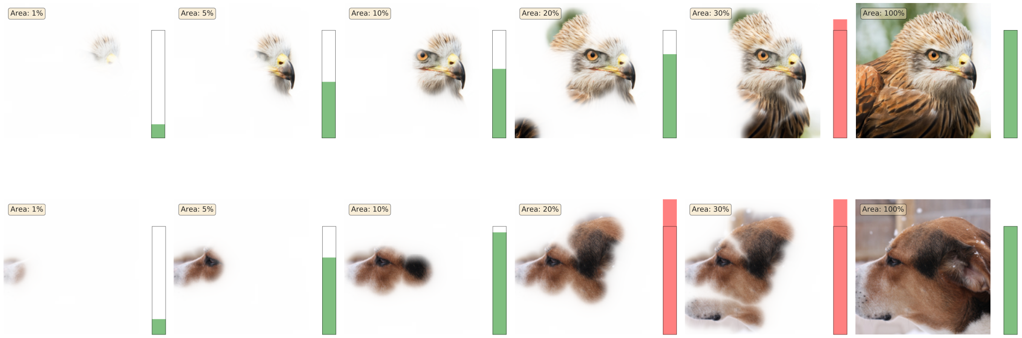
facebook
Artificial Intelligence

Foreground evidence is usually sufficient

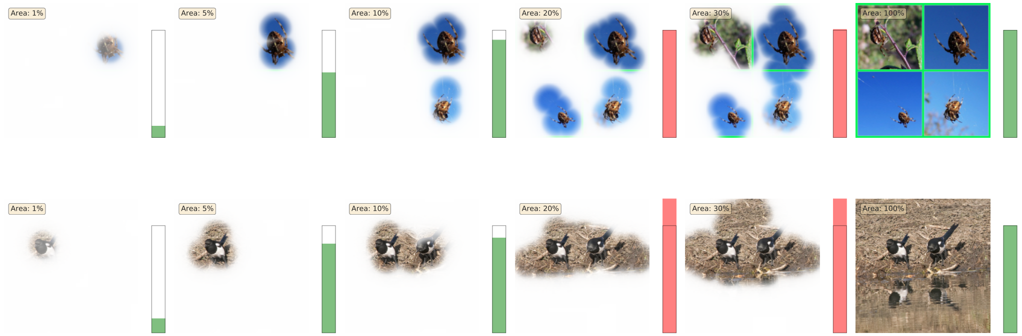


facebook
Artificial Intelligence

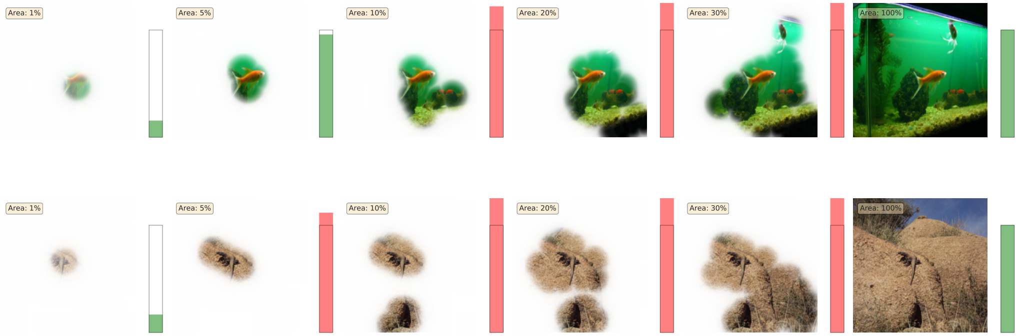
Large objects are recognised by their details



Small objects contribute cumulatively

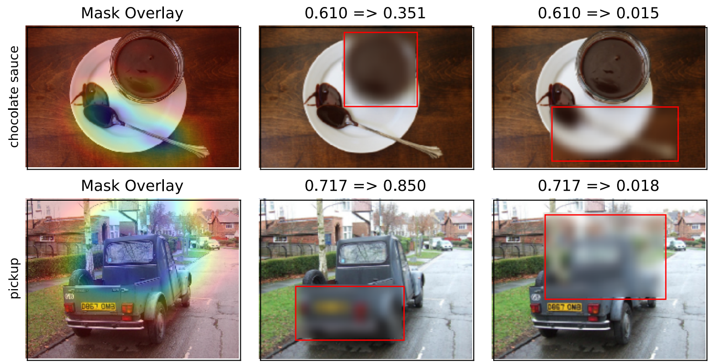


Suppressing the background may overdrive the network



Diagnosing networks

Example: the hot chocolate is recognized via the spoon and the truck vs the license plate

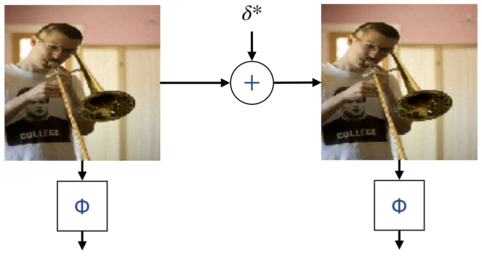


CNN fragility

Let $\mathbf{y} = \Phi(\mathbf{x})$ be the label predicted for image $\mathbf{x}$ by the deep net

