Digest

Self-supervised learning and Pseudo-labelling

Material sourced from 4F12 (Computer Vision) 2021 lecture series

Digest by Samuel Albanie, April 2022

Outline

- Self-supervised learning
- Pseudo-labelling

Self-supervised Learning - Motivation

Motivation - the state of the (machine perception) nation

Reasons to be cheerful

Deep learning has achieved remarkable progress through the supervised learning paradigm:

- Gather a large collection of data and manually annotate it
- Supervise a model with the resulting (data, annotation) pairs.

Major gains on vision benchmarks!

Cause for concern

Despite these successes, we still seem to have a long way to go:

- Even the *highest capacity* models trained on the *largest annotated datasets* continue to make "*silly*" mistakes
- It seems we can never get enough labelled data to get close to the human perception system

Question: *Can we take inspiration from the early stages of development of human perception?*

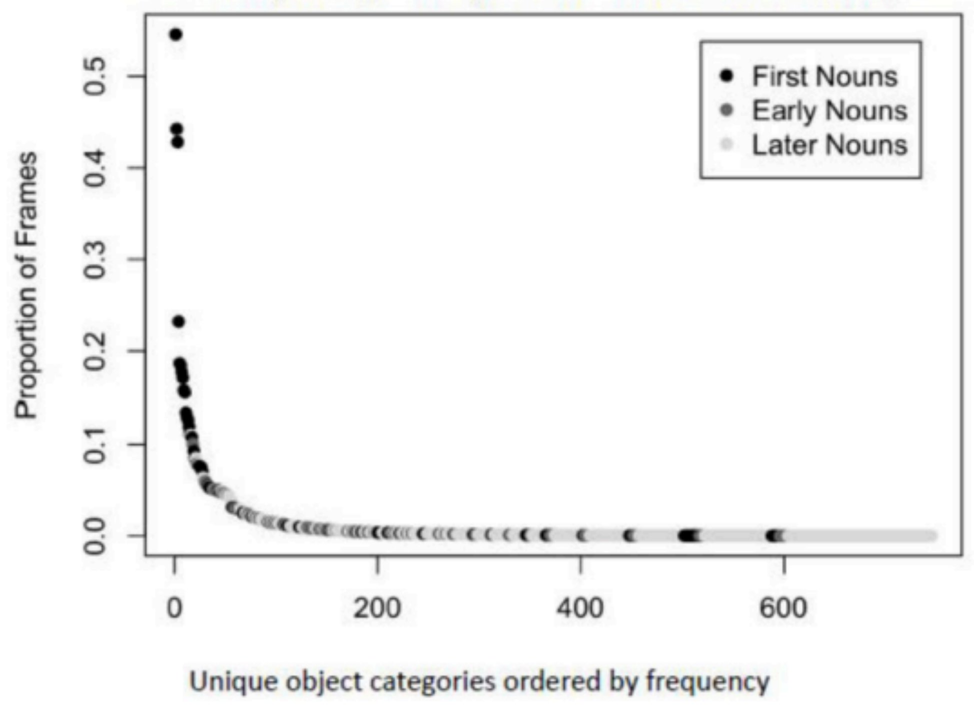
Lessons from Embodied Cognition

Human baby learning is:

Incremental	Social	Phsyical
Exploratory	Language-based	Multi-modal

We will discuss self-supervised methods that are (partly) inspired by human multi-modal learning (exploiting redundant signal).

Babies build curricula



Heavy focus on a small number of objects

Practical Challenges

"In order that the machine should have a chance of finding things out for itself it should be allowed to roam the countryside, and the danger to the ordinary citizen would be serious." *Turing, 1948*

There are some practical challenges to embodied learning.

Simulation may help.

References/Image credits

L. B. Smith and M. Gasser, "The Development of Embodied Cognition: Six Lessons from Babies," Artificial Life (2005)

L. B. Smith et al., "The Developing Infant Creates a Curriculum for Statistical Learning", *Trends in Cognitive Sciences* (2018)

A. M. Turing, "Intelligent Machinery", (1948)

Self-supervised Learning - creating your own supervision

Learning via prediction - Helmholtz

Each movement we make by which we alter the appearance of objects should be thought of as an experiment designed to test whether we have understood correctly the invariant relations of the phenomena before us, that is, their existence in definite spatial relations.

Helmholtz, 1878

Generate labels by predicting the future

Redundancy provides knowledge - Barlow

Learning requires previous knowledge: To detect a new association (e.g. event C precedes event U), requires knowledge of the prior probabilities of C and U. We can then learn new associations as occurrences of C followed by U more frequently than would happen by chance.

