Self-Supervised Learning

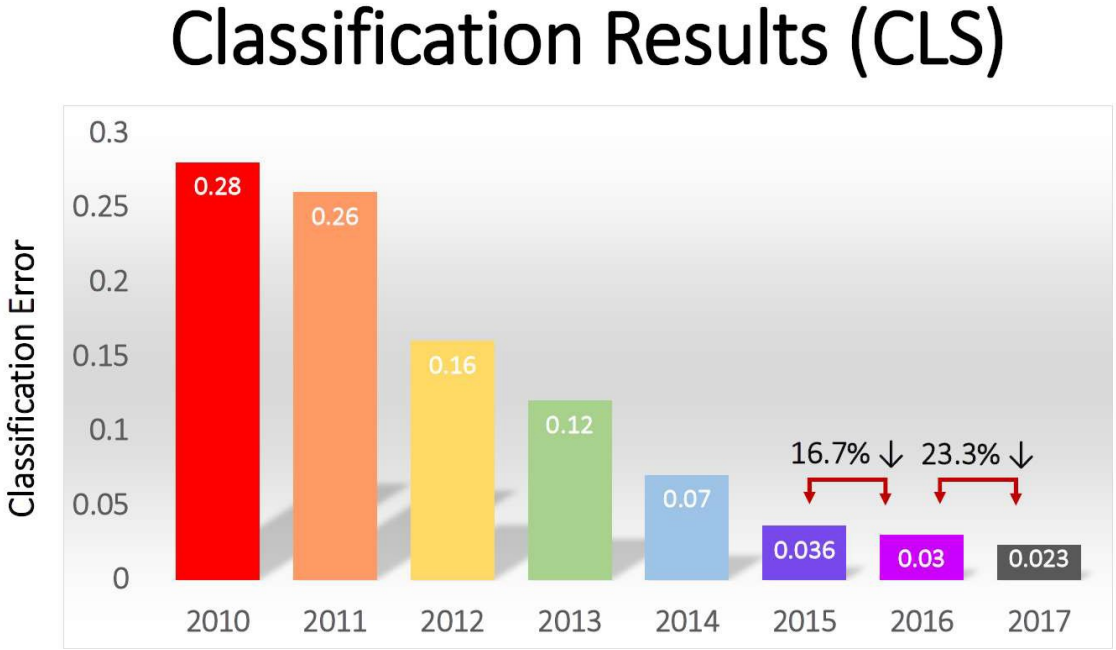
Andrew Zisserman, Relja Arandjelović

The ImageNet Challenge Story …

1000 categories

- Training: 1000 images for each category
- Testing: 100k images

The ImageNet Challenge Story … strong supervision



Strong supervision:

- Features from networks trained on ImageNet can be used for other visual tasks, e.g. detection, segmentation, action recognition, fine grained visual classification

- To some extent, any visual task can be solved now by:
  1. Construct a large-scale dataset labelled for that task
  2. Specify a training loss and neural network architecture
  3. Train the network and deploy

- Are there alternatives to strong supervision for training? Self-Supervised learning ….

Why Self-Supervision?

1. Expense of producing a new dataset for each new task

2. Some areas are supervision-starved, e.g. medical data, where it is hard to obtain annotation

3. Untapped/availability of vast numbers of unlabelled images/videos
   - Facebook: one billion images uploaded per day
   - 300 hours of video are uploaded to YouTube every minute

4. How infants may learn …

Self-Supervised Learning

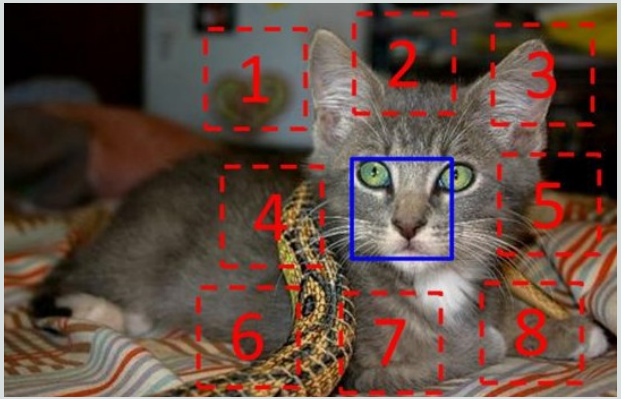
The Scientist in the Crib: What Early Learning Tells Us About the Mind
by Alison Gopnik, Andrew N. Meltzoff and Patricia K. Kuhl

The Development of Embodied Cognition: Six Lessons from Babies
by Linda Smith and Michael Gasser

What is Self-Supervision?

- A form of unsupervised learning where the data provides the **supervision**
- In general, withhold some part of the data, and task the network with predicting it
- The task defines a proxy loss, and the network is forced to learn what we really care about, e.g. a semantic representation, in order to solve it

Inferring structure



[“Unsupervised visual representation learning by context prediction”, Doersh et al. 15]

Context prediction

Can you guess the spatial configuration for the two pairs of patches?

Question 1:

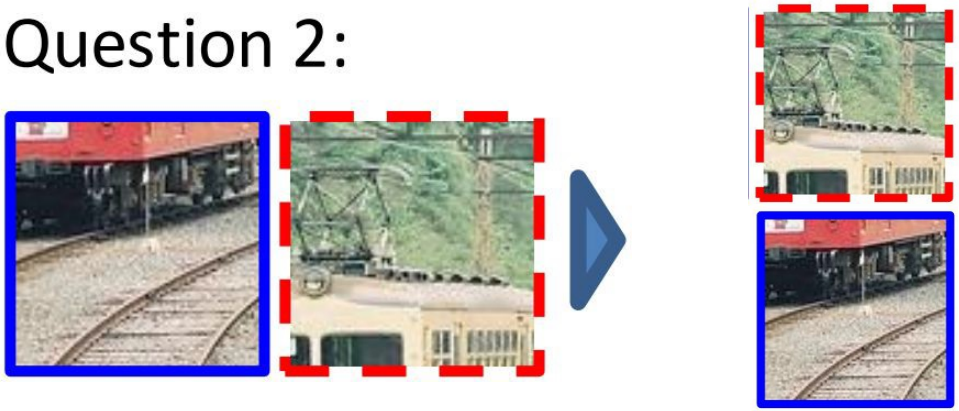
Question 2:

Context prediction

Can you guess the spatial configuration for the two pairs of patches? Much easier if you recognize the object!

