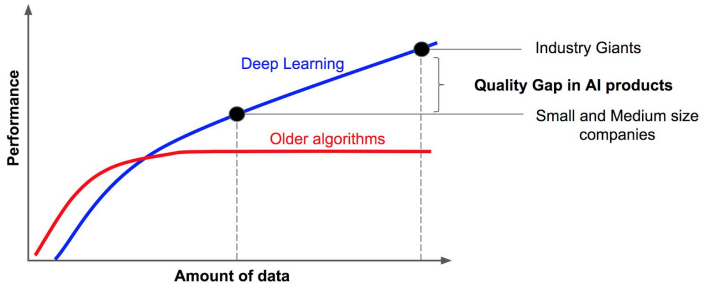
# Learning without fully-labelled data

*Vision for Robotics 2020*

# Outline




## Self-supervision

- Image transformations, colorization, jigsaw puzzle
- Autoencoders
- Contrastive learning

Image-Net Classification challenge


## Weak-supervision

- Multiple instance learning
- Physical models
- Other practical examples

Note: To understand these methods value, imagine yourself as a manual labeller or stakeholder

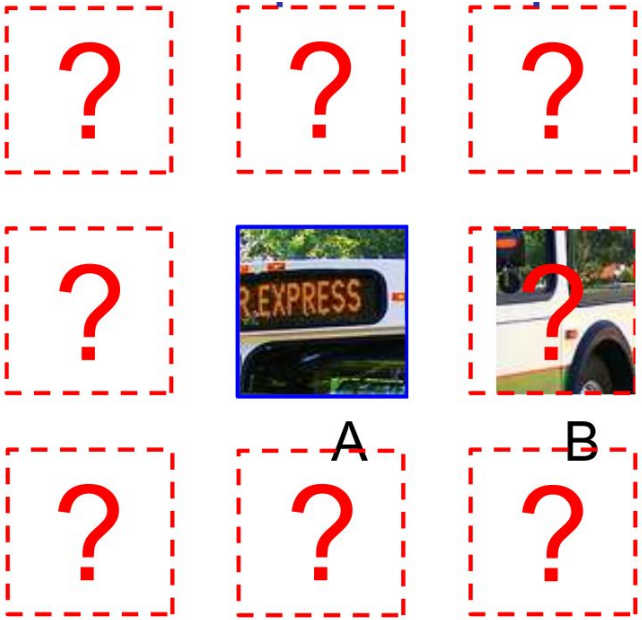
# Fully-supervised labels

- Features from networks trained on ImageNet can be used for other visual tasks, e.g. detection, segmentation, action recognition, fine grained visual classification
- Any visual task can be solved to some extent by:
  - Construct a large-scale dataset labelled for that task
  - Specify a training loss and neural network architecture
  - Train the network and deploy
- Main issue: Time and Financially consuming!
- Alternatives? Self-supervision

# Self-supervision

- Expense of producing a new dataset for each new task
- Some areas are supervision-starved, e.g. medical data, where it is hard to obtain annotation
- Availability of unlabelled images/videos
- A form of unsupervised learning where the **data** provides the supervision
- Create artificial task that network would predict
- 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
- How infants may learn …

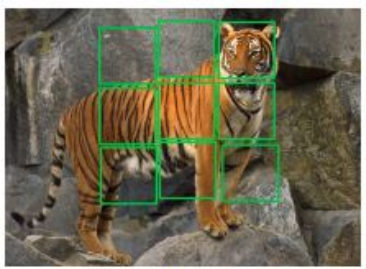
# Relative position - jigsaw puzzle

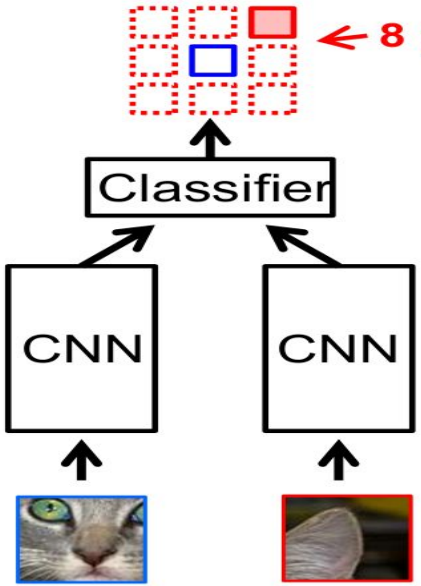


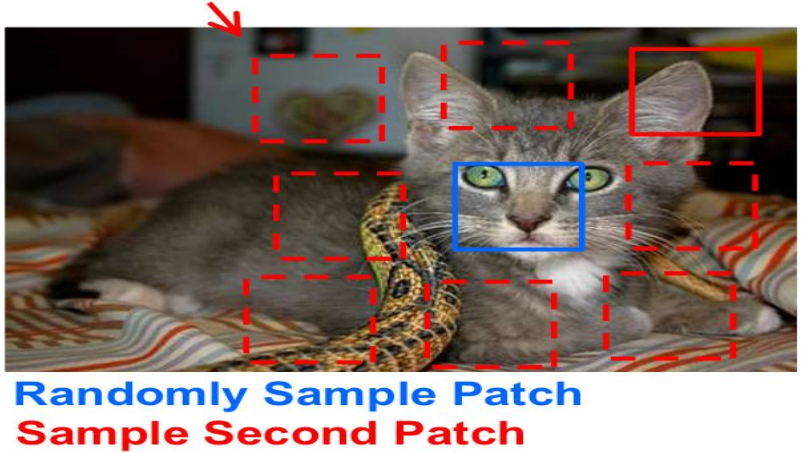

Semantics from non-semantic tasks


Unsupervised visual representation learning by context prediction, Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

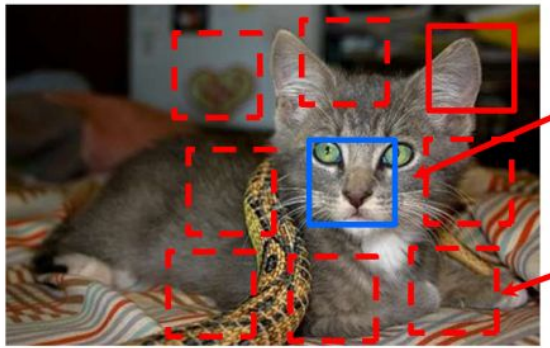
# Relative positioning





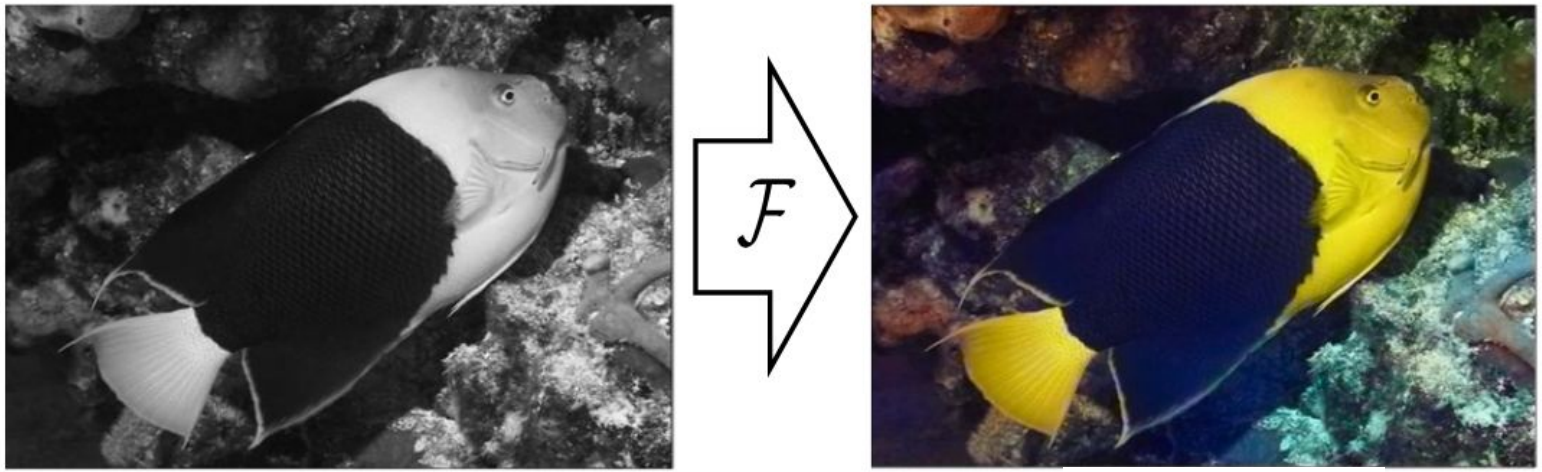
# Avoiding trivial shortcuts



Introduce gap between the patches

Jitter / noise the positions of the patches

# Colourization




Grayscale image: *L* channel

$$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$$

Concatenate (L,ab)