Redundancy: To know "what usually happens", we need redundancy or "structure" in the input signal (e.g. sensory messages of the same event from different modalities). Redundant signal (by definition) can be predicted from the remaining signal.

Generate labels from redundant signal

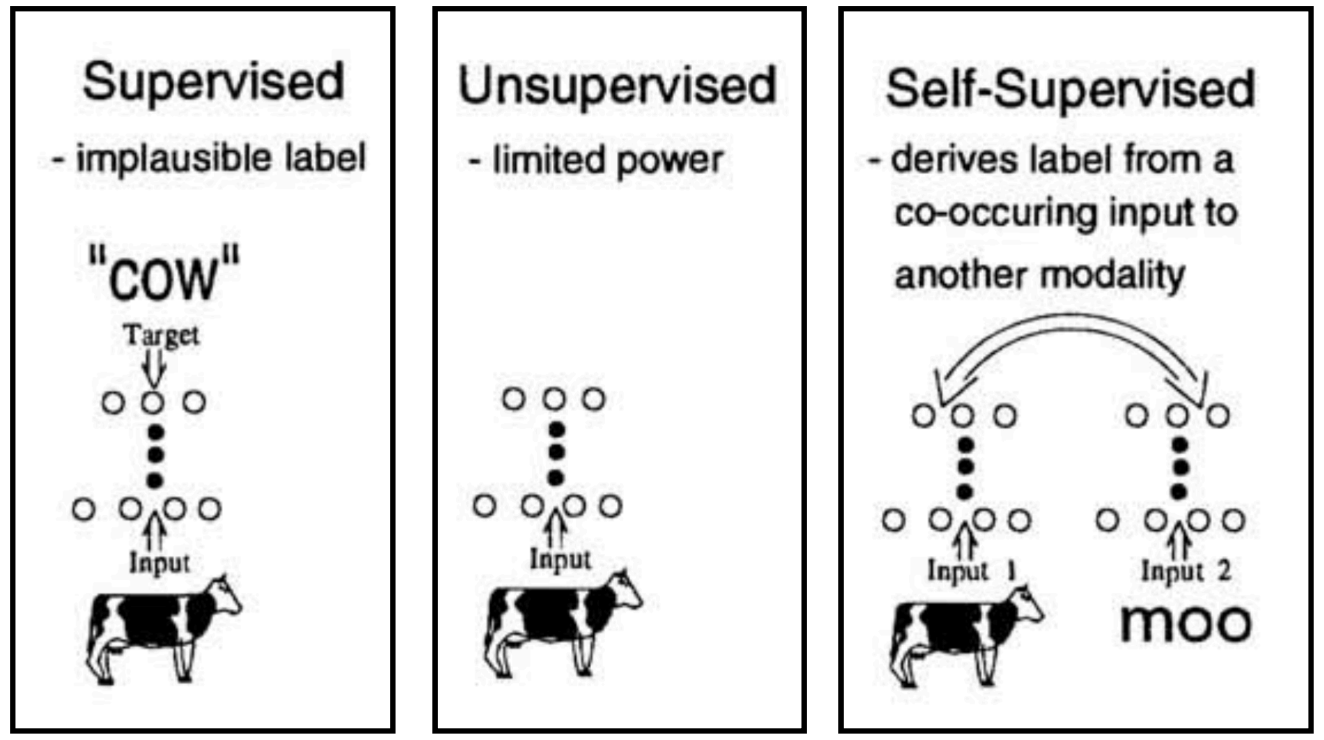
Computational trick: factorial codes for learning new associations

When learning pairwise associations between N events, we need to store N^2 co-occurrence probabilities.