Question 1:

Question 2:

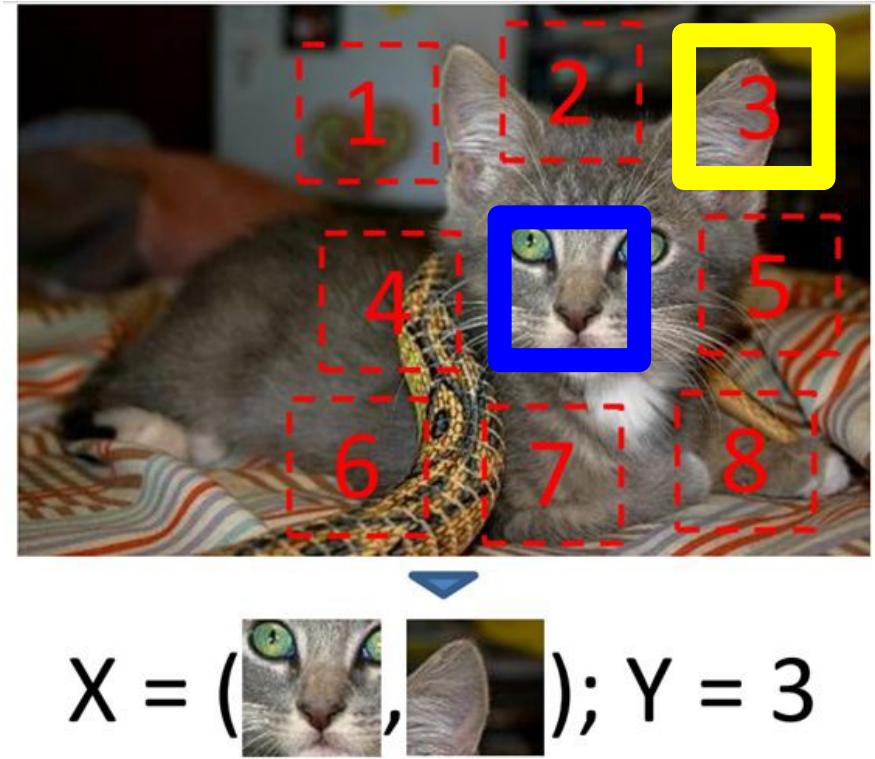


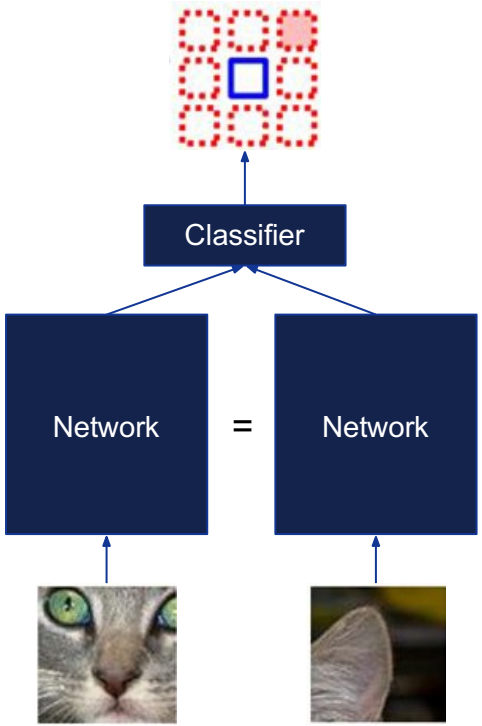
Intuition

- The network should learn to recognize object parts and their spatial relations

[“Unsupervised visual representation learning by context prediction”, Doersh et al. 15]

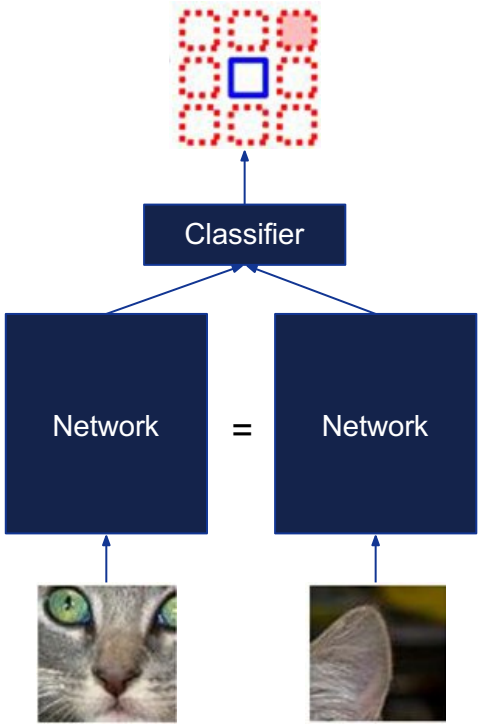
Context prediction





[“Unsupervised visual representation learning by context prediction”, Doersh et al. 15]

Context prediction



Pros

- (arguably) The first self-supervised method
- Intuitive task that should enable learning about object parts

