CS 103: Representation Learning, Information Theory and Control

Lecture 1, Jan 11, 2019

What is a task

Making a decision based on the data

Classification: Decide the class of an image (the prototypical supervised problem)

Survival: Decide the best actions to take to survive (Reinforcement Learning)

Reconstruction: Decide which information to store to reconstruct the data (generative models, unsupervised learning)

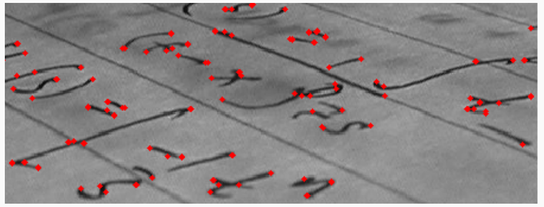
What is a representation

Any function of the data which is useful for a task.

Brightness

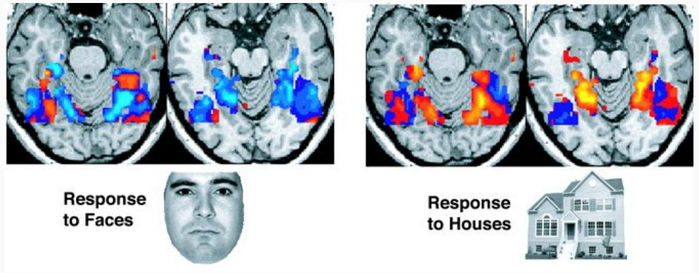
A simple organism may only need the light source direction.

Corners



Popular in Computer Vision before DNNs, central to visual inertial systems and AR.

Neuronal activity



Hidden Layer

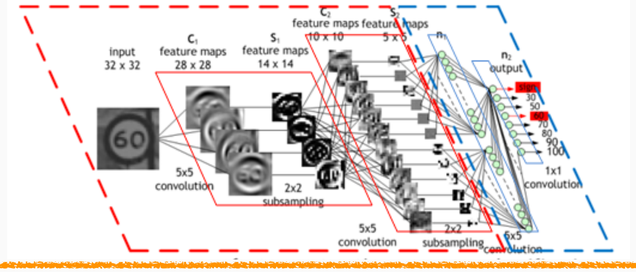
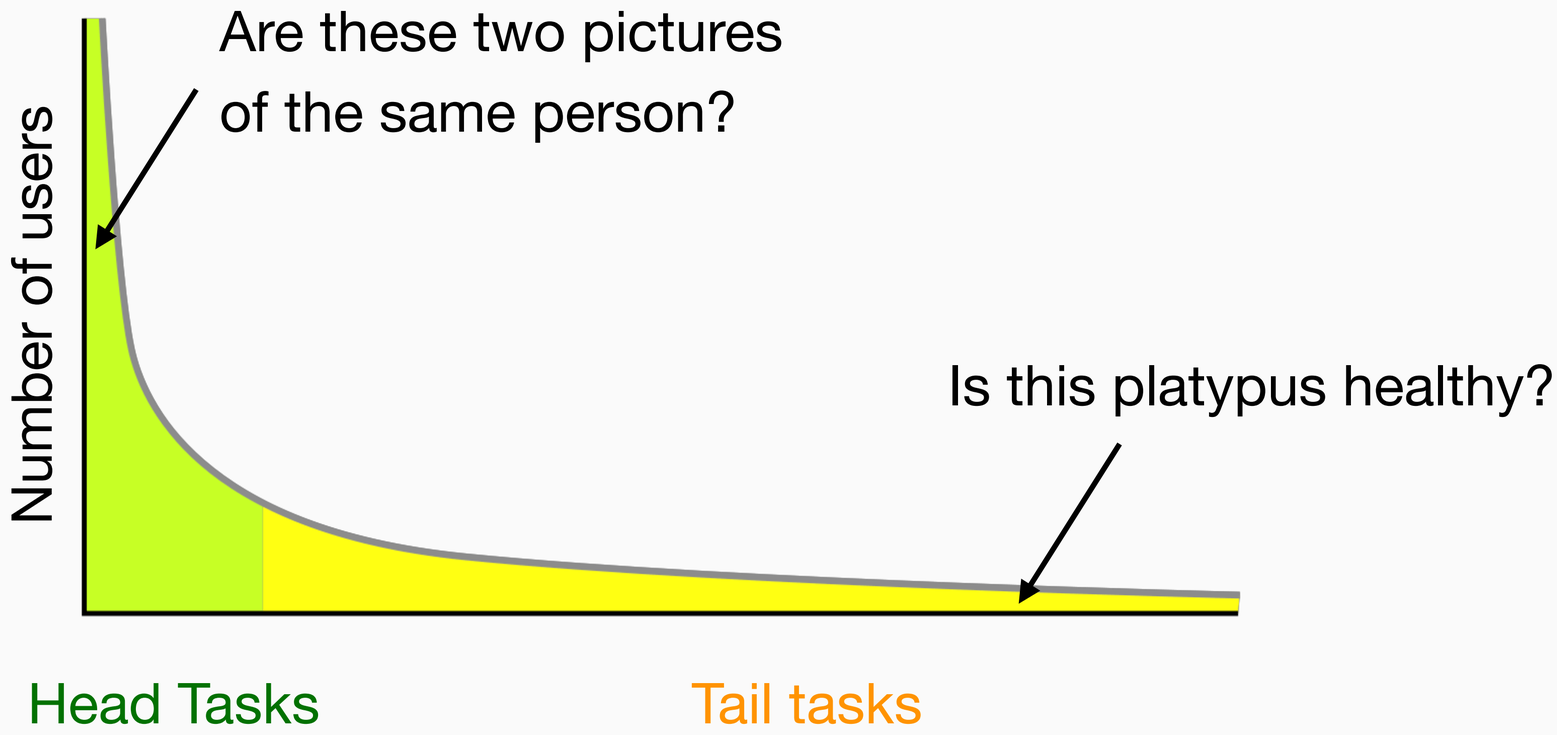