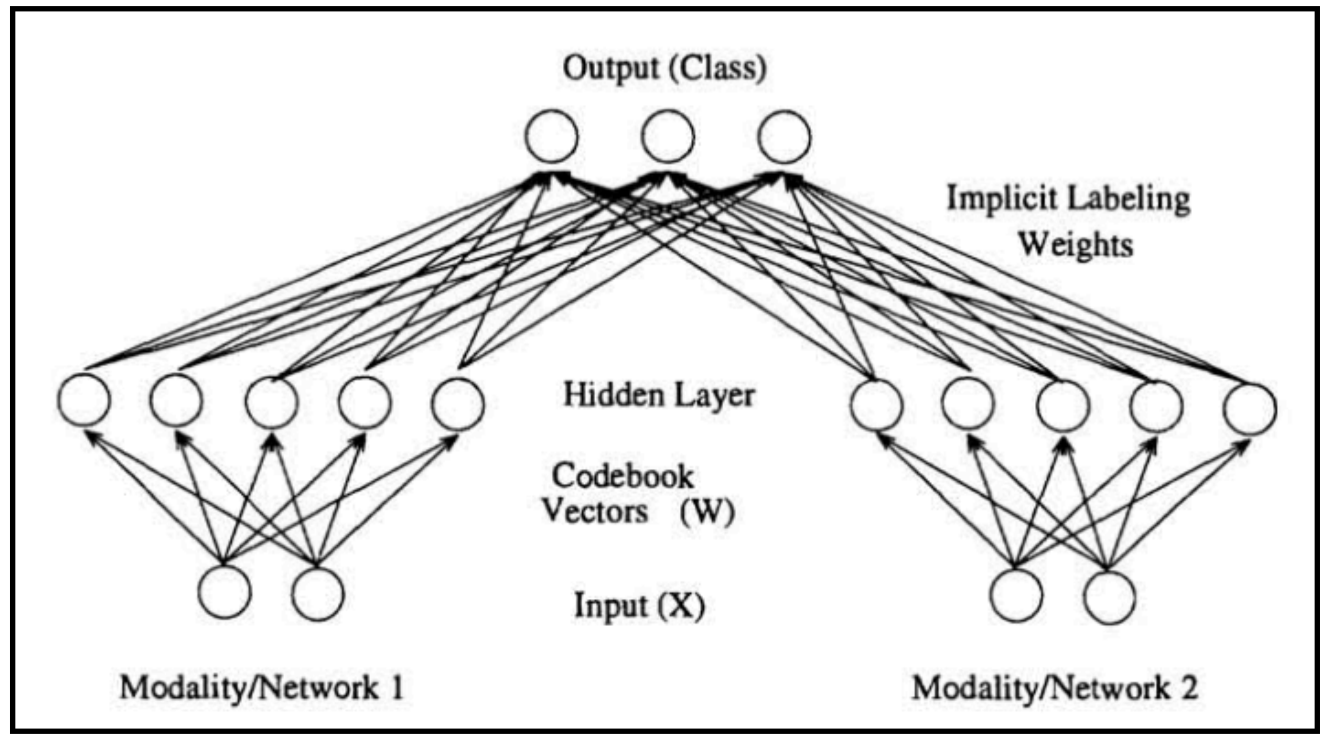
If our representations of events C and U are statistically independent, we can compute the chance co-occurrence of C and U from their marginals: i.e. $P(C)P(U)$, so we need only store N event probabilities!

Barlow suggested *Minimum Entropy Coding* to obtain such factorial representations - but this principle applies more generally.

Exploiting Multi-modal Correlation - de Sa



Learning signal: Minimise disagreement between class labels predicted from each modality:



Note: in the modern literature, the distinction between self-supervised and unsupervised methods can be blurry.

References/Image credits:

H. L. F. Helmholtz, "The Facts in Perception" (1878)

H. B. Barlow, "Unsupervised learning", Neural computation (1989)

V. R. de Sa, "Learning Classification with Unlabeled Data", *NeurIPS* (1993)

Self-supervised Learning - context as supervision

Natural Language Processing

Unlabelled text corpora have long been used to provide (relatively) low-level supervision for neural networks, with the hope that their distributed representations will enable generalisation.

Autoregressive models

Factor the probability of a sequence, x_1^T, as conditionals:

$$P(x_1^T) = \prod_{t=1}^{T} P(x_t \mid x_1^{t-1})$$

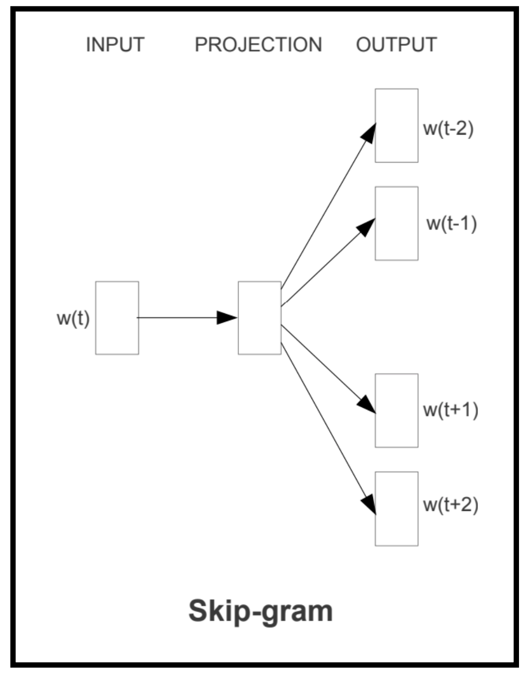
Train a network to maximise likelihood of text corpus.