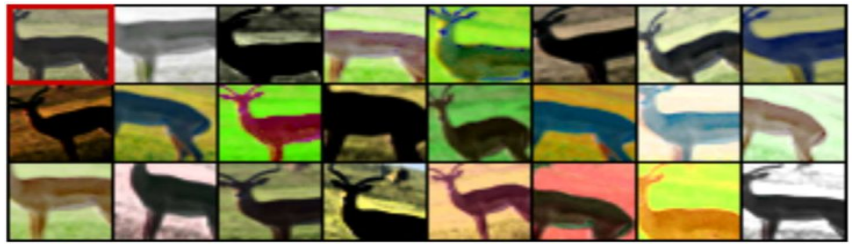
$$(\mathbf{X}, \hat{\mathbf{Y}})$$

# Colourization examples

# Sideways: Usage of Colorization

# Exemplar tasks

- Perturbation or distortion of image patches
- Cropping and affine transformations (torchvision in pytorch)
- Train to classify these exemplars as same class



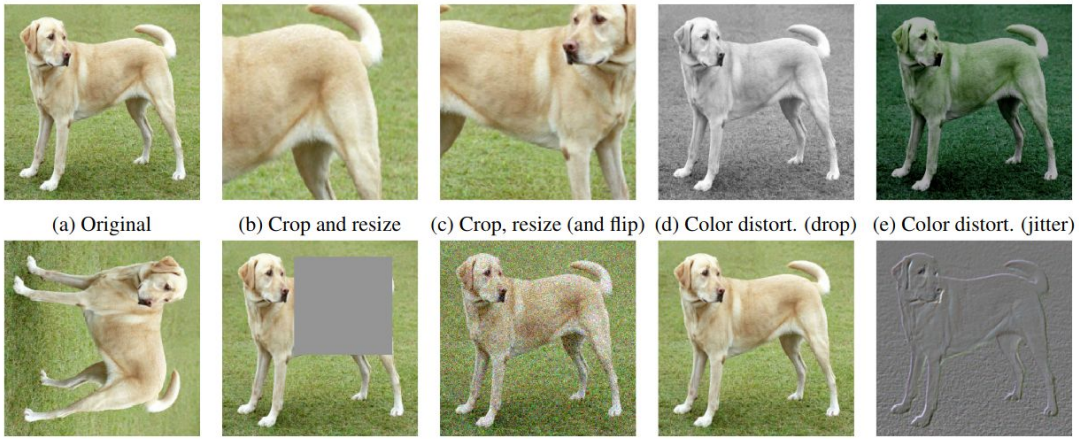
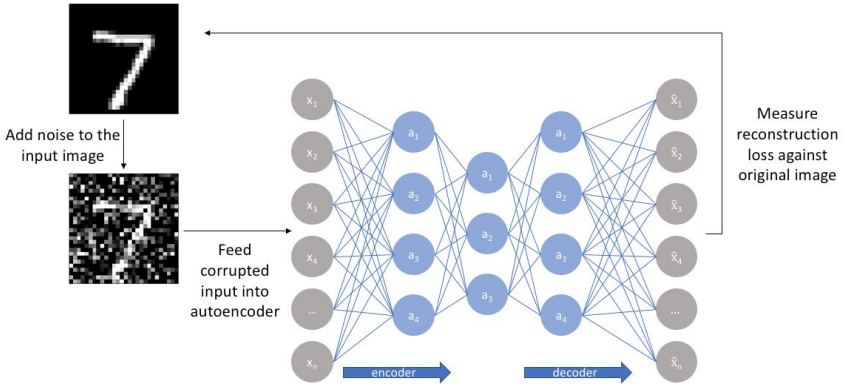


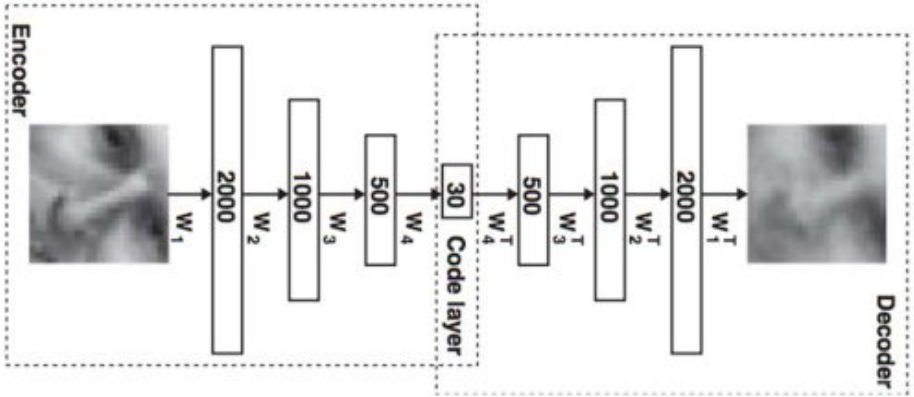

*Figure 4.* Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)

Ting Chen, et. al: A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020

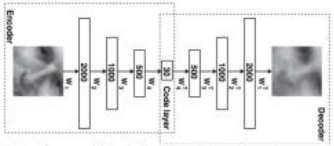
# Autoencoders

- Learn efficient data encoding
- Learn representations for dimensionality reduction and denoising
- Gather useful features from input data
- Variational encoders, generative models ...



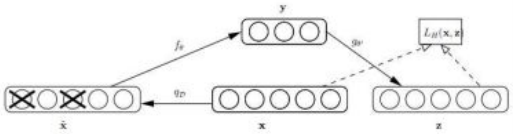



## Autoencoders



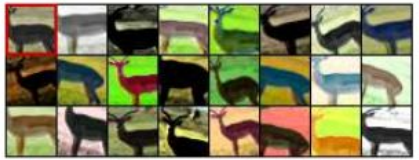

Hinton & Salakhutdinov. Science 2006.

## Denoising Autoencoders



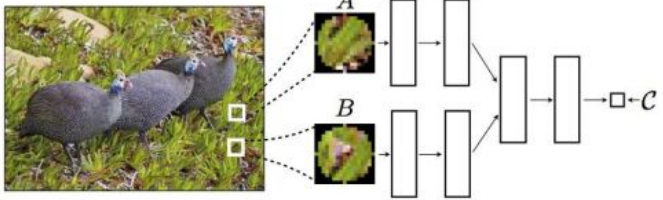
Vincent *et al.* ICML 2008.

## Exemplar networks



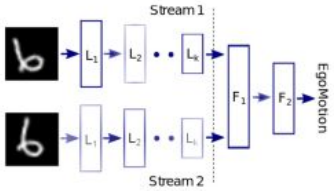
Dosovitskiy *et al.*, NIPS 2014

## Co-Occurrence



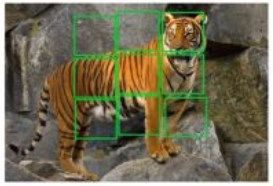

Isola *et al.* ICLR Workshop 2016.

## Egomotion




Agrawal *et al.* ICCV 2015 Jayaraman *et al.* ICCV 2015

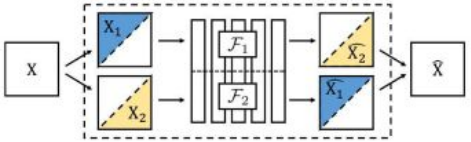
## Context



Noroozi et al 2016

Pathak *et al.* CVPR 2016

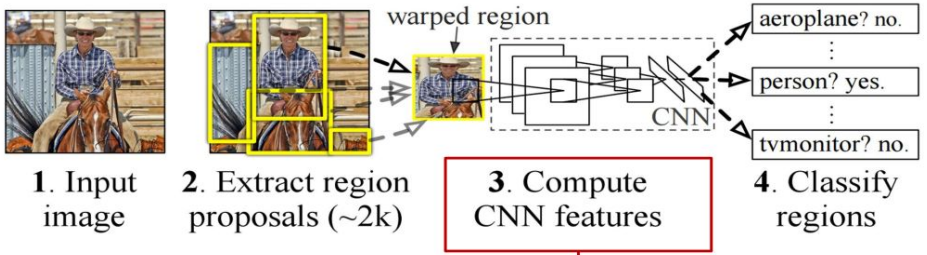
## Split-brain auto-encoders




Zhang *et al.* CVPR 2017

# Multi-task learning

- On ImageNet dataset we train network self-supervised, then fix parameters and learn classifier on extracted features
- On PASCAL dataset we train net with self-supervision and then train faster-RCNN
- ImageNet labels == fully supervised
- Everything on same backbone network ResNet-101