Image sources https://en.wikipedia.org/wiki/Functional_magnetic_resonance_imaging#/media/File:Haxby2001.jpg, https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/

Representation as a Service

We can try to solve to the most common tasks, but what about the tails?



Idea: Provide the user with a powerful and flexible representation that allows them to easily solve their task.

Representation as a Service

Google Cloud Why Google Solutions Products Pricing Getting started

AI & Machine Learning Products

Powerful image a

Cloud Vision offers both pretrained
custom models using AutoML Visio
use case.

Cloud Vision API enables developer
by encapsulating powerful machine
API. It quickly classifies images into
"sailboat"), detects individual object
printed words contained within ima
image catalog, moderate offensive
scenarios through image sentiment

Microsoft Azure Contact Sales: 1-800-867-1389 Search My account Portal Sign in

Overview Solutions Products Documentation Pricing Training Marketplace Partners Support Blog More Free account

Analyze an image

This feature returns information about visual content found in an image. Use tagging, domain-specific models, and descriptions in four languages to identify content and label it with confidence. Apply the adult/racy settings to help you detect potential adult content. Identify image types and color schemes in pictures.

VALUE

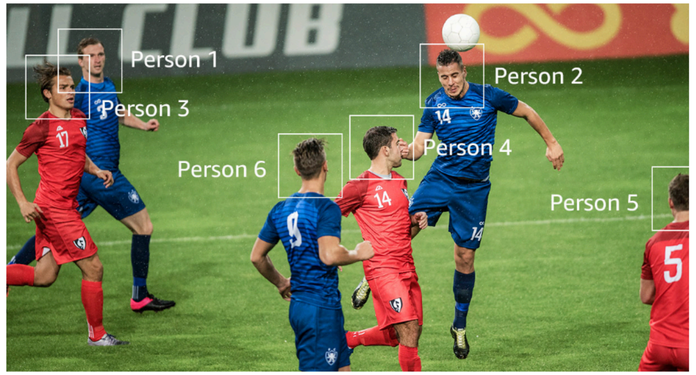
{ "tags": ["train", "platform", "station", "building", "indoor", "subway", "track", "walking", "waiting", "pulling", "board", "people", "man", "luggage", "standing", "holding", "large", "woman", "yellow", "suitcase"], "captions": [{ "text": "people waiting at a train station", "confidence": 0.8330993 }] }

[{ "name": "train", "confidence": 0.9975446 }, { "name": "platform", "confidence": 0.995543063 }, {

aws Contact Sales Support English My Account Create an AWS Account

Products Solutions Pricing Documentation Learn Partner Network AWS Marketplace Explore More

Amazon Rekognition Overview Features Pricing Getting Started Resources FAQs Customers



Facial analysis

You can analyze the attributes of faces in images and videos you provide to determine things like happiness, age range, eyes open, glasses, facial hair, etc. In video, you can also measure how these things change over time, such as constructing a timeline of the emotions of an actor.

Pathing

You can capture the path of people in the scene when using Amazon Rekognition with video files. For example, you can use the movement of athletes during a game to identify plays for post-game analysis.