Predict next character (Schmidhuber et al., 1996)

Predict next word (Bengio et al., 2003)

Predicting context

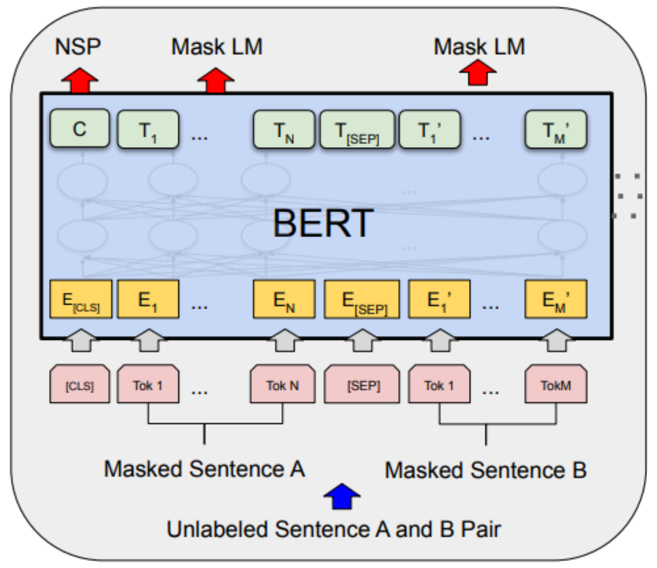
Word2Vec was trained to predict surrounding words.



This work highlighted the critical importance of having *lots* of training data.

Multitask masking

BERT was trained to predict *randomly masked words* and next sentence prediction.



Pre-training

BERT showed the benefits of using a high-capacity bi-directional transformer.

Computer Vision

In vision, we can train the network by tasking it with playing a game (often called a pretext task).

We typically don't care about performance on the pretext task itself, but we hope that by solving it, a model learns good representations of the visual world.

Example:

Question 1:

Question 2:

Key idea: a model can only solve these questions once it learns about cats, buses and trains. No labelling is required!

Warning: sometimes the model won't solve the task in the way you wanted!

Doersch et al. found that the network could "cheat" by exploiting chromatic aberration to solve the puzzle unless it was prevented from doing so.

References/Image credits

J. Schmidhuber and S. Heil, “Sequential neural text compression”, IEEE Trans. on Neural Networks (1996)

Y. Bengio et al., “A Neural Probabilistic Language Model”, JMLR (2000)

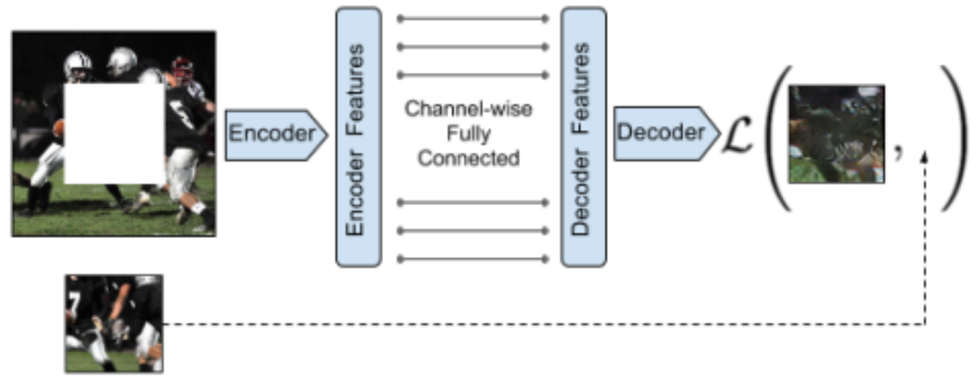
T. Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”, ICLR (2013)

J. Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, NAACL (2019)

C. Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV (2015)