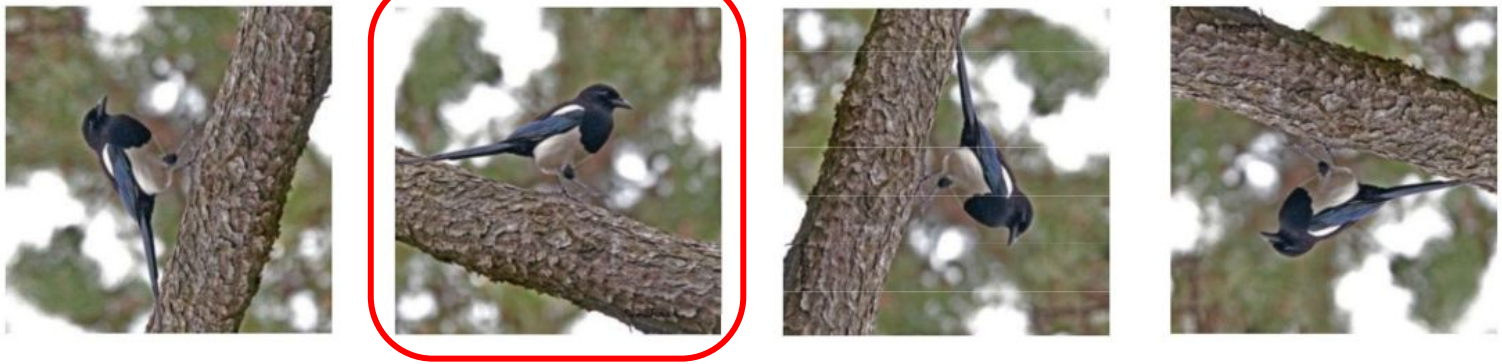

Pre-train on relative-position task, w/o labels

| Self-supervision task | ImageNet Classification top-5 accuracy | PASCAL VOC Detection mAP |
|---|---|---|
| Rel. Pos | 59.21 | 66.75 |
| Colour | 62.48 | 65.47 |
| Exemplar | 53.08 | 60.94 |
| Rel. Pos + colour | 66.64 | 68.75 |
| Rel. Pos + Exemplar | 65.24 | 69.44 |
| Rel. Pos + colour + Exemplar | 68.65 | 69.48 |
| ImageNet labels | 85.10 | 74.17 |

Multi-task self-supervised visual learning, C Doersch, A Zisserman, ICCV 2017

# Image Transformations

- Which image has a correct rotation?



Unsupervised representation learning by predicting image rotations, Spyros Gidaris, Praveer Singh, Nikos Komodakis, ICLR 2018

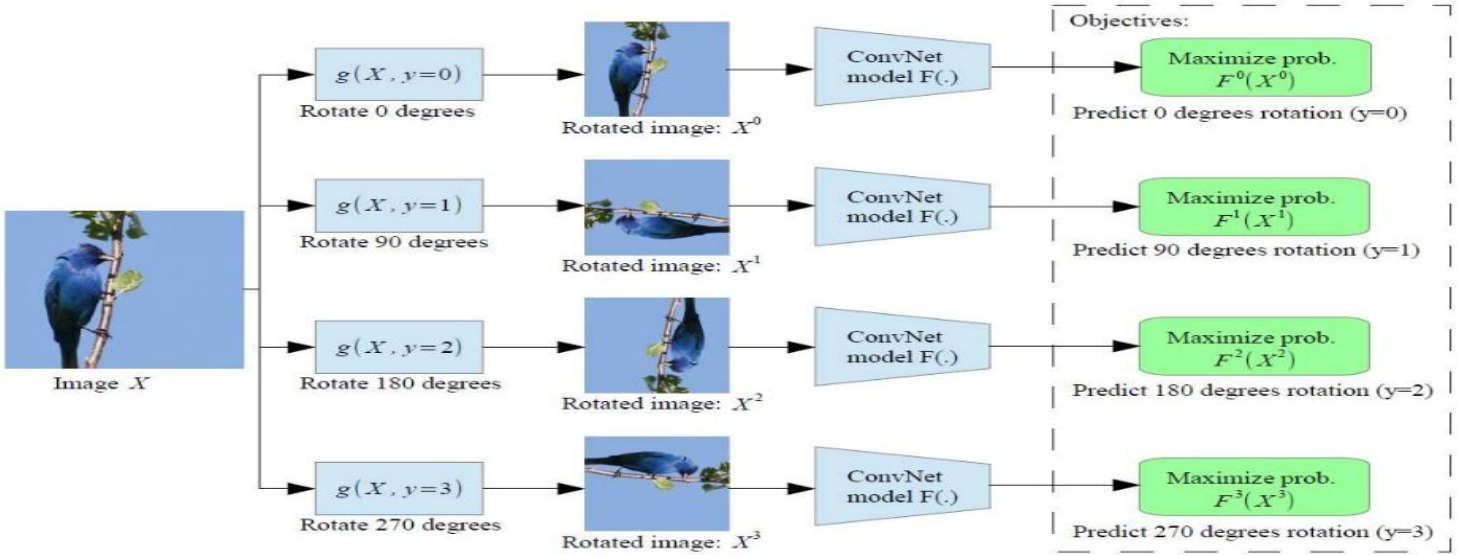
# Image transformations


Figure 1: Images rotated by random multiples of 90 degrees (e.g., 0, 90, 180, or 270 degrees). The core intuition of our self-supervised feature learning approach is that if someone is not aware of the concepts of the objects depicted in the images, he cannot recognize the rotation that was applied to them.

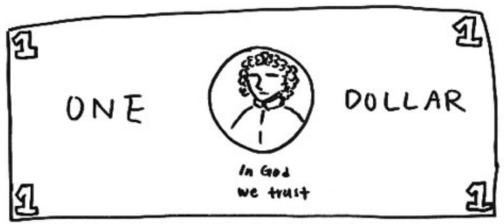
Unsupervised representation learning by predicting image rotations, Spyros Gidaris, Praveer Singh, Nikos Komodakis, ICLR 2018

# Learning the rotations

| | PASCAL VOC Detection mAP |
|---|---|
| Random | 43.4 |
| Rel. Pos. | 51.1 |
| Colour | 46.9 |
| Rotation | 54.4 |
| ImageNet Labels | 56.8 |



# Contrastive learning

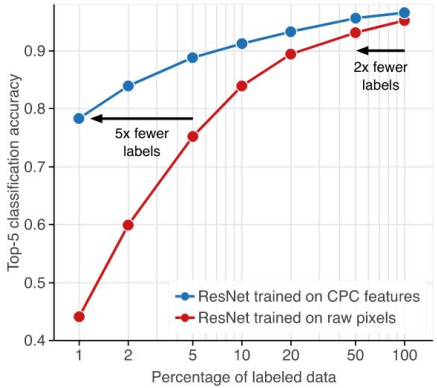




*Epstein (2016)*

- “Despite having seen a dollar bill countless number of times, we don’t retain a full representation of it.”
- We really only retain enough features of the bill to distinguish it from other objects.
- Can we build representation learning algorithms that don’t concentrate on pixel-level details, and only encode high-level features sufficient enough to distinguish different objects?

# Generative vs. contrastive




*Henaff aff et al., 2019*

## Generative / Predictive


Loss measured in the output space

Examples: Colorization, Auto-Encoders

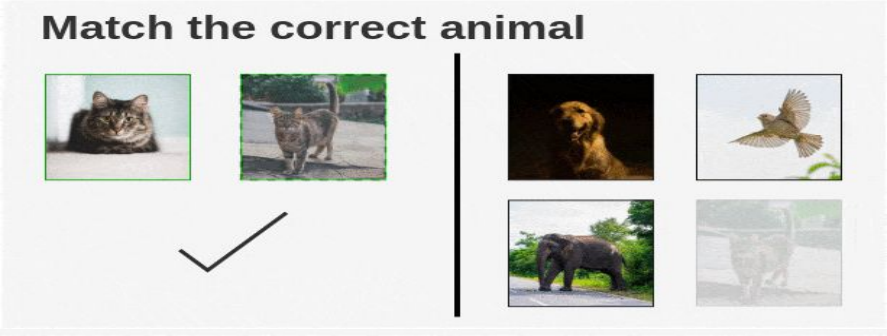
## Contrastive


Loss measured in the representation space

Examples: TCN, CPC, Deep-InfoMax

- Contrastive methods learn representations by contrasting positive and negative examples
- Pixel-level losses can lead to focus on pixel-based details, rather than latent factors.
- Pixel-based objectives often assume independence between each pixel, thereby reducing their ability to model correlations or complex structure

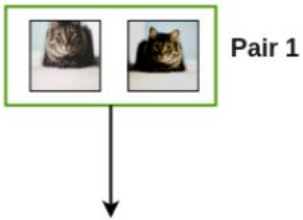
# Contrastive




- Contrastive objective causes representations of corresponding views to “attract” each other, while representations of non-corresponding views “repel” each other.