Cons

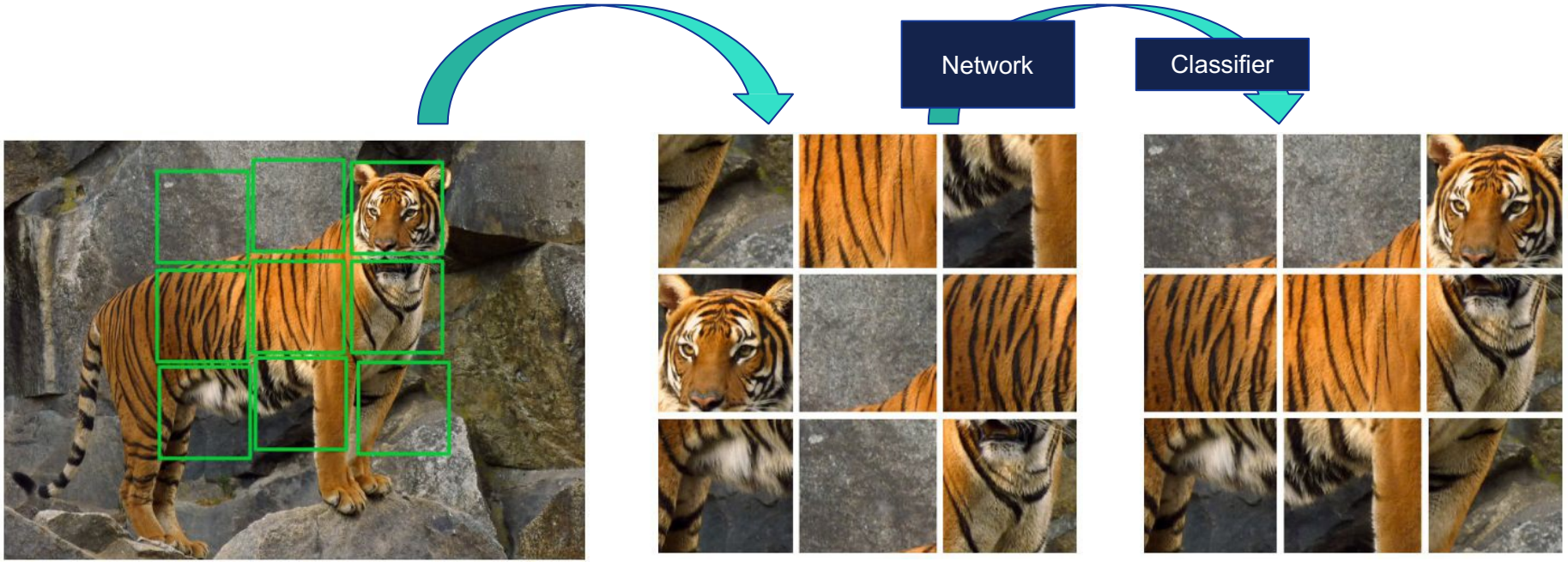
- Assumes training images are photographed with canonical orientations (and canonical orientations exist)
- Training on patches, but trying to learn image representations
- Networks can “cheat” so special care is needed [discussed later]
  - Further gap between train and eval
- Not fine-grained enough due to no negatives from other images
  - e.g. no reason to distinguish cat from dog eyes
- Small output space - 8 cases (positions) to distinguish?

[“Unsupervised learning of visual representations by solving jigsaw puzzles”, Noroozi et al. 17]

Jigsaw puzzles

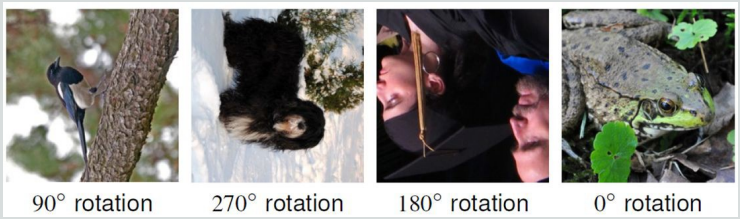
Divide image into patches and permute them

Predict the permutation



Pros & Cons: Same as for context prediction apart from being harder

Transformation prediction

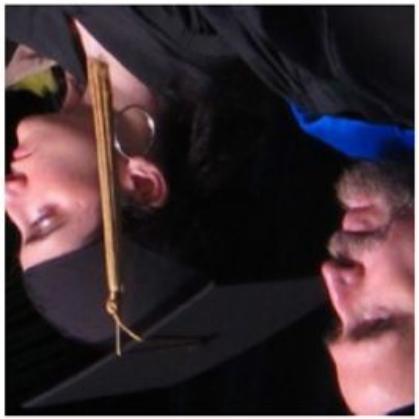


[“Unsupervised representation learning by predicting image rotations”, Gidaris et al. 18]

Rotation prediction

Can you guess how much rotated is applied?





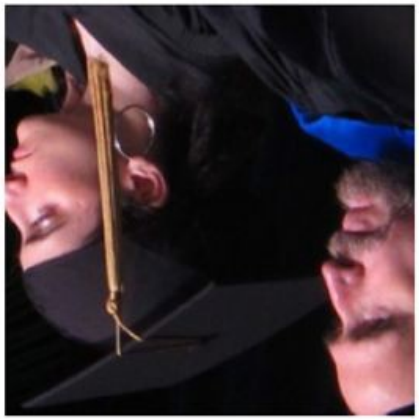

[“Unsupervised representation learning by predicting image rotations”, Gidaris et al. 18]

Rotation prediction

Can you guess how much rotated is applied? Much easier if you recognize the content!

90° rotation

270° rotation

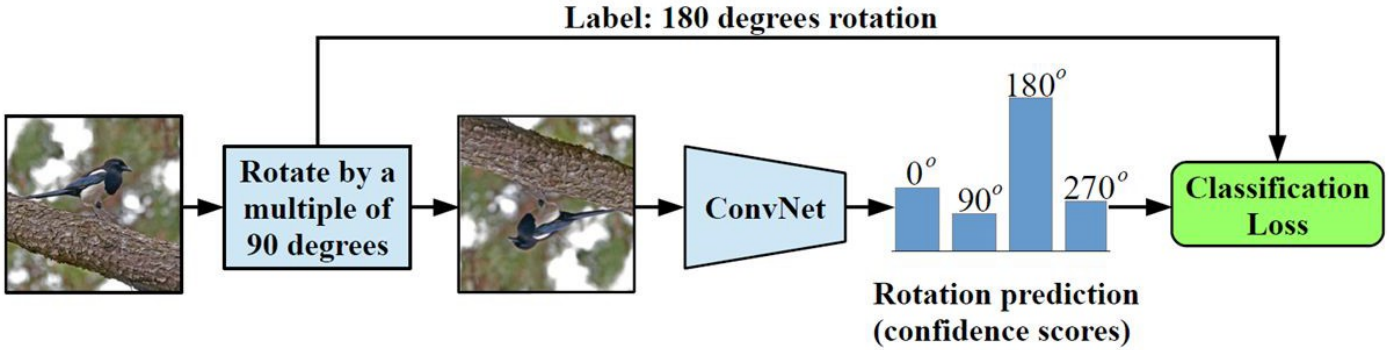


180° rotation

0° rotation

[“Unsupervised representation learning by predicting image rotations”, Gidaris et al. 18]