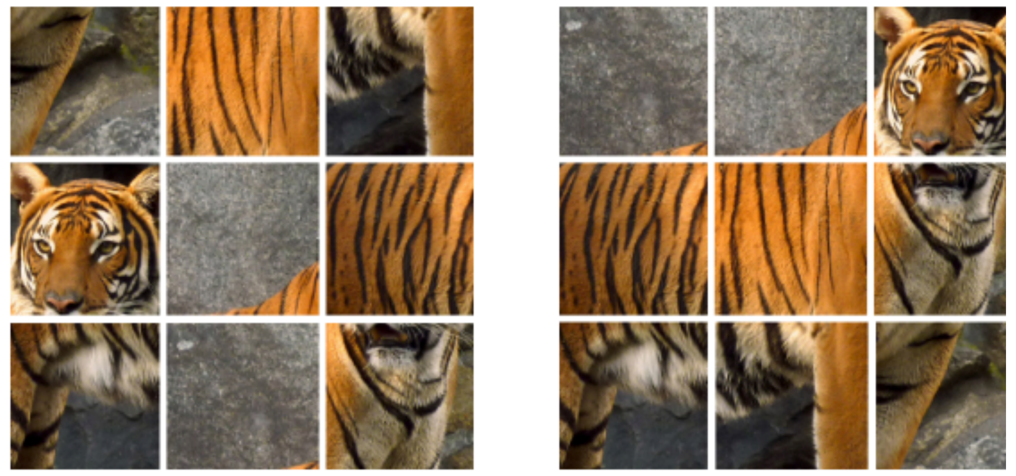
Self-supervised Learning - pretext tasks