Procedure:

- First, generate batches of a certain size, say N from the raw images
- For each images, a random transformation / crop function is applied to get a pair
- Each augmented image in a pair is passed through an encoder to get image representations.
- For each augmented image in the batch, get an embedding vector




$$\text{score}(f(x), f(x^+)) >> \text{score}(f(x), f(x^-))$$

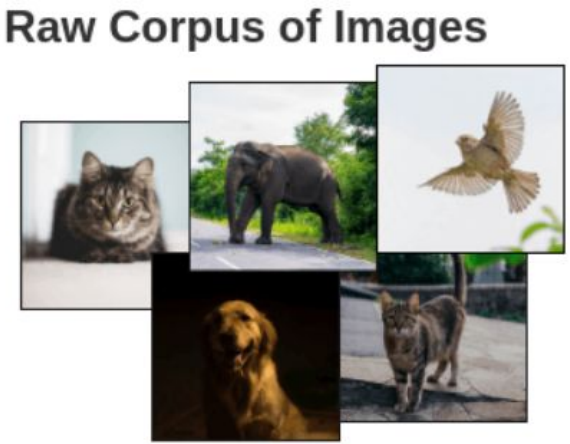
- x+ is a data point similar to x (from transformation) ... a positive sample
- x− is a data point dissimilar to x (not part of original), ... a negative sample

Softmax = $\frac{e^{\text{similarity}(\,,\,)}}{e^{\text{similarity}(\,,\,)} + e^{\text{similarity}(\,,\,)} + e^{\text{similarity}(\,,\,)}}$

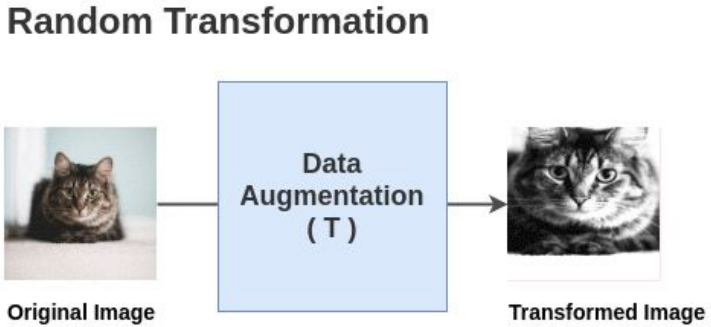
# Contrastive learning example

*T.Chen; SimCLR. 2020*

Gather images



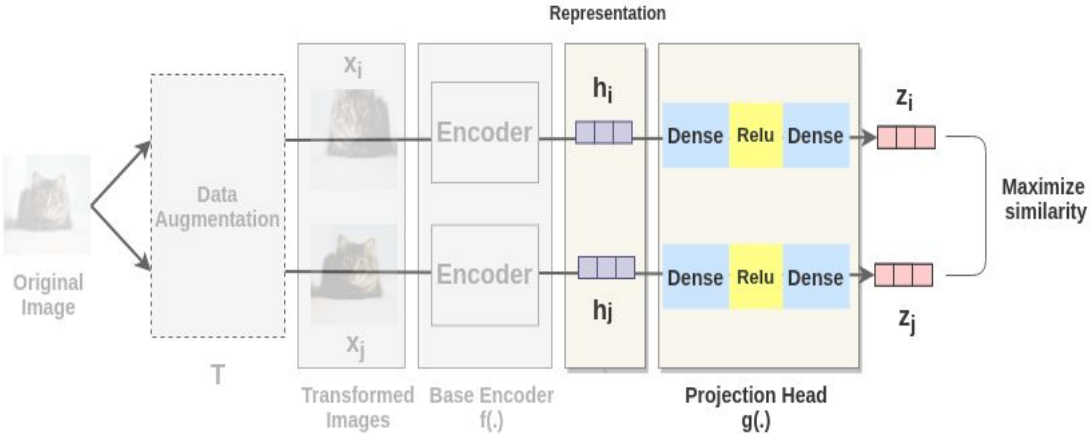

Apply transformations (Crop, color jitter, rotate, translate)

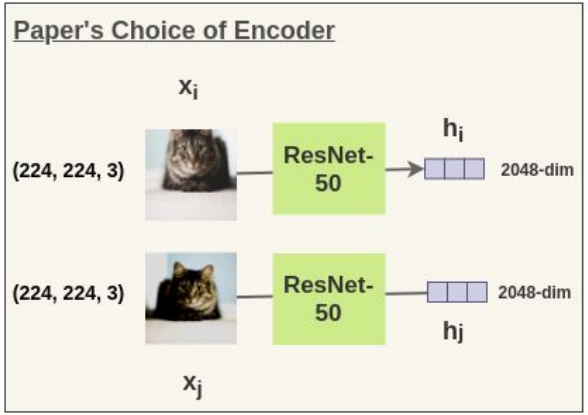


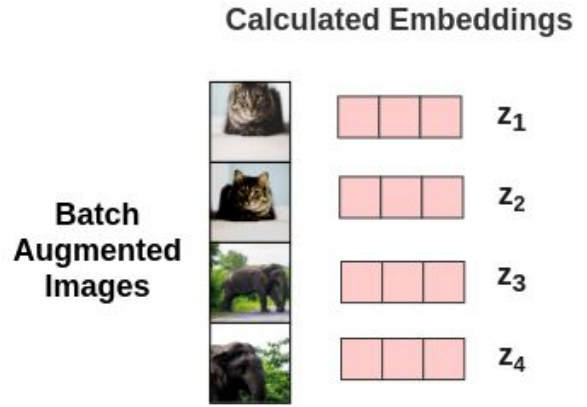

https://amitness.com/2020/03/illustrated-simclr/

# Contrastive learning example

- Used encoder ResNet-50 (shared weights)



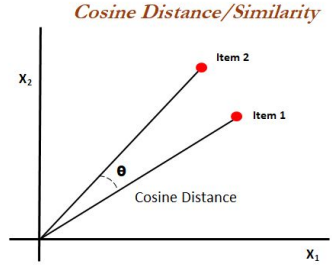



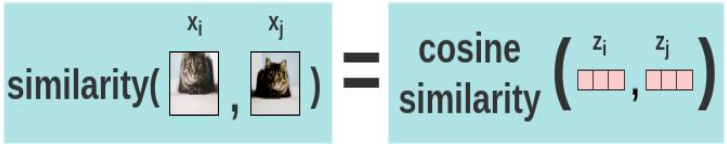




https://amitness.com/2020/03/illustrated-simclr/

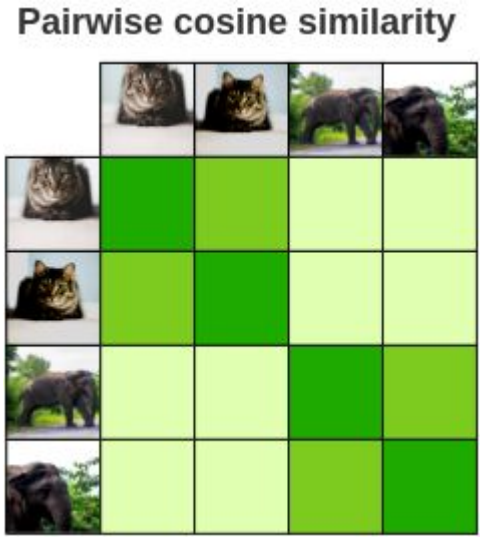
# Contrastive learning example - loss

- Cosine similarity



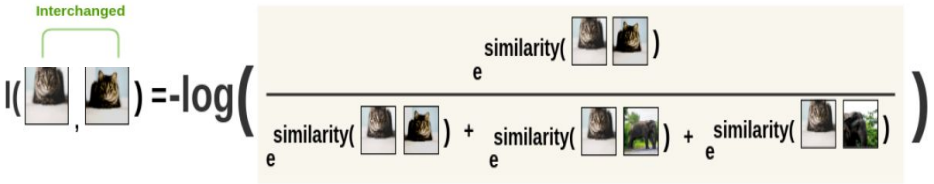




$$s_{i,j} = \frac{z_i^T z_j}{(\tau ||z_i|| ||z_j||)}$$

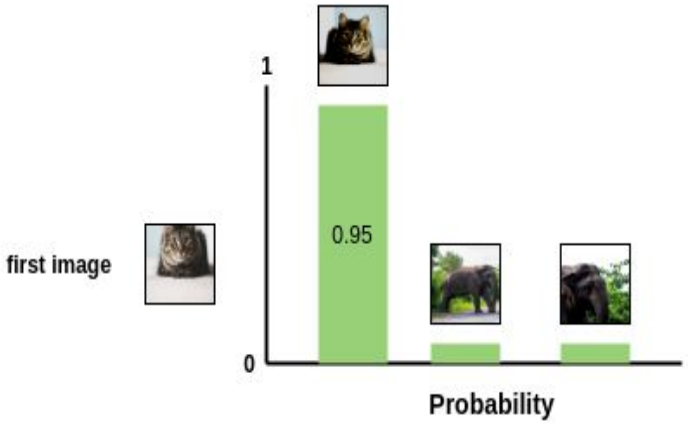


# Contrastive learning example - loss

- Cross-entropy loss

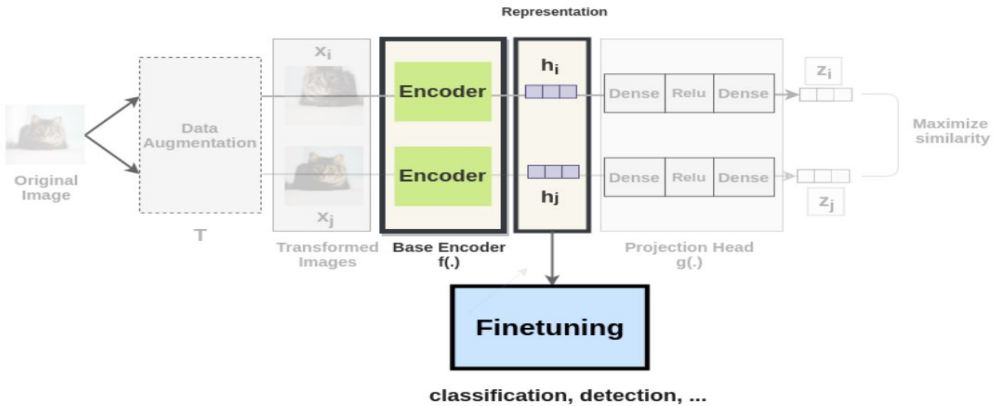



$$l(i,j) = -log\frac{exp(s_{i,j})}{\sum_{k=1}^{2N} l_{[k!=i]} exp(s_{i,k})}$$



# Contrastive learning example - Classification

- Once the SimCLR model is trained on the contrastive learning task, it can be used for transfer learning.
- The representations from the encoder are used, not from projection head