Empirically, we can find tiny perturbations $\mathbf{x} + \delta$ that change $\mathbf{y}$ arbitrarily

$$\delta^* = \underset{\|\delta\|<\epsilon}{\operatorname{argmin}} \|\mathbf{y}_{\text{arbitrary}} - \Phi(\mathbf{x} + \delta)\|$$

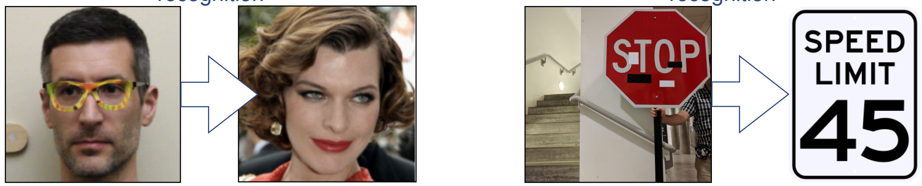


Intriguing properties of neural networks
Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus. CoRR, 2013

Dangerous adversaries

Adversarial glasses fooling face recognition

Adversarial stickers fooling sign recognition



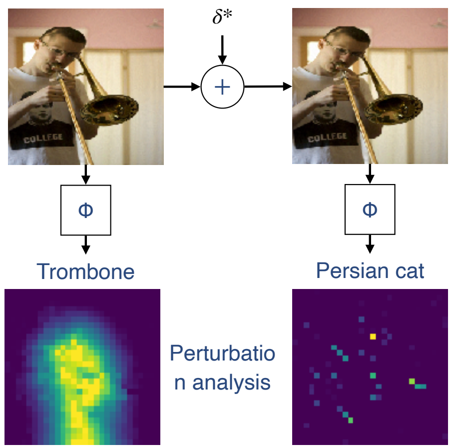
Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. Sharif, Bhagavatula, Bauer, Reiter. Proc. CSS, 2016.

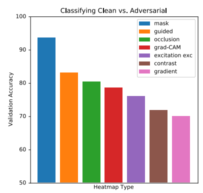
Robust physical-world attacks on machine learning models. Evtimov, Kevin Eykholt, Li, Prakash, Rahmati, Song. arXiv, 2017.

Adversarial defence

Method: recognize genuine vs adversarial images by learning a classifier on top of the saliency maps

(Illustrative of attribution, not really a recommended defence strategy!)





Assessing attribution

Assessing attribution: pointing game & weak localisation

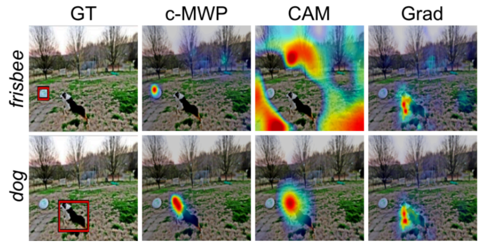
Goal: measure the spatial correlation between attribution maps and object occurrences

If the correlation is strong:

- the diagnosed model “understand” the object an
- the attribution method can tell

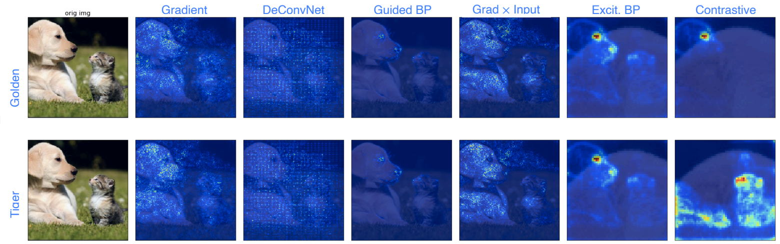
However, if the correlation is poor, *either*:

- the diagnoses model does not understand the object
  or
- the attribution method fails to tell



Assessing attribution: neuron sensitivity

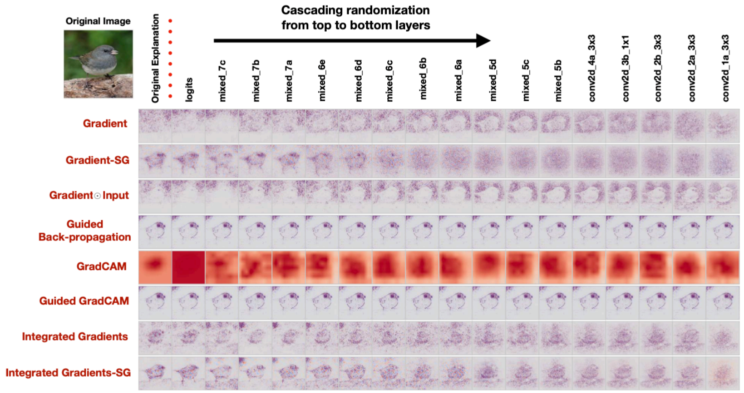
Attribution should generally result in a different output depending on which neon one wishes to visualise.



