Learning perception independent of goals: self-supervised learning

- An embodied agent can’t depend on labels
- Reinforcement learning goals are inevitably tied to particular goals / tasks
- Need another way to build good feature representation.
Pretext task

• Labels are unavailable
• Idea: create your own labels from data
• “Pretext” task
• Hope: Solving the task leads to good feature representations
Pretext tasks

• Transform input, task network with predicting transformation
Pretext tasks

- Remove data, then task network with predicting it
Pretext tasks

• Clustering

Use pseudo-labels to produce representation

Use representation to cluster dataset
Pretext tasks

- Use some source with additional data
- E.g. videos
1. Collect videos

2. Segment using motion

3. Train ConvNet

What do humans do?

What do humans do?

What do humans do?

What do humans do?


Patients recovering from blindness
The kitten carousel

• Both kittens see same visual input
• Active kitten learns well, passive kitten does not. Why?
  • Knowledge of motion?
  • Actively choosing action?
  • Paid more attention?


Ego-motion ↔ vision: view prediction

After moving:

Slide credit: Dinesh Jayaraman
Approach idea: Ego-motion equivariance

Training data
Unlabeled video + motor signals

Equivariant embedding
organized by ego-motions

Pairs of frames related by similar ego-motion should be related by same feature transformation

Slide credit: Dinesh Jayaraman
Self supervision by reconstructing hidden data

Doersch et al, ICCV 2015

Pathak et al, CVPR 2016

Zhang et al, ECCV 2016
Self supervision by using external information

Wang et al, ICCV 2015


Pathak et al, CVPR 2017

Owens et al, CVPR 2016
The unreasonable effectiveness of ImageNet
The unreasonable effectiveness of ImageNet
Why is ImageNet so good?

• Classification?
• Curated images?
• Many diverse classes?
• Related to target tasks?
Lessons from human cognition

• Babies experience of the world is profoundly multi-modal
• Babies develop incrementally, and they are not smart at the start.
• Babies live in a physical world, full of rich regularities that organize perception, action, and ultimately thought.
• Babies explore – they move and act in highly variable and playful ways that are not goal-oriented and are seemingly random.
• Babies act and learn in a social world in which more mature partners guide learning and add supporting structures to that learning
• Babies learn a language, a shared communicative system that is symbolic.

Weak supervision

Beating supervised learning with tons and tons of data

Two phases of supervision

Base task (ImageNet)

Representation learning

Model

Finetuning

New task

New Model

Evaluation
Different forms of reduced supervision

• Weakly supervised
  • Use less rich annotation / noisy annotation

• Semi supervised
  • Use a few labeled images and a bank of unlabeled images

• Few-shot
  • Use a few labeled images

• Zero-shot
  • Use no labeled images but side-information
Weak supervision for detection

• Can we learn object detection / semantic segmentation with only image level labels?
• Idea: image label = “cat” => somewhere in the image there is a cat
Multiple instance learning
Multiple instance learning

- Bag is negative if all instances are negative
- Bag is positive if one or more instances are positive

\[ p_i = \max_j p_{ij} \quad \text{and} \quad p_i = 1 - \prod_j (1 - p_{ij}) \]
Few-shot learning
The challenge: Intra-class variation
“Train set”

Philippine Tarsier

“Test set”

Philippine Tarsier

Mouse lemur

Beaver
Key cue: shared modes of variation
How do humans do this?

Bird, grey color, long beak, long legs, black markings on head,…

More invariant representations

Inductive biases during learning
Better representations: metric learning

True class boundary
Better representations: metric learning

“One-shot” class boundary
Metric learning

• Pull same-class pairs closer and different-class pairs apart
• Contrastive loss (DrLIM)
  • $= d(x, x')^2$ if $y = y'$
  • $= \max(0, m - d(x, x'))^2$ if $y \neq y'$
• Triplet loss
  • $= \max(d(x, x_+) - d(x, x_-) + \gamma, 0)$

Meta-learning

• Given:

  Small training set
  (few training examples)

• Produce:

• Idea: Make this a learnable function!
Meta-learning

\[ S_{\text{train}} \]

\[ S_{\text{test}} \]

\[ h \]

\[ \hat{p} \]

\[ m = \mathcal{A}(S_{\text{train}}) \]

\[ \hat{p} = m(x) \]
Meta-learning

\[ S_{train} \]

\[ h \]

\[ S_{test} \]

\[ \hat{p} \]
Meta-learning: training

$D$

$S_{train}$

$S_{test}$

$h$

$w$

$\hat{p}$

Image

Label
An army of meta-learners

Meta-learning: Prototypical Networks