# Weak-supervision

- Not fully descriptive, noised, limited labels provided
- Insufficient datasets
- Inexpensive way to learn
- Multiple instance learning
- Other knowledge about problem (for example Physics constraints, heuristics, demonstrations)

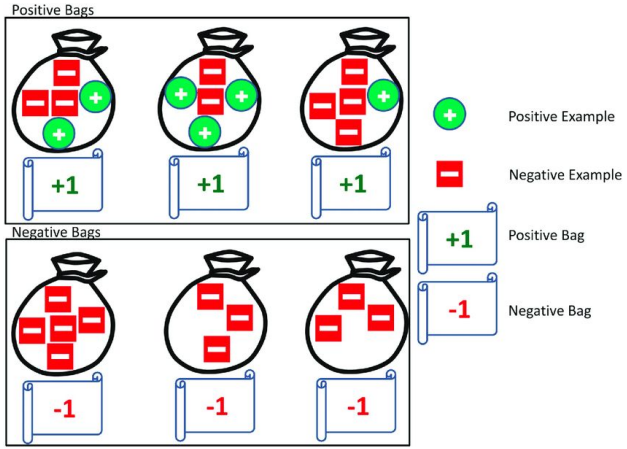
# Multiple Instance learning

Its in the bag!

- Training instances are arranged in sets, called **bags**.
- A label is provided for **entire bags but not for instances**.

What it is not:

- Fully-supervised learning
- Self-supervised learning



# Illustration of MIL problem

Can enter the secret room

Can not enter the secret room

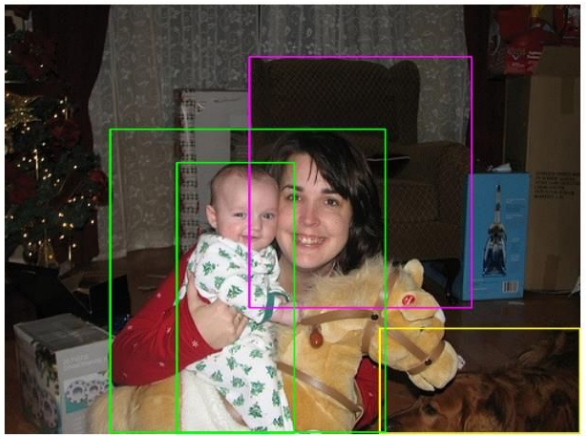
Can I the secret room???

What is the magic key???

# Why Multiple instance learning

It has been proposed because:

- Some problems are naturally formulated as MIL
- It deals with weakly annotated data.
- This reduces the annotation cost.
- Algorithms can now learn from a greater quantity of training data.



# Definition of the standard MIL assumption

- Training instances are arranged in sets generally called bags.
- A label is given to bags but not to individual instances.
- Negative bags do not contain positive instances.
- Positive bags contain at least one or specific combination of positive instances.
- Positive and negative bags can differ by their instance distributions

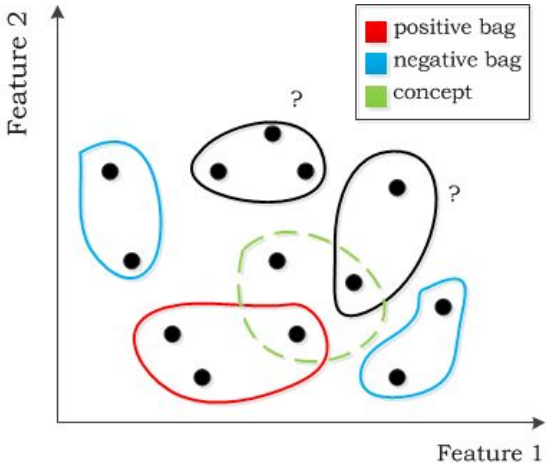



Image from : http://www.miproblems.org/mi-learning/

# Example of MIL

- Bag: Image with beach
- Instance: Sand, water
- Classify beach images
- Both sand and water segments are positive instances for beach pictures.
- However, picture of beach must contain both segments of sand and water. Otherwise, they can be pictures of desert or sea.

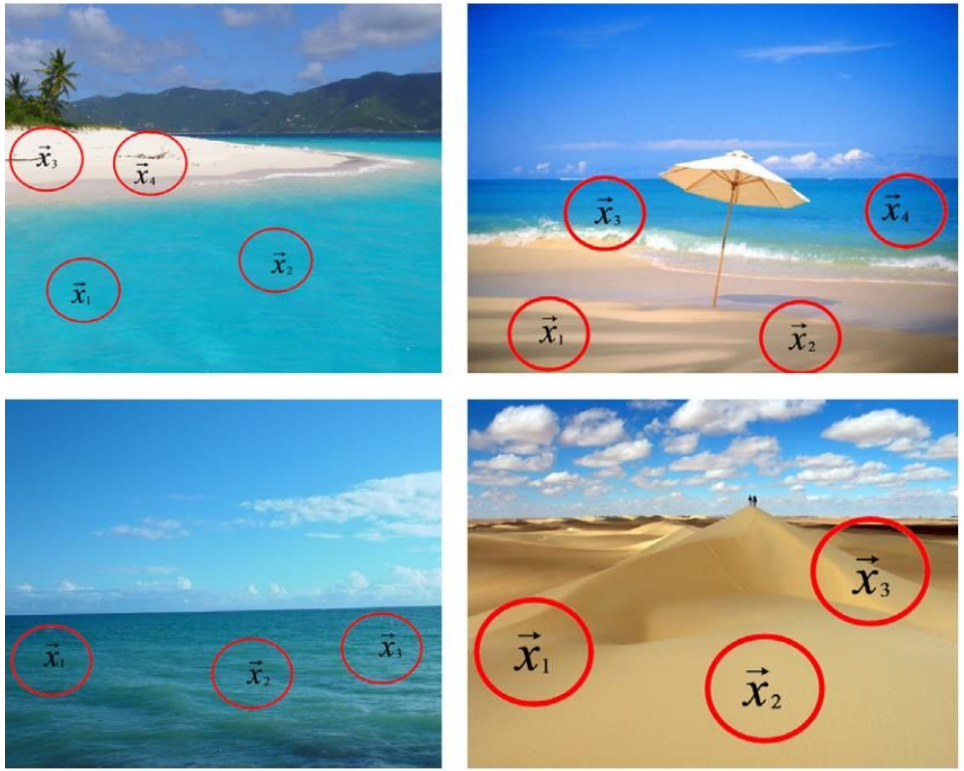



Image from : J. Amores, “Multiple instance classification: Review, taxonomy and comparative study,” Artif. Intell., vol. 201, pp. 81–105, Aug. 2013.