Rotation prediction



Pros

- Very simple to implement and use, while being quite effective

Cons

- Assumes training images are photographed with canonical orientations (and canonical orientations exist)
- Train-eval gap: no rotated images at eval
- Not fine-grained enough due to no negatives from other images
  - e.g. no reason to distinguish cat from dog
- Small output space - 4 cases (rotations) to distinguish [not trivial to increase; see later]
- Some domains are trivial e.g. StreetView ⇒ just recognize sky

[“AET vs. AED: Unsupervised representation learning by auto-encoding transformations rather than data”, Zhang et al. 19]

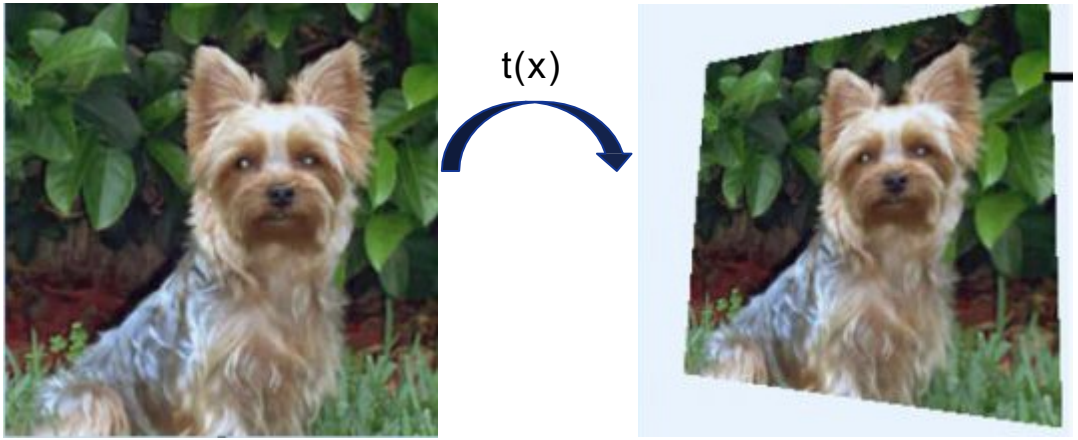
Relative transformation prediction

Estimate the transformation between two images.

[“AET vs. AED: Unsupervised representation learning by auto-encoding transformations rather than data”, Zhang et al. 19]

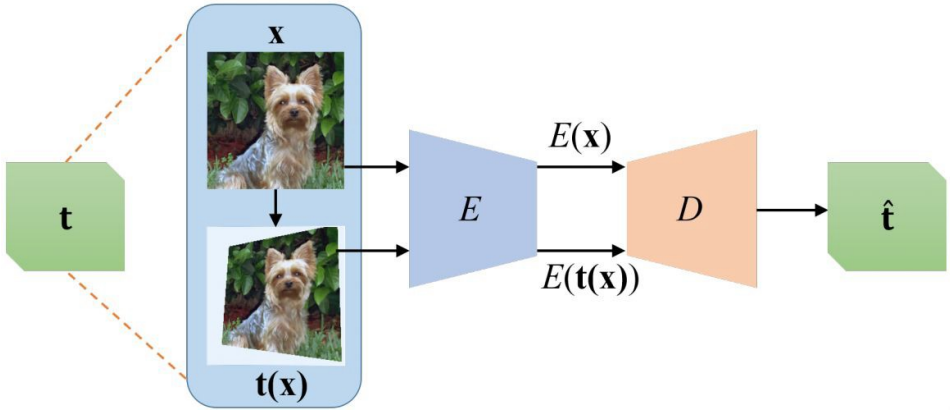
Relative transformation prediction

Estimate the transformation between two images. Requires good features



[“AET vs. AED: Unsupervised representation learning by auto-encoding transformations rather than data”, Zhang et al. 19]

Relative transformation prediction



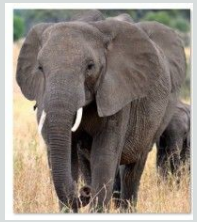
Pros

- In line with classical computer vision, e.g. SIFT was developed for matching

Cons

- Train-eval gap: no transformed images at eval
- Not fine-grained enough due to no negatives from other images
  - e.g. no reason to distinguish cat from dog
- Questionable importance of semantics vs low-level features (assuming we want semantics)
  - Features are potentially not invariant to transformations

Reconstruction

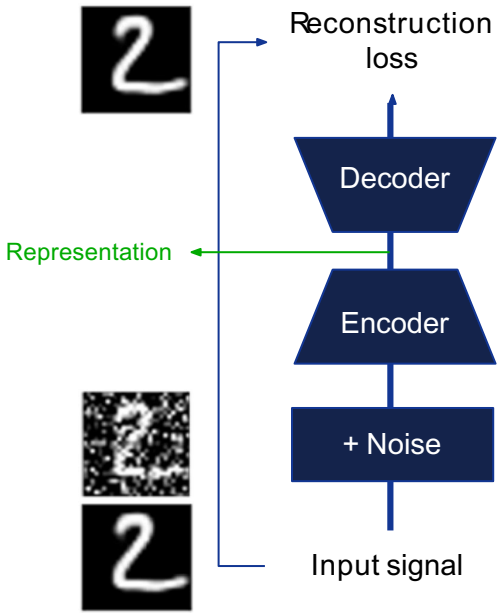


[“Extracting and composing robust features with denoising autoencoders”, Vincent et al. 08]

Denoising autoencoders

What is the noise and what is the signal?

Recognizing the digit helps!