Representation as a Service

1. What is the best representation for a task?
2. Which tasks can we solve using a given representation?

   The representation used by an health provider is probably not useful to a movie recommendation system.
3. Can we build a “universal” representation?
4. Can we fine-tune a representation for a particular task?
5. Can we provide the user with error bounds? Privacy bounds?

But what is a good representation?

Data Processing Inequality:

No function of the data (representation) can be better than the data themself for decision and control (task).

However, most organisms and algorithms use complex representations that deeply alter the input. In Deep Learning we regularly torture the data to extract the results:

Three main ingredients of DNNs: Convolutions, ReLU, Max-Pool

Destroy information

Questions

Is the destruction of information necessary for learning?

Why some properties (invariance, hierarchical organization) emerge naturally in very different systems?

Why do we need to forget?

Let's assume we want to learn a classifier $p(y \mid x)$ given an input image x.

Curse of dimensionality: In general, to approximate $p(y \mid x)$ the number of samples should scale exponentially with the number of dimensions.

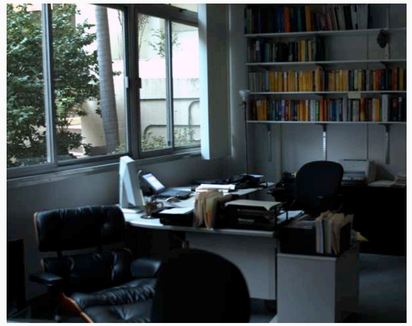
If x is a 256x256 image, this means we would need ~10^{28462} samples.

Then, how can we learn on natural images?

1. Nuisance invariance (reduce the dimension of the **input**)
2. Compositionally (reduce the dimension of the **representation space**)
3. Complexity prior on the solution (reduce the dimension of **hypothesis space**)