Assessing attribution: parameter sensitivity

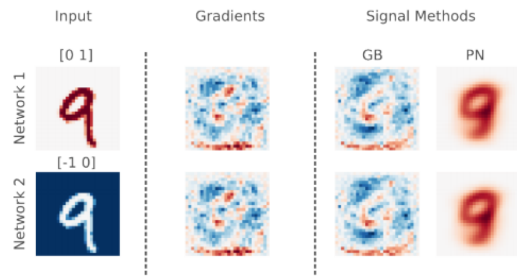
Attribution should also produce a different output if the model weights are different — e.g. random

Sanity checks for saliency maps. Adebayo, Gilmer, Muelly, Goodfellow, Hardt, Kim. Proc. NeurIPS, 2018.



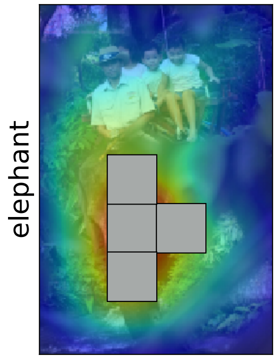
Assessing attribution: shift invariance

Learning how to explain neural networks: PatternNet and PatternAttribution. Kindermans, Schütt, Alber, Müller, Erhan, Kim, Dähne. Proc. ICLR, 2018.
Making convolutional networks shift-invariant again. Zhang. Proc. ICML, 2019.



Assessing attribution: perturbation analysis

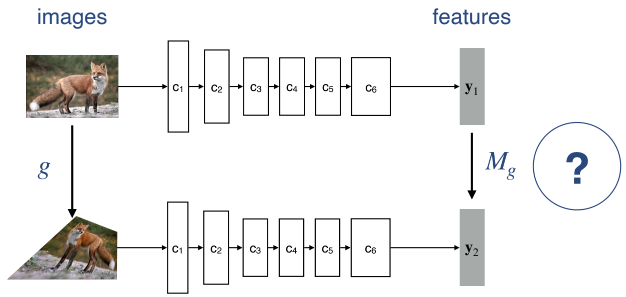
Display



Equivariance

Short answer: warping image usually reduces to sparse linear tf in feature space.

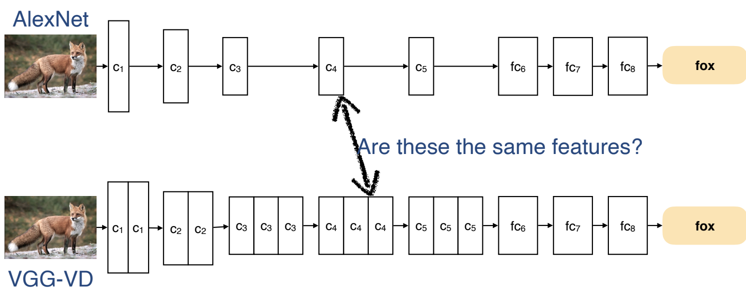
Long answer:
Understanding image representations by measuring their equivariance and equivalence. Lenc Vedaldi. CVPR 2015 & IJCV 2018



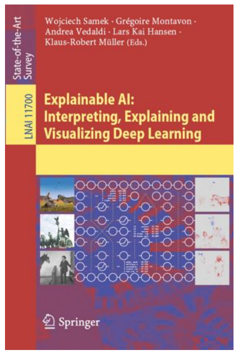
Equivalence

Short answer: there generally are corresponding features in different networks (up to 1x1 linear tfs).

Long answer
Understanding image representations by measuring their equivariance and equivalence. Lenc Vedaldi. CVPR 2015 & IJCV 2018



Collected references



Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Samek, Montavon, Vedaldi, Hansen, Muller, editors. Springer, 2019

Software

Captum

https://pytorch.org/captum/

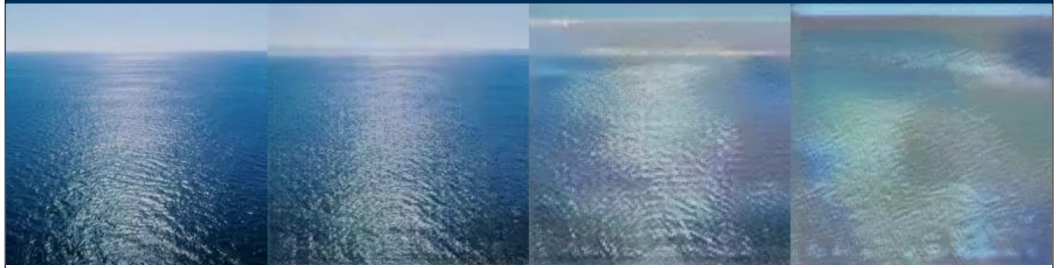
More than just vision

TorchRay

https://github.com/facebookresearch/TorchRay

Attribution, reproducibility, benchmarks

Summary



Universal Representation
- Compact representation families

Unsupervised Representation
- Self-supervision for learning features
- Self-supervision for learning structure
- What's in the prior

Understandable Representations
- Iconic visualizations
- Attribution
- Semantic identification