Pros

- Simple classical method
- Apart from representations, we get a denoiser for free

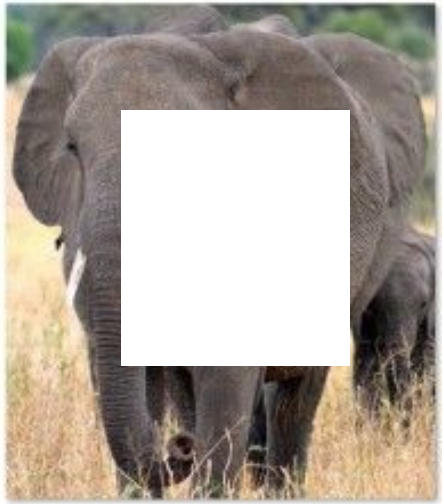
Cons

- Train-eval gap: training on noisy data
- Too easy, no need for semantics - low level cues are sufficient

[“Context encoders: Feature learning by inpainting”, Pathak et al. 16]

Context encoders

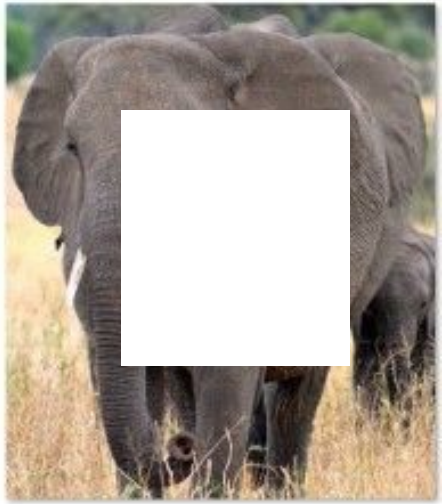
What goes in the middle?

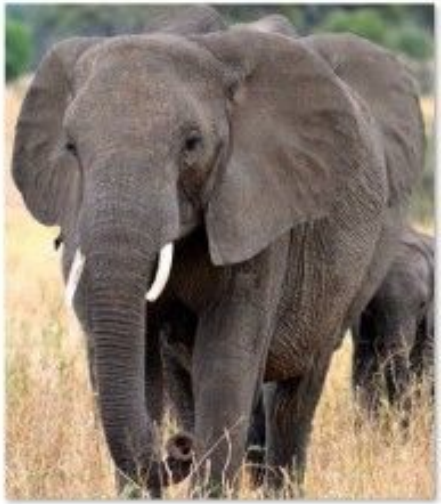


[“Context encoders: Feature learning by inpainting”, Pathak et al. 16]

Context encoders

What goes in the middle? Much easier if you recognize the objects!





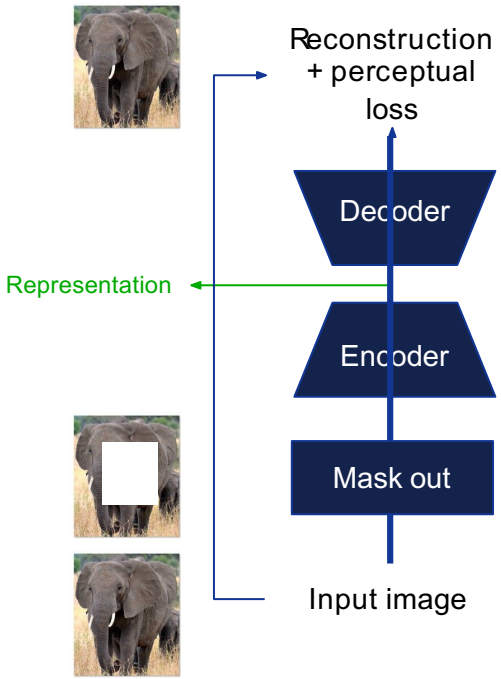
Natural language processing (e.g. word2vec, BERT)

All [MASK] have tusks. ⇒ All elephants have tusks.

[“Distributed representations of words and phrases and their compositionality”, Mikolov et al. 13]
[“BERT: Pre-training of deep bidirectional transformers for language understanding”, Devlin et al. 18]

[“Context encoders: Feature learning by inpainting”, Pathak et al. 16]

Context encoders



Pros

- Requires preservation of fine-grained information

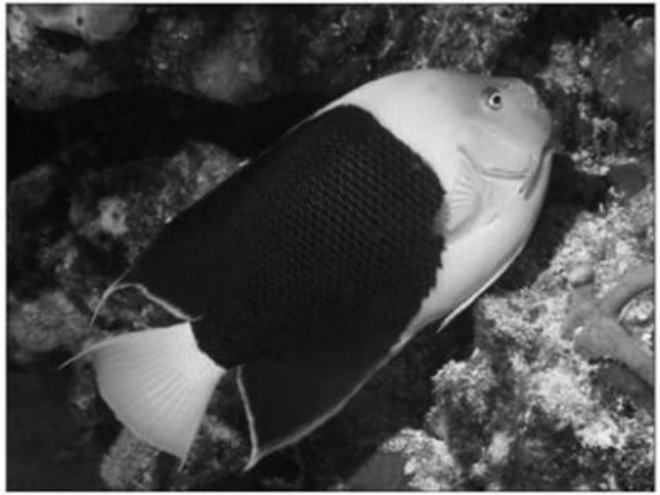
Cons

- Train-eval gap: no masking at eval
- Reconstruction is too hard and ambiguous
- Lots of effort spent on “useless” details: exact colour, good boundary, etc

[“Colorful image colorization”, Zhang et al. 16]

Colorization

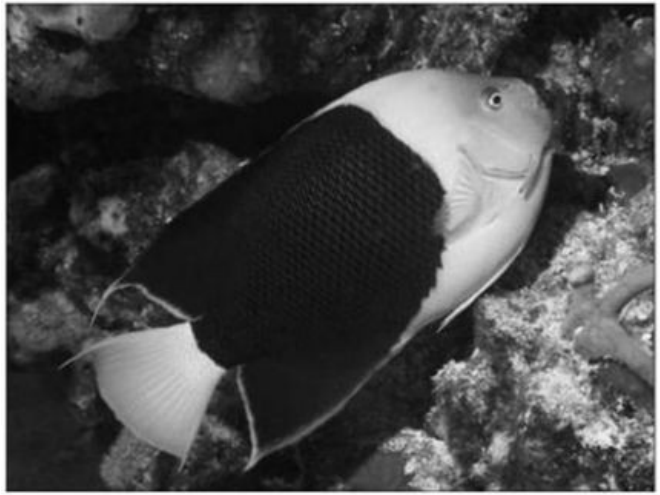
What is the colour of every pixel?