Nuisance invariance

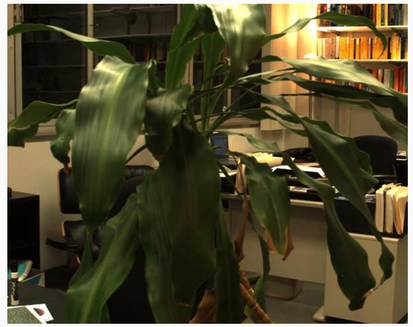
Nuisance variability



$I = h(\xi, \nu)$

Change of nuisance

$\tilde{I} = h(\xi, \tilde{\nu}), \;\; \tilde{\nu} = \text{illumination}$



$\tilde{\nu} = \text{visibility}$

$\tilde{\nu} = \text{viewpoint}$

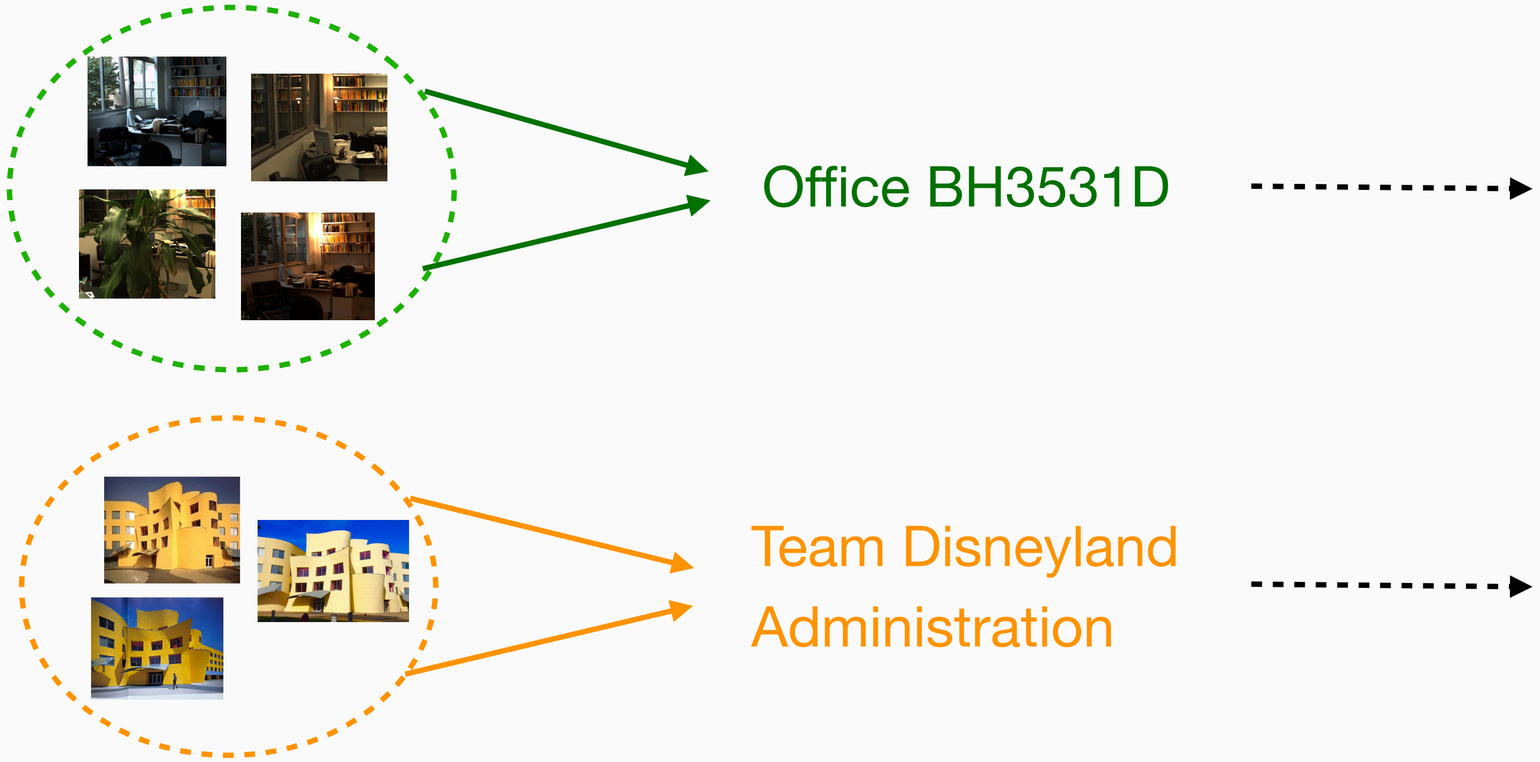
$\tilde{I} = h(\tilde{\xi}, \tilde{\nu}), \;\; \tilde{\xi} \neq \xi$

Change of identity

Images from *Steps Toward a Theory of Visual Information*, S. Soatto, 2011

How to use nuisance variability

A good representation should collapse images differing only for nuisance variability.



Quotienting with respect to nuisances reduces the dimensionality of the space of images, and simplifies learning the successive parts of the pipeline.

Group nuisances

Examples: Translations, rotations, change of scale/contrast, small diffeomorphisms

Given a group *G* acting on the space of data *X*, we say that a representation f(x) is invariant to *G* if:

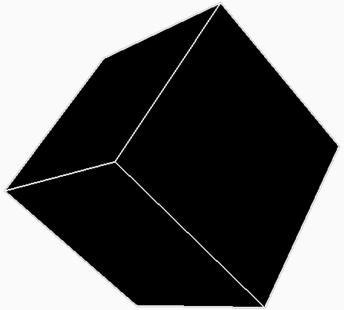
$$f(x) = f(g \circ x) \qquad \text{for all} \quad g \in G, x \in X$$

A representation is *maximal invariant* if all other invariant representations are a function of it.

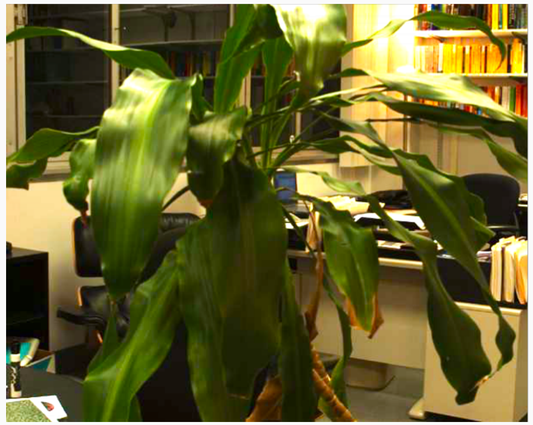
Well understood for translation and scale (week 2). The solution inspired and justifies the use of convolutions and max-pooling.

Problems with group nuisances

1. Rapidly becomes difficult for more complex groups
2. Groups acting on 3D objects do not act as groups on the image



3. Not all nuisances are groups (*e.g.,* occlusions)



More general nuisances

Idea: A nuisance as everything that does not carry information about the task.