Learning by Inpainting



Pathak et al., 2016

Jigsaw Puzzles

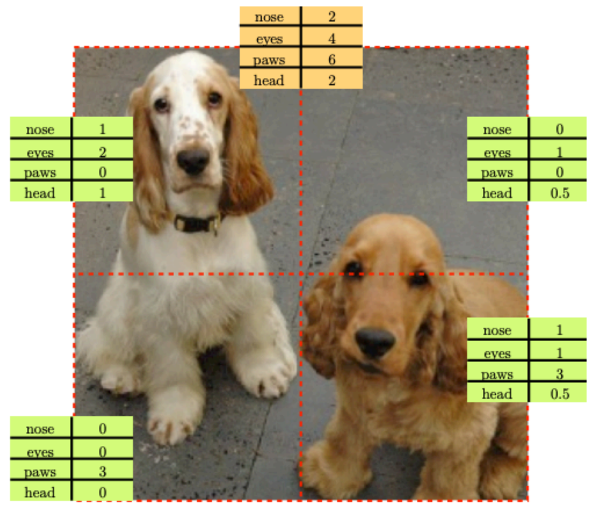


Noroozi and Favaro, 2016

Colourisation

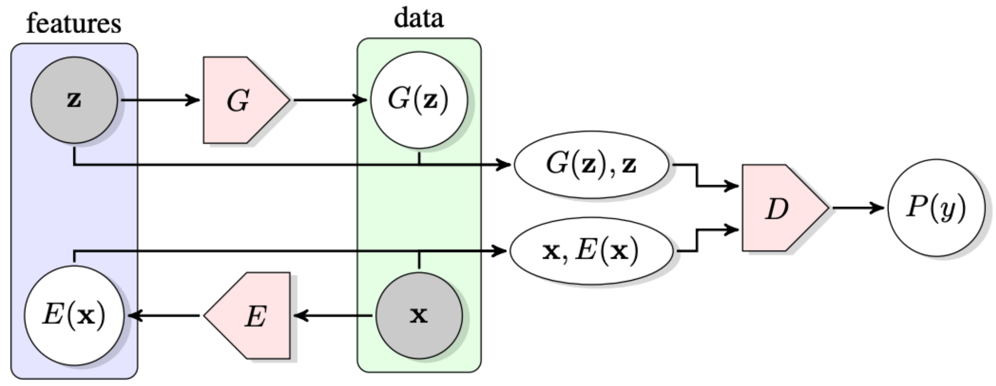
Zhang et al., 2016

Counting



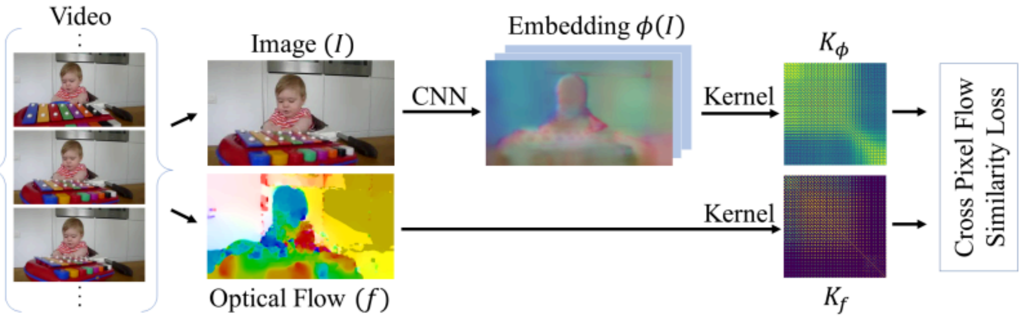
Noroozi et al., 2017

Inverting an Image GAN



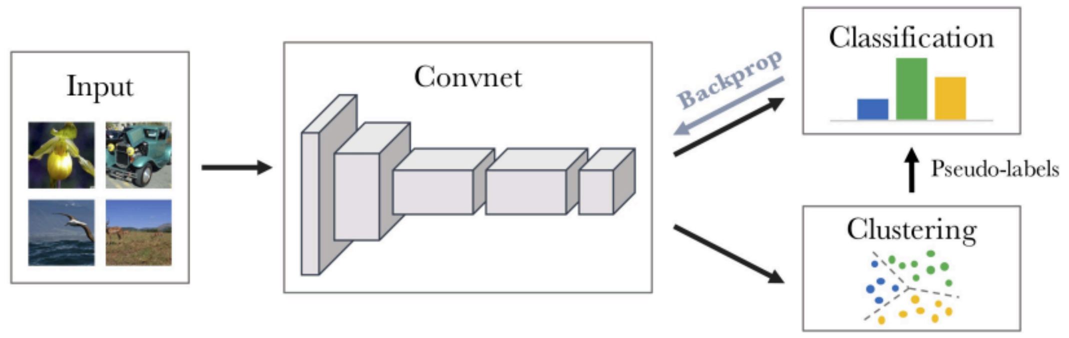
Donahue et al, 2017

Grouping via Common Fate



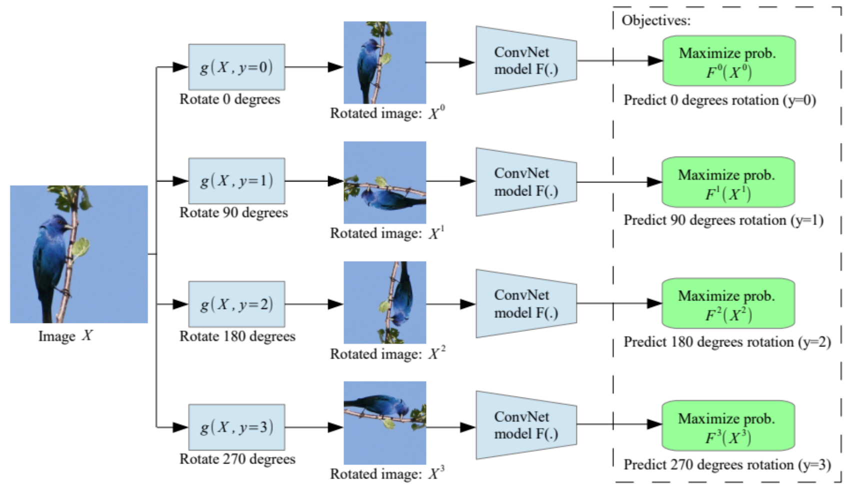
Mahendran et al, 2018

Clustering



Caron et al., 2018

Rotation Prediction



Gidaris et al., 2018

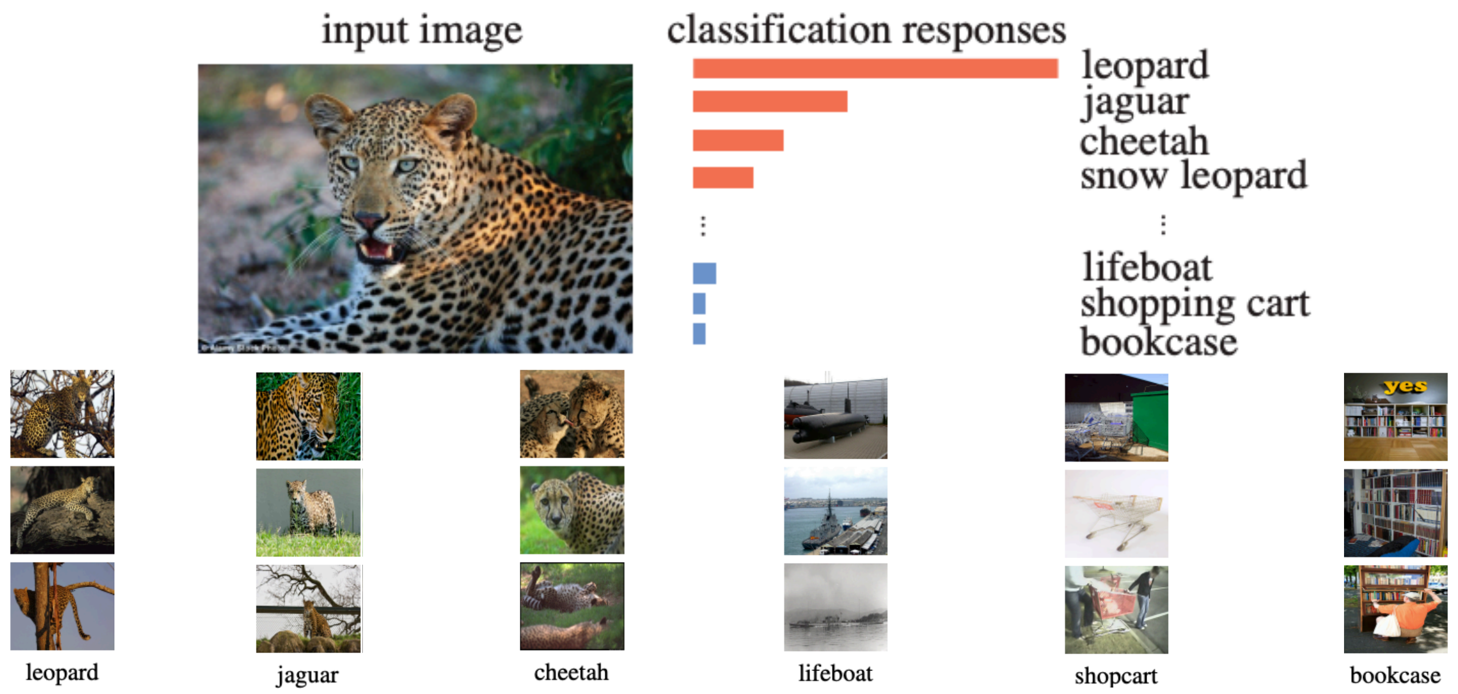
References/Image credits

D. Pathak et al., “Context Encoders: Feature Learning by Inpainting”, CVPR (2016)
M. Noroozi and P. Favaro, “Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles”, ECCV (2016)
R. Zhang et al., “Colorful Image Colorization”, ECCV (2016)