# Tasks that can be performed in MIL

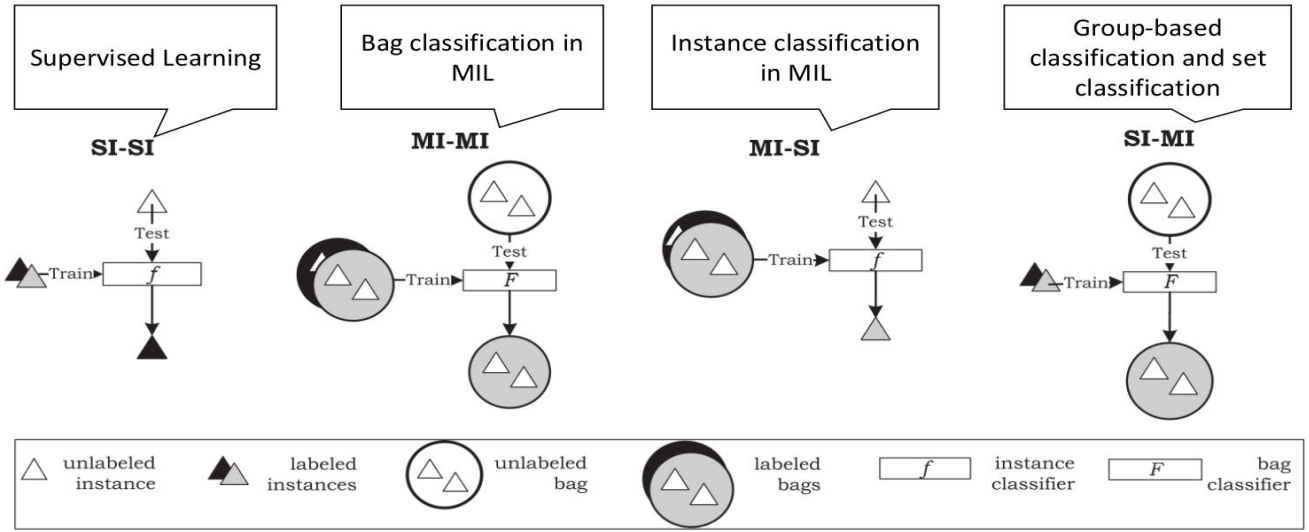
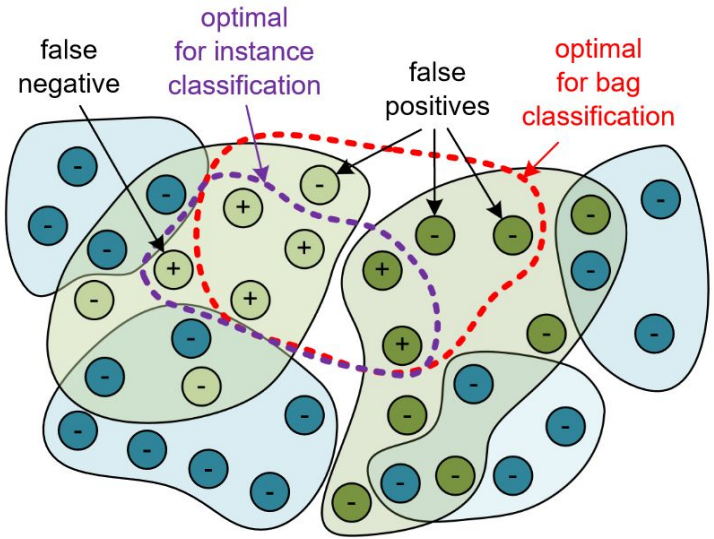



Image from: V. Cheplygina, D. M. J. Tax, and M. Loog, “On classification with bags, groups and sets,” Pattern Recognition Letters, vol. 59, pp. 11–17, Jul. 2015.

# Difference against instances in optimality

- Instance and bag classification are two different tasks.
- It has been observed by many authors that the best algorithm for instance classification is rarely the best for bag classification.
- “The key difference is the instance misclassifying cost.”

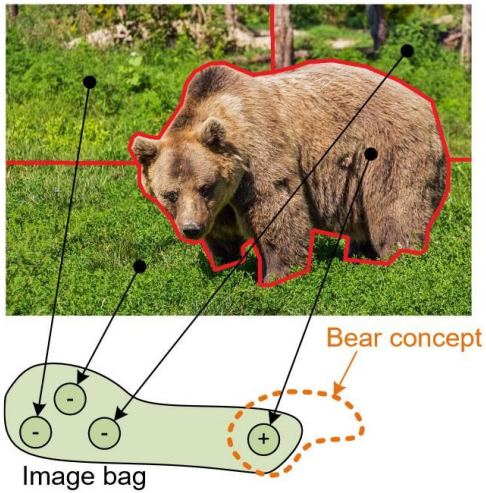


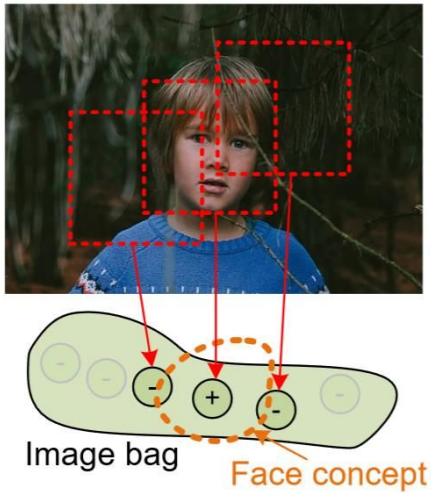

G. Vanwinckelen, V. do O, D. Fierens, and H. Blockeel, “Instance-level accuracy versus bag-level accuracy in multi-instance learning,” Data Mining Knowledge Discovery, 2015

Image from: M.-A. Carbonneau, E. Granger, and G. Gagnon, “Decision Threshold Adjustment Strategies for Increased Accuracy in Multiple Instance Learning,” in Proc. The 6th International Conference on Image Processing Theory, Tools and Applications (IPTA), 2016.

# Bag Composition

- Depending on the applications, bags can differ in:
  - The proportion of positive instances in positive bags
  - The size of the bags.
  - Instance Co-occurrences
  - Intra-bags similarities

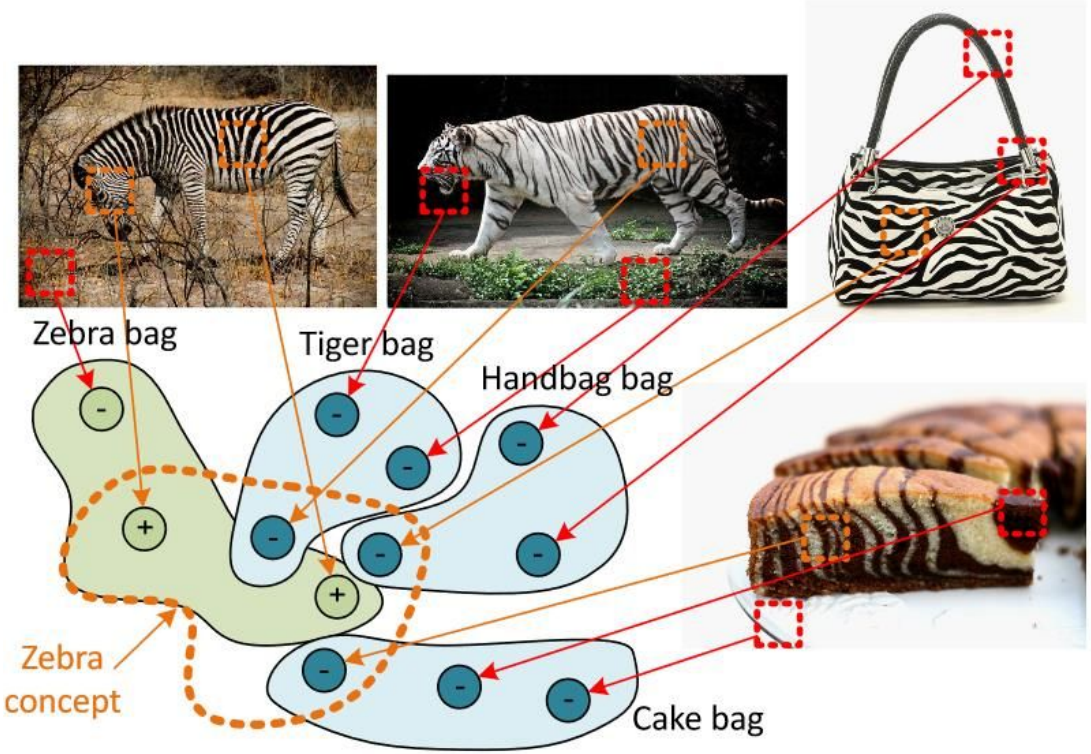



Images from: M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, “Multiple Instance Learning: A Survey on Problems Characteristics and Applications,” to be submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

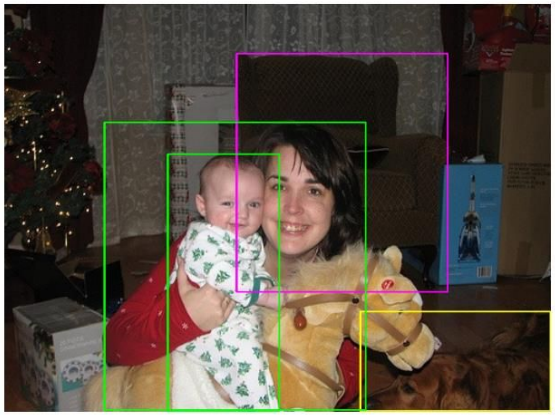
# Label Ambiguity

- Weak supervision implies label ambiguity
  - Noise.
  - Lack of clear classes at instance level.
  - Ambiguous representation.
  - Classes can share the same type of instances.



# Examples: Object localization

- Objective: Find objects in images.
- Bag: Collection of candidate annotation boxes
- Instance: Sub-image corresponding to candidate windows.
- Justification: A large quantity of data can be used to learn because costly strong annotations are not necessary.



H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell, “On learning to localize objects with minimal supervision,” International Conference on Machine Learning, 2014

# Examples: Computer diagnosis

- Objective: Predict if a subject is diseased or healthy.
- Bags: Collection segments or patches extracted from a medical image.
- Instances: Image segments or patches.
- Justification: A large quantity of images can be used to train. Only a diagnosis is required per image. Expert local annotation are no longer required.