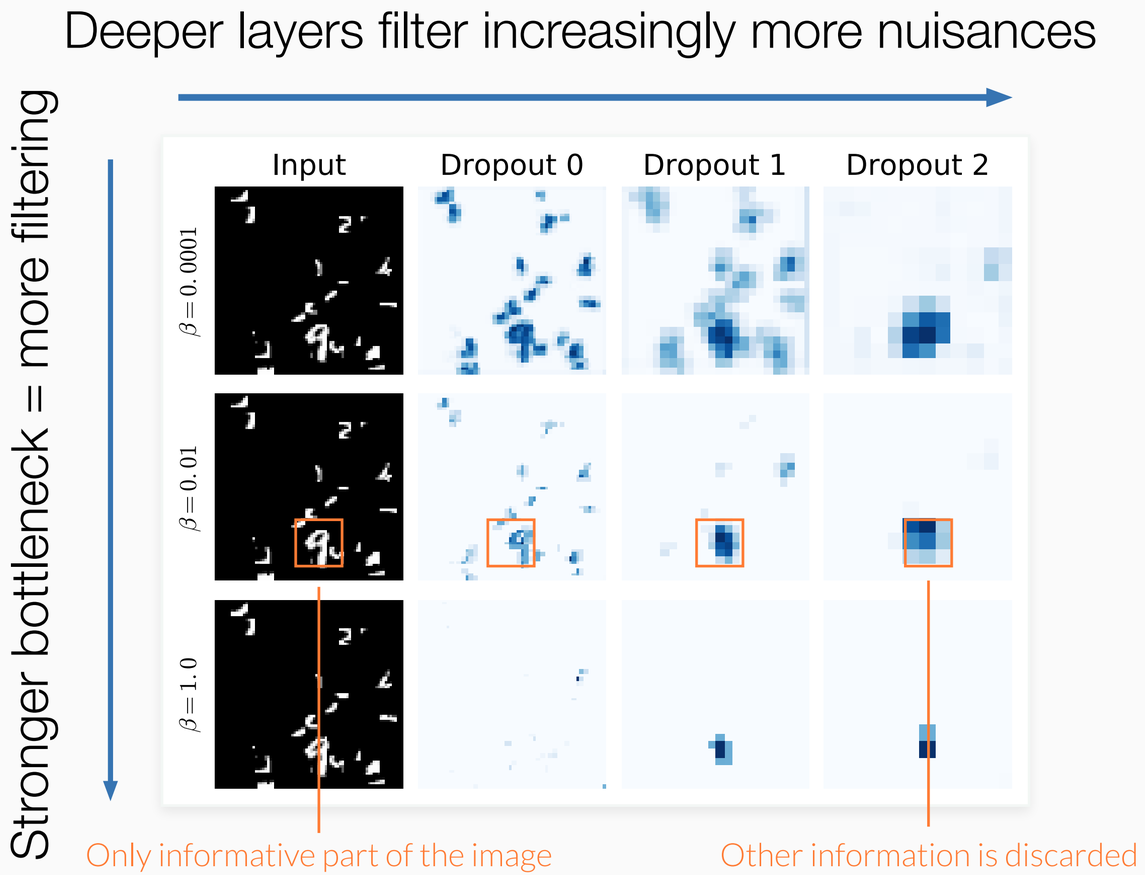
Introduce the Information Bottleneck Lagrangian:

$$\min_f \underbrace{I(f(x); x)}_{\text{Total information}} - \lambda \underbrace{I(f(x); \text{task})}_{\text{Information the representation has about the task}}$$

where *I(x; y)* is the mutual information. The solution to the Lagrangian (for $\lambda \to +\infty$) is a maximally invariant representation for all nuisances (week 4).

We can thus rephrase the problem of nuisance invariance as a much simpler variational optimization problem.

Learning invariant representations



Achille and Soatto, *"Information Dropout: Learning Optimal Representations Through Noisy Computation"*, PAMI 2018 (arXiv 2016)