[“Colorful image colorization”, Zhang et al. 16]

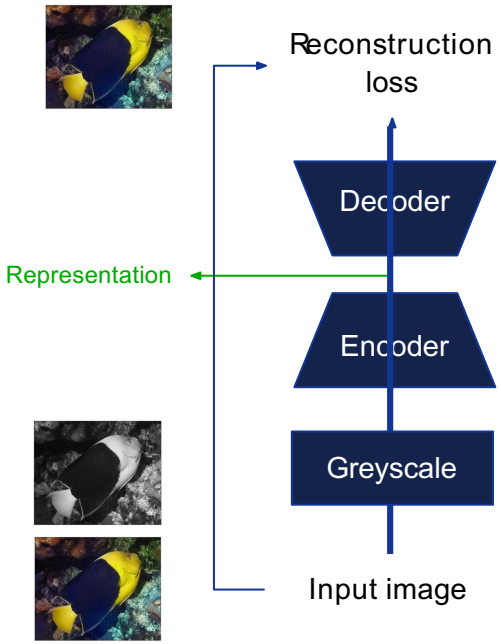
Colorization

What is the colour of every pixel? Hard if you don’t recognize the object!



[“Colorful image colorization”, Zhang et al. 16]

Context encoders



Pros

- Requires preservation of fine-grained information

Cons

- Reconstruction is too hard and ambiguous
- Lots of effort spent on “useless” details: exact colour, good boundary, etc
- Forced to evaluate on greyscale images, losing information

Exploiting time

[“Learning features by watching objects move”, Pathak et al. 16]

Watching objects move

Which pixels will move?

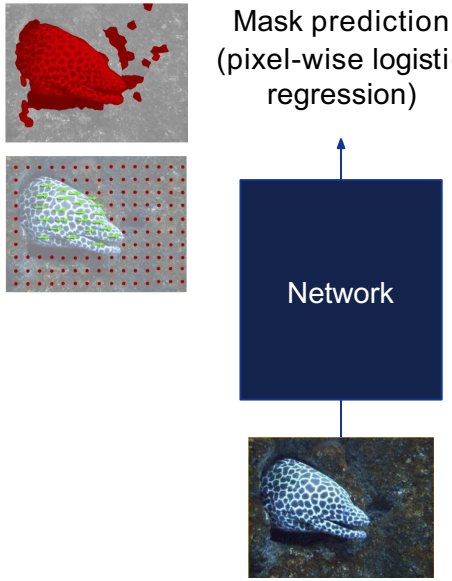
[“Learning features by watching objects move”, Pathak et al. 16]

Watching objects move

Which pixels will move? Easy if we can segment objects!

[“Learning features by watching objects move”, Pathak et al. 16]

Watching objects move



Pros

- Emerging behaviour: segmentation
- No train-eval gap

Cons

- “Blind spots”: stationary objects
- Potential focus on large salient objects
- Depends on an external motion segmentation algorithm
- Cannot be extended to temporal nets (pretext task would be trivial)

[“Tracking emerges by colorizing videos”, Vondrick et al. 18]

Tracking by colorization

Given an earlier frame, colourize the new one.

[“Tracking emerges by colorizing videos”, Vondrick et al. 18]

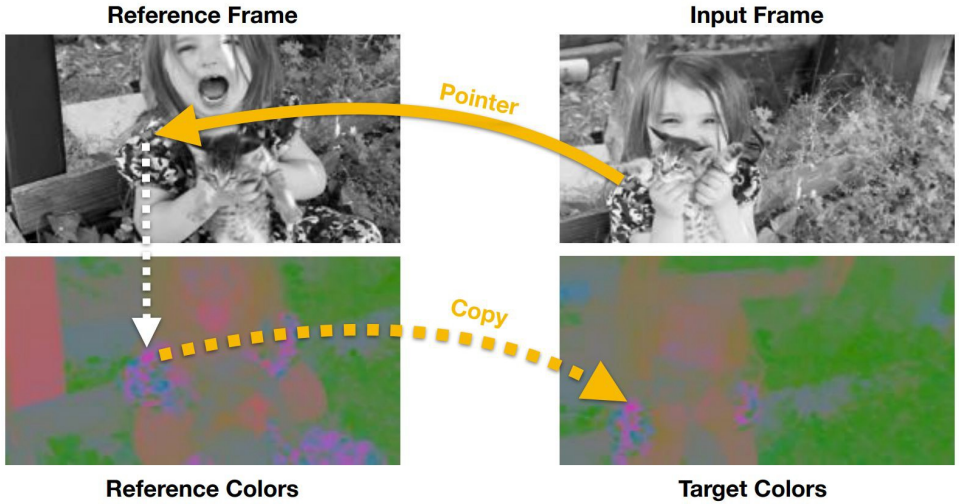
Tracking by colorization

Given an earlier frame, colourize the new one. Easy if everything can be tracked!




[“Tracking emerges by colorizing videos”, Vondrick et al. 18]

Tracking by colorization



Pros

- Emerging behaviour: tracking, matching, optical flow, segmentation

Cons

- Low level cues are effective - less emphasis on semantics
- Forced to evaluate on greyscale frames, losing information

[“Shuffle and learn: Unsupervised learning using temporal order verification”, Misra et al. 16]

Temporal ordering

Is this sequence of frames correctly ordered?