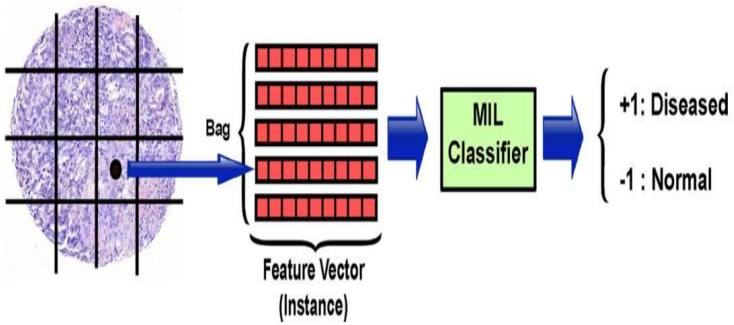



Image from: M. Kandemir and F. A. Hamprecht, “Computer-aided diagnosis from weak supervision: a benchmarking study.,” Comput. Med. Imaging Graph., vol. 42, pp. 44–50, Jun. 2015.

# Example: Sentiment Analysis in Text (or any other Text analysis …)

- Objective: Predict if a text/sentence expresses positive or negative sentiment.
- Bags: Texts/paragraphs.
- Instances: Sentences.
- Justification: Large quantity of text can be harvested from the web. A sentiment is usually given to a complete text while it may contain positive and negative sentences/words.

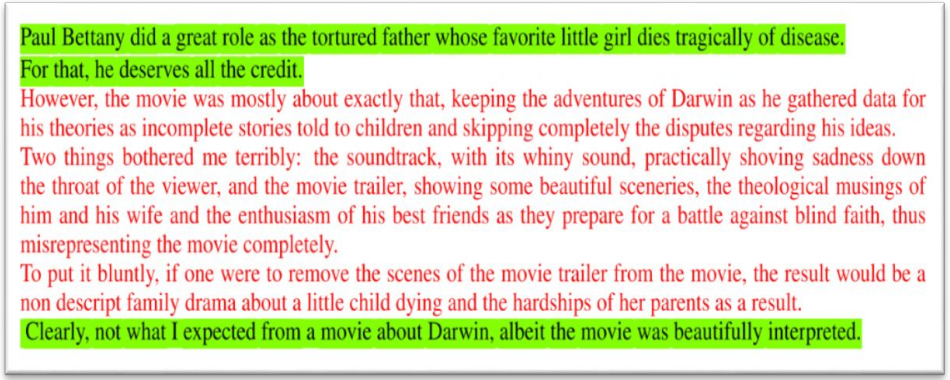

Paul Bettany did a great role as the tortured father whose favorite little girl dies tragically of disease.
For that, he deserves all the credit.
However, the movie was mostly about exactly that, keeping the adventures of Darwin as he gathered data for his theories as incomplete stories told to children and skipping completely the disputes regarding his ideas.
Two things bothered me terribly: the soundtrack, with its whiny sound, practically shoving sadness down the throat of the viewer, and the movie trailer, showing some beautiful sceneries, the theological musings of him and his wife and the enthusiasm of his best friends as they prepare for a battle against blind faith, thus misrepresenting the movie completely.
To put it bluntly, if one were to remove the scenes of the movie trailer from the movie, the result would be a non descript family drama about a little child dying and the hardships of her parents as a result.
Clearly, not what I expected from a movie about Darwin, albeit the movie was beautifully interpreted.

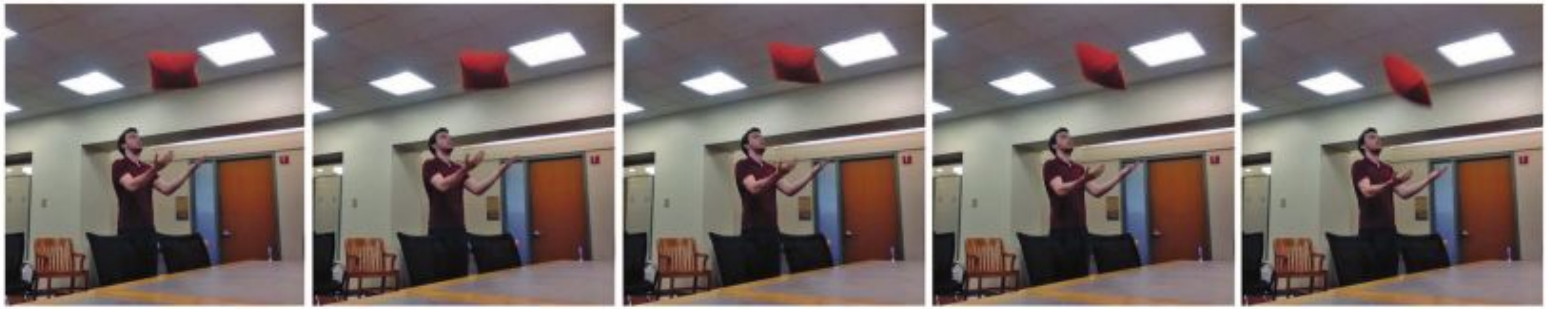
Image from: D. Kotzias, M. Denil, P. Blunsom, and N. de Freitas, “Deep Multi-Instance Transfer Learning,” CoRR, vol. abs/1411.3, 2014.

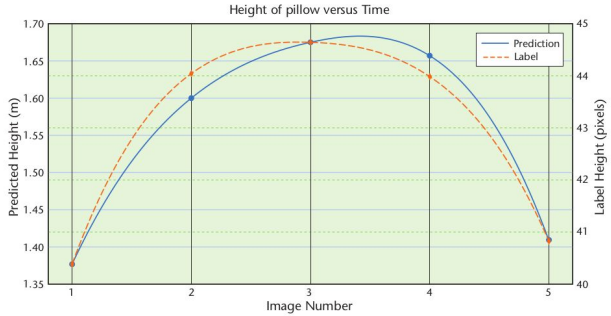
# Cases for MIL

- Data points are grouped in sets
- Weak supervision is provided
- Problems are naturally formulated as MIL.
- Strong supervision is costly to obtain or a large quantity of weakly labeled data can be leveraged.

Hongyu Ren,Learning with Weak Supervision from Physics and Data-Driven Constraints, AAAI 2018

# Physical constraints



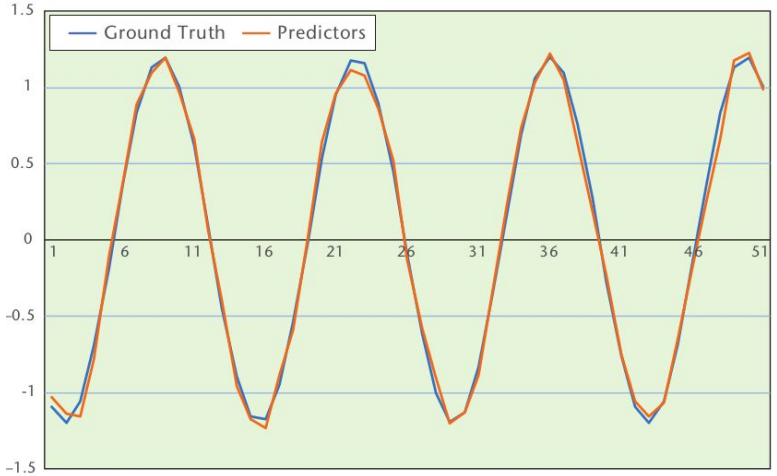



$$\mathbf{y}_i = y_0 + v_0(i\Delta t) + a(i\Delta t)^2$$

- This equation provides a necessary constraint, which the correct mapping must satisfy.
- Minimize difference between constraint and prediction (fit parabola)

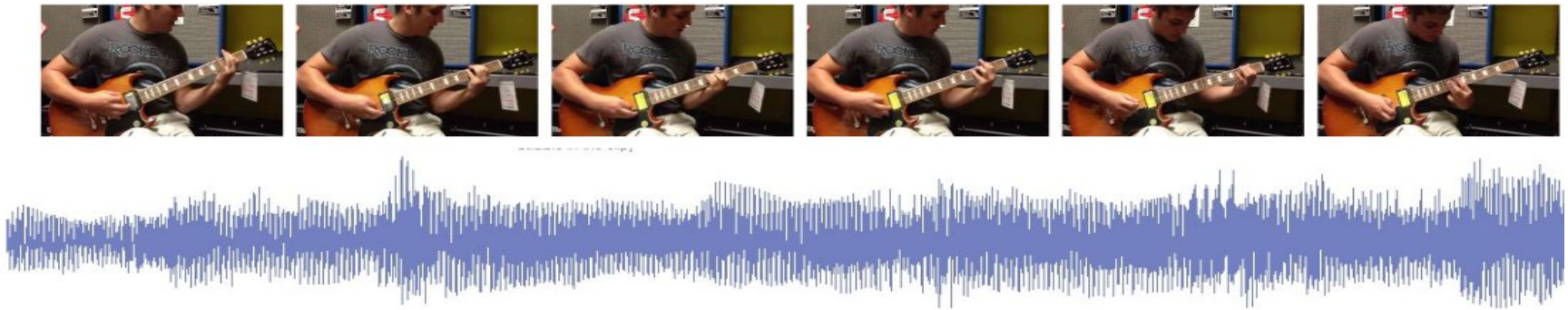
# Physical constraints

- The network is trained to predict angles that cannot be distinguished from the simulated dynamics, encouraging it to track the metal ball over time.