Compositional representations

Compositional representations

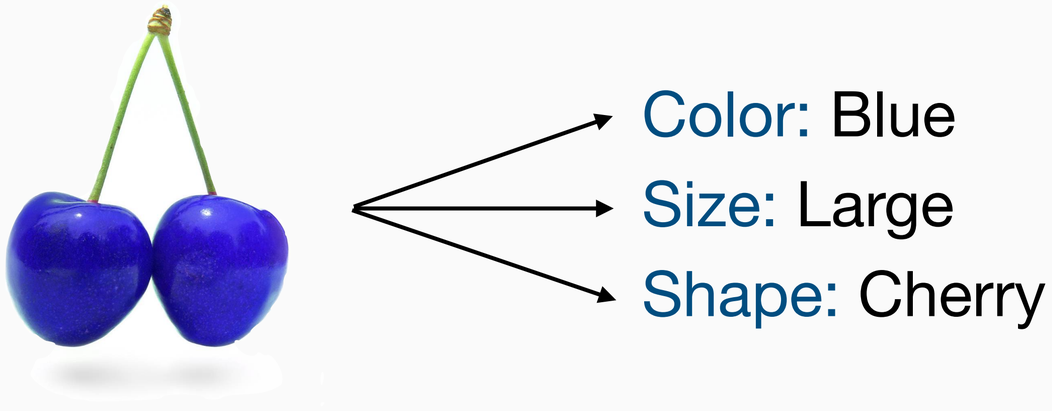
Humans can easily solve task by combining concepts:

“Find a blue large cherry”

We can easily solve this task, even if we have never seen a blue cherry before.

Compositionally requires disentanglement

To learn a good compositional representation, we first need to learn to decompose the image in reusable semantic factors:



This mitigates the curse of dimensionality: each factor is easy to learn, but combined they yield exponentially many objects.

Factors of variation can be learnt in succession in a life-long learning setting and used in the future for one-shot or zero-shot learning.

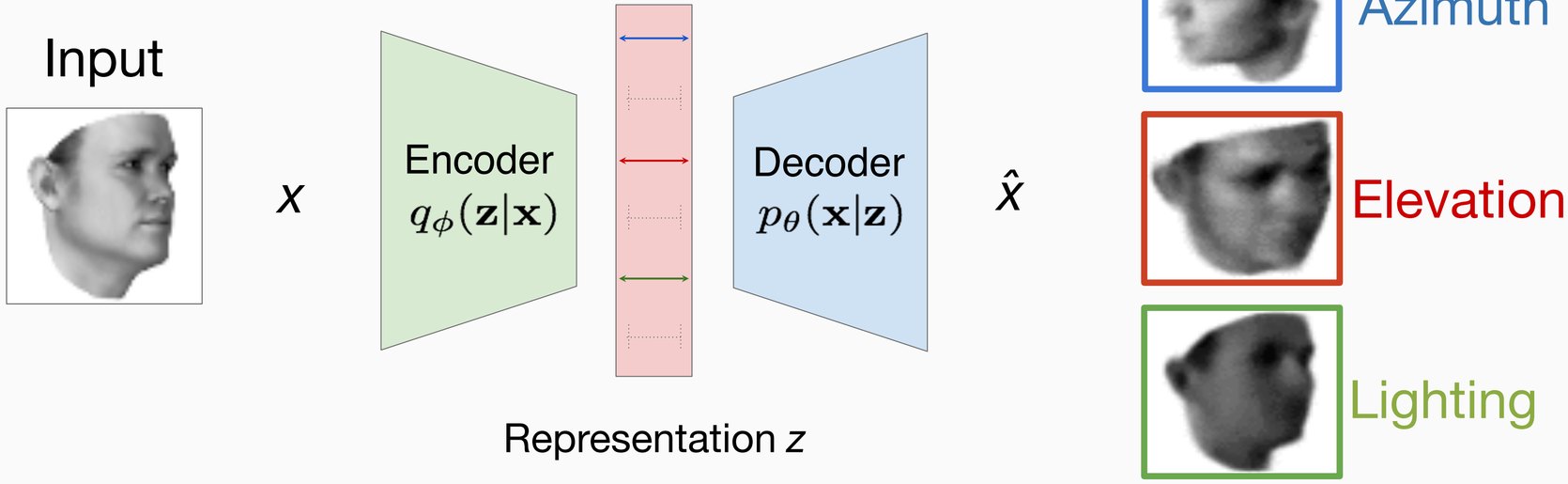
Problem. But what are “semantic factors of variation”?

Learning disentangled representations

(Higgins et al., 2017, Burgess et al., 2017)

Possible answer through the Minimum Description Length principle (week 7):

$$\mathcal{L}_{\text{MDL}}(\phi,\theta) = \underbrace{\mathbb{E}_{\mathbf{z}^s \sim q_\phi(\cdot|\mathbf{x}^s)}[-\log p_\theta(\mathbf{x} \mid \mathbf{z}^s, s)]}_{\text{Reconstruction error}} + \gamma \Big| \underbrace{\mathbb{KL}(q_\phi(\mathbf{z}^s|\mathbf{x}^s)||p(\mathbf{z}))}_{\text{Representation capacity}} - \underbrace{C}_{\text{Target}} \Big|^2$$