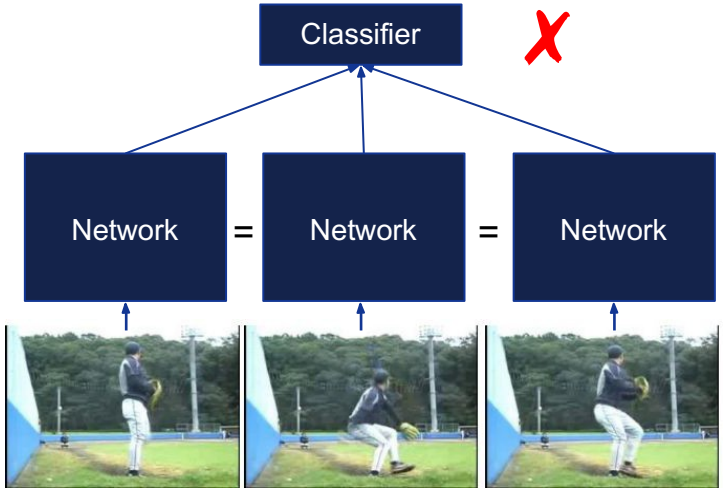
[“Shuffle and learn: Unsupervised learning using temporal order verification”, Misra et al. 16]

Temporal ordering

Is this sequence of frames correctly ordered? Easy if we recognize the action and human pose!

[“Shuffle and learn: Unsupervised learning using temporal order verification”, Misra et al. 16]

Temporal ordering



Pros

- No train-eval gap
- Learns to recognize human pose

Cons

- Mostly focuses on human pose - not always sufficient
- Questionable if it can be extended to temporal nets (potentially task becomes too easy)

Extensions

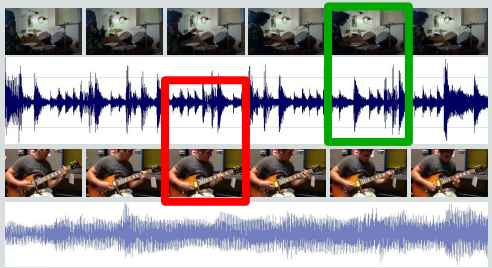
- N frames with one randomly placed - find it

  [“Self-supervised video representation learning with odd-one-out networks”, Fernando et al. 16]

- Ranking loss: embeddings should be similar for frames close in time and dissimilar for far away frames

  [“Time-contrastive networks: Self-supervised learning from video”, Sermanet et al. 17]

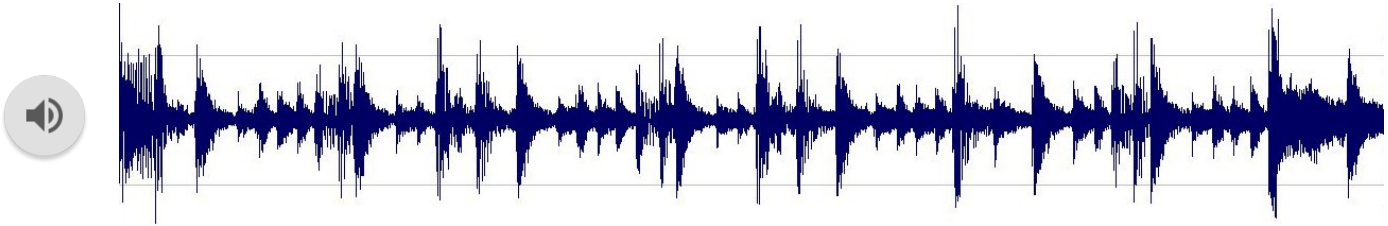
Multimodal



[“Look, Listen and Learn”, Arandjelović et al. 17]

Audio-visual correspondence

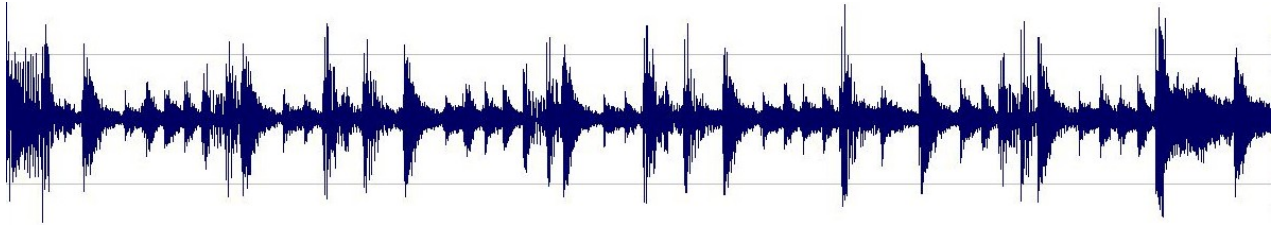
Does the sound go with the image?



[“Look, Listen and Learn”, Arandjelović et al. 17]