M. Noroozi et al., “Representation Learning by Learning to Count”, ICCV (2017)
J. Donahue et al., “Adversarial Feature Learning”, ICLR (2017)
A. Mahendran et al., “Cross Pixel Optical Flow Similarity for Self-Supervised Learning”, ACCV (2018)
M. Caron et al., “Deep Clustering for Unsupervised Learning of Visual Features”, ECCV (2018)
S. Gidaris et al. “Unsupervised Representation Learning by Predicting Image Rotations”, ICLR (2018)

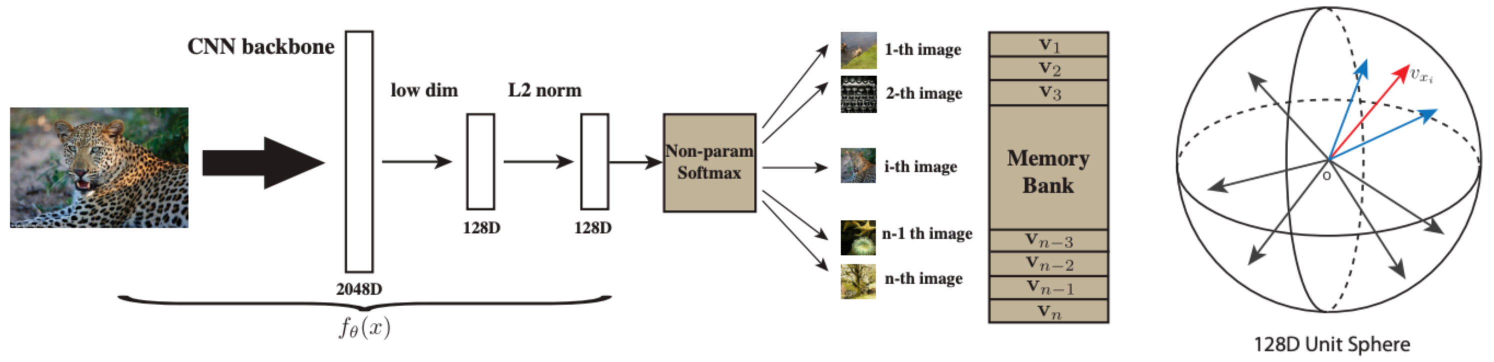
Self-supervised Learning - instance discrimination

Learning via Non-Parametric Instance Discrimination

Motivation: despite training with semantic labels, fully-supervised CNNs appear to capture the visual similarity between instances:



Can we learn a representation that captures similarity among instances, by training it to discriminate individual **instances**, rather than semantic classes?