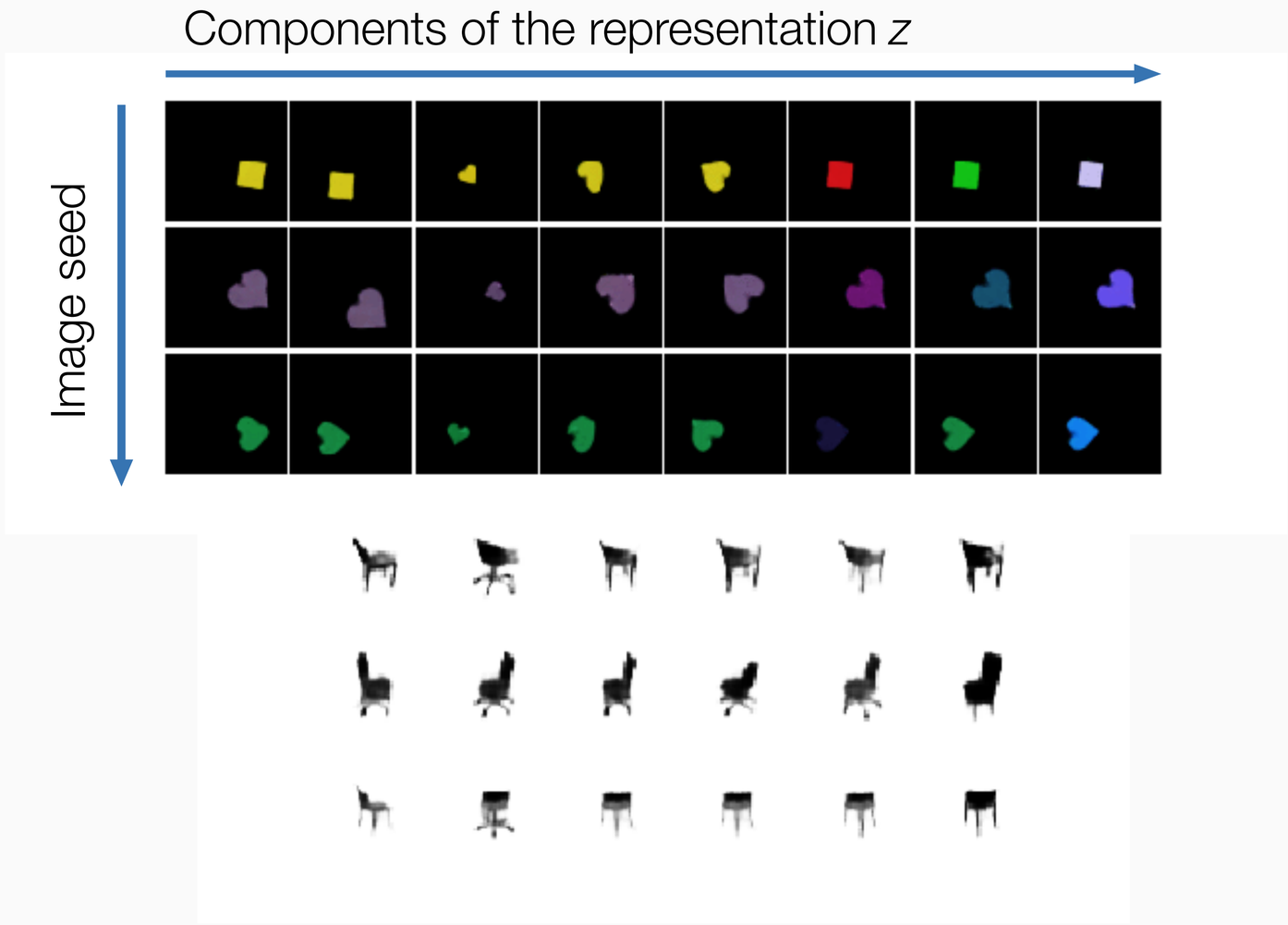
Higgins et al., β-*VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*, 2017
Burgess et al., *Understanding Disentangling in beta-VAE"* 2017

Pictures courtesy of Higgins et al., Burgess et al.

Learning disentangled representations

(Higgins et al., 2017, Burgess et al., 2017)

Possible answer through the Minimum Description Length principle (week 7):



Higgins et al., β-*VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*, 2017
Burgess et al., *Understanding Disentangling in beta-VAE"* 2017

Pictures courtesy of Higgins et al., Burgess et al.