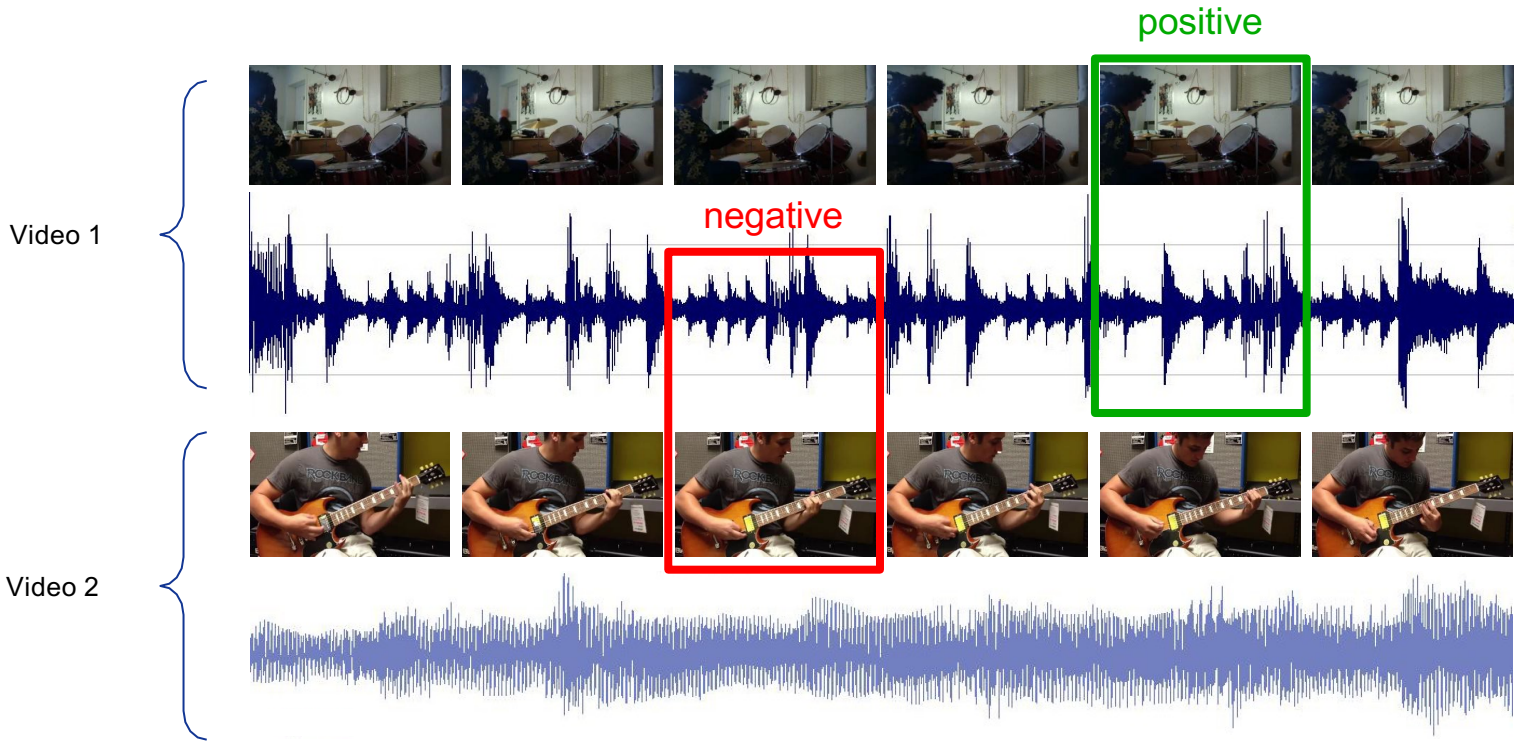
Audio-visual correspondence

Does the sound go with the image? Easy if we recognize what is happening in both the frame and the audio

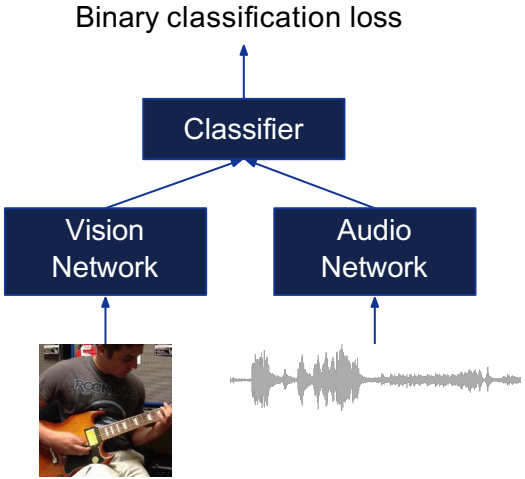


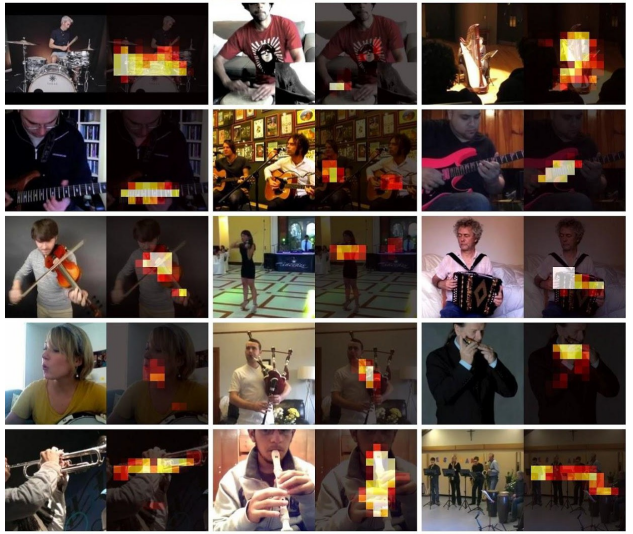
[“Look, Listen and Learn”, Arandjelović et al. 17]

Audio-visual correspondence



Audio-visual correspondence





["Look, Listen and Learn", Arandjelović et al. 17]

["Objects that sound", Arandjelović et al. 18]

Pros

- Natural different views of the training data, no need for augmentations
- No train-eval gap
- Representations in both modalities for free

Cons

- "Blind spots": not everything makes a sound
- Exemplar based: videos of the same class or instance are negatives
- Small output space - two cases (corresponds or not)
  - Can be improved by contrastive approaches

[“End-to-end learning of visual representations from uncurated instructional videos”, Miech et al. 19]

Leveraging narration

Does the narration go with the video?

(Text obtained from automatic speech recognition)

[“End-to-end learning of visual representations from uncurated instructional videos”, Miech et al. 19]

Leveraging narration

Does the narration go with the video? Easy if we recognize what is happening in the video and narrations

(Text obtained from automatic speech recognition)

Complication compared to the audio-visual case:

- Narration and visual content are less aligned

Summary

- Self-Supervised Learning from images/video
  - Enables learning without explicit supervision
  - Learns visual representations – on par with ImageNet training
- Self-Supervised Learning from videos with sound
  - Intra- and cross-modal retrieval
  - Learn to localize sounds
  - Tasks not just a proxy, e.g. synchronization, attention, applicable directly
- Applicable to other domains with paired signals, e.g.
  - face and voice
  - Infrared/visible
  - RGB/D
  - Stereo streams …

Recommend reading list:

1. https://lilianweng.github.io/lil-log/2019/11/10/self-supervised-learning.html
2. Self Supervised Learning: What is Next? ECCV 2020 : https://sslwin.org/