# Audio-Visual Correspondence

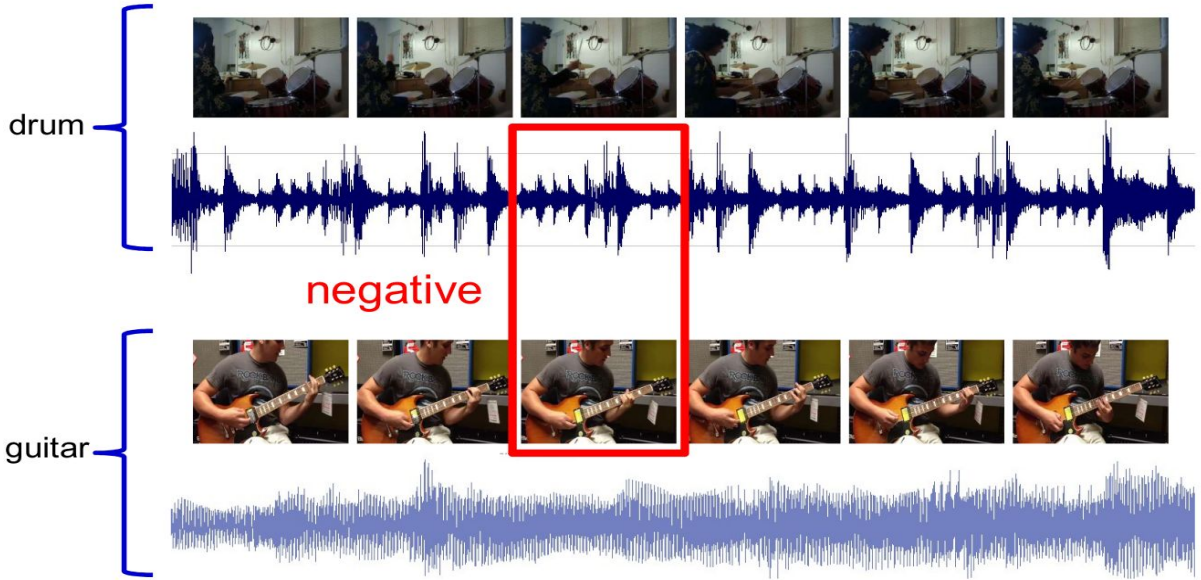
- What can be learnt by looking at and listening to a large number of unlabelled videos?
  - the networks are able to learn useful semantic concepts
  - the two modalities can be used to search one another
  - the object making the sound can be localised.



https://deepmind.com/blog/article/objects-that-sound

# Audio-Visual Correspondence

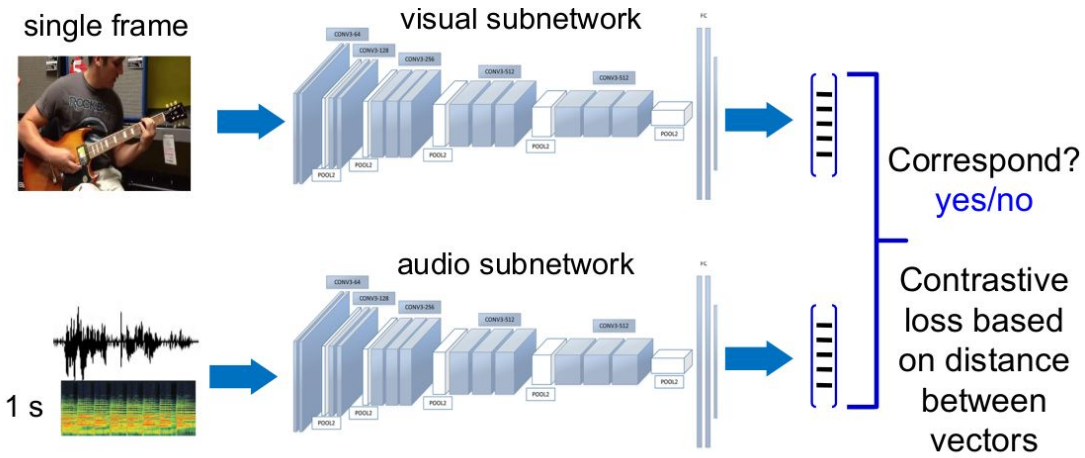
- Two types of proxy task
  - Predict audio-visual correspondence
  - Predict audio-visual synchronization
- No Classification labels
- “Guitar” naturally emerges in both modalities.



# Audio-Visual Embedding (AVE-Net)

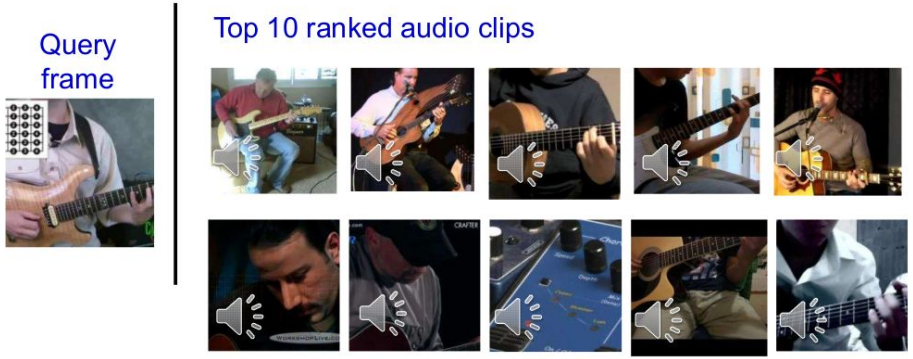
- The only way for a system to solve this binary classification task is by learning to detect various semantic concepts in both the visual and the audio domain

- Distance between audio and video vectors
  - Small ... Positive
  - Large ... Negative

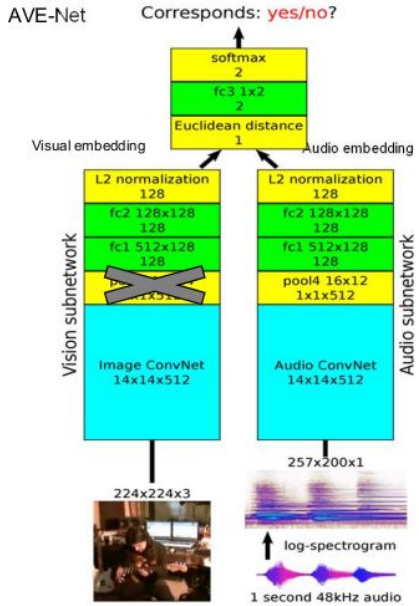


# Cross-modal retrieval

- Since the correspondence score is computed purely based on the distance, the two embeddings are forced to be aligned (i.e. the vectors live in the same space, and so can be compared meaningfully), thus facilitating cross-modal retrieval:



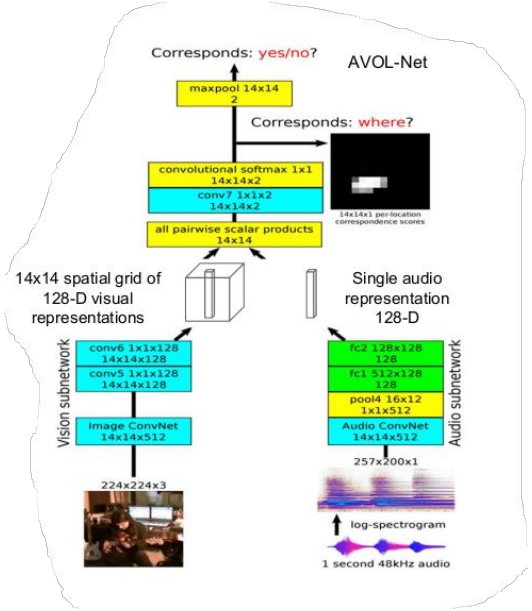
# Can we use it to localize object in the image?




Apply Visual ConvNet convolutionally

Multiple instance learning

[VIDEO]



# Objects that sound - ImageNet classification

- Evaluation procedure for self-supervised setting
  - Use method to extract features
  - Linear classification learned on ImageNet
- On par with state-of-the-art methods
- The only method that never seen ImageNet images
  - Probably did not see image with “Tibetian Terrier”
  - Video frames have different quality than images

| Method | Top 1 accuracy |
|---|---|
| Random | 18.3% |
| Pathak *et al.* [21] | 22.3% |
| Krähenbühl *et al.* [14] | 24.5% |
| Donahue *et al.* [7] | 31.0% |
| Doersch *et al.* [6] | 31.7% |
| Zhang *et al.* [34] (init: [14]) | 32.6% |
| Noroozi and Favaro [18] | 34.7% |
| Ours random | 12.9% |
| Ours | 32.3% |

# Simulators ---> VLC