Store instance features in a *memory bank*.
Learn to spread them out across a hypersphere.

No labels are required, but strong representations emerge.

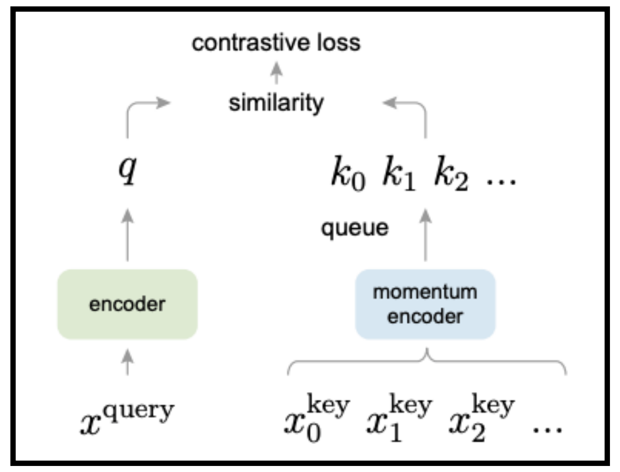
Momentum Contrast

Motivation: Instance discrimination works well, but memory banks have an issue:

- Re-computing the feature bank (one feature per image in the dataset) every time the CNN changes is prohibitively expensive.
- If memory bank instances are not updated, they grow increasingly stale with every optimisation step during training (*suboptimal* for instance discrimination).

MOCO (Momentum Contrast) aims to avoid staleness this by:

1. Replacing the memory bank with a queue of recently encoded samples (fewer than the full dataset).
2. Encoding queue samples with a *momentum encoder (a slow moving average of query encoder weights)*



MOCO uses some terminology:

- "keys" to refer to instances encoded in the queue with the momentum encoder
- "queries" are instances to be compared against keys
- Positives pairs - queries and keys originating from the same image.

The instance discrimination task is to uniquely match queries against keys that form their positive pairs (optimising an InfoNCE loss). The resulting resulting query encoder then provides a useful representation for downstream tasks.

References/Image credits

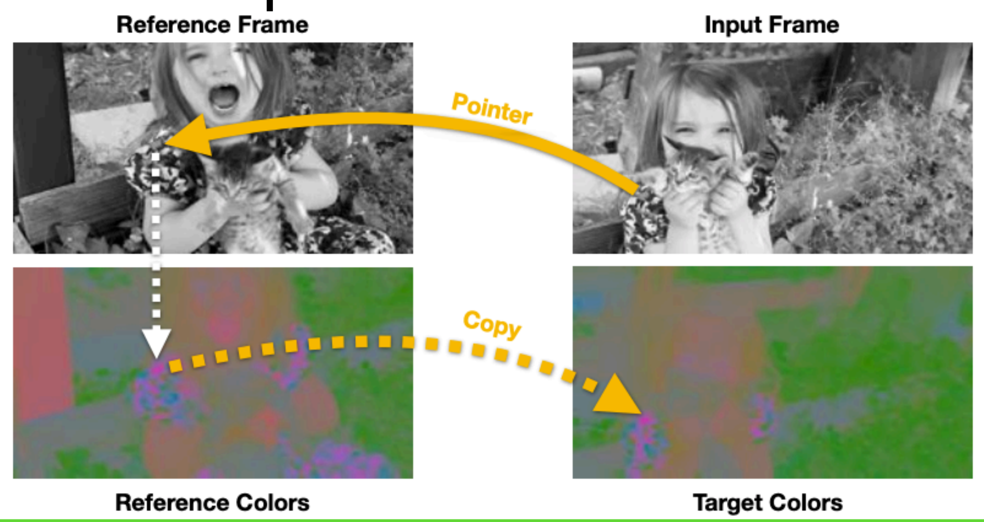
Z. Wu et al., "Unsupervised Feature Learning via Non-parametric Instance Discrimination", *CVPR* (2018)
A. van den Oord et al., "Representation Learning with Contrastive Predictive Coding", arxiv (2018)
K. He et al., "Momentum Contrast for Unsupervised Visual Representation Learning" CVPR (2020)

Self-supervised Learning - Beyond Image Representations

Learning tracking by cololurisation