Complexity of the classifier

1. Nuisance invariance (reduce the dimension of the input)
2. Compositionally (reduce the dimension of the representation)
3. **Complexity prior on the solution** (reduce the dimension of hypothesis space)

We can define the (Kolmogorov) complexity of a classifier as the length of the shortest program implementing it. Leads to the PAC-Bayes bound:

PAC-Bayes bound (Catoni, 2007; McAllester 2013).

$$L_{\text{test}}(q(w|\mathcal{D})) \leq \frac{1}{N(1-1/2\beta)} \underbrace{\left[H_{p,q}(y|x,w) + \beta\ \text{KL}(q(w|\mathcal{D})\|p(w))\right]}_{\text{IB Lagrangian for the weights}}$$

Emergence of invariant and disentangled representations

Weeks 5-6

Theorem 1 (informal). Stochastic gradient descent biases the optimization process toward recovering low-complexity solutions.

$$p(w_f, t_f | w_0, t_0) = e^{-\Delta\mathcal{L}(w;\mathcal{D})} \int_{w_0}^{w_f} e^{-\frac{1}{2D}\int_{t_0}^{t_f} \frac{1}{2}\dot{u}(t)^2 + V(u(t))dt} du(t)$$

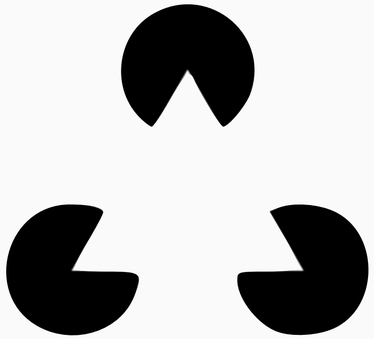
Theorem 2 (informal). In DNNs, low-complexity classifier have invariant and disentangled representations.

Corollary (Theorem 1 + 2). DNNs are biased toward learning invariant and disentangled representations.

Information and actions

The MDL principle allows top-down inference

The MDL principle allows correct interpretation of low-level features through the interpretation that makes it easier to explain the global image.



Which sometimes can go wrong:

Inputs are ambiguous, fortunately we can move

Single inputs are often hard or impossible to interpret correctly. However, intelligent agents can move to acquire more information.

Without assuming a prior, we can’t detect objects from a single image.

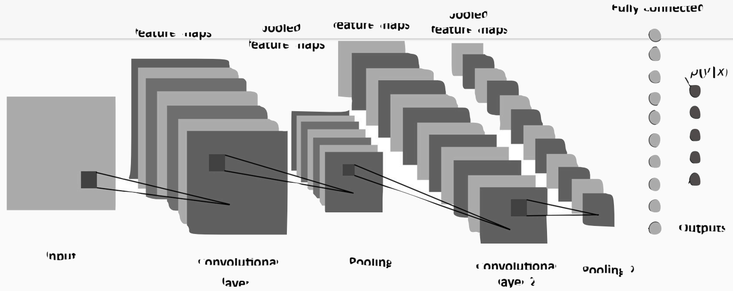
Image courtesy of Preventable.com

The connection between intelligence and control

Tunicate, is an organism capable of mobility until it finds a suitable rock to cement itself in place. Once it becomes stationary, it digests its own cerebral ganglion cells.

Image and caption from *Steps Toward a Theory of Visual Information*, S. Soatto, 2011

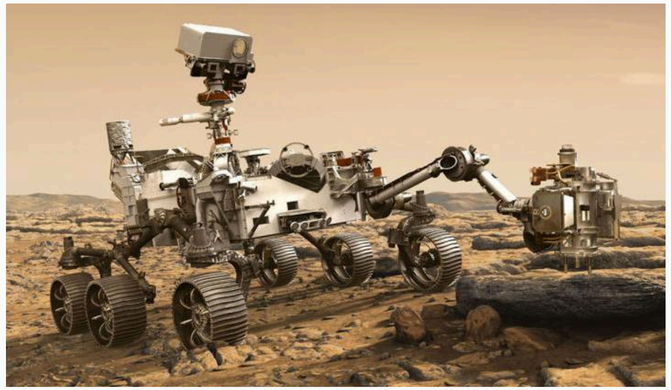
Embodied Intelligence



Cognition

Sensing

Action



Representations for Embodied Intelligence

Unlike standard machine learning, we can act on the environment to collect more data or modify the state of the system.

The representation we learn should interact with control. In particular:

1. What is the best action to take to minimize the uncertainty of the representation?
2. Is the representation grounded in the environment? For example, what happens if we move one single object? Will only one component of the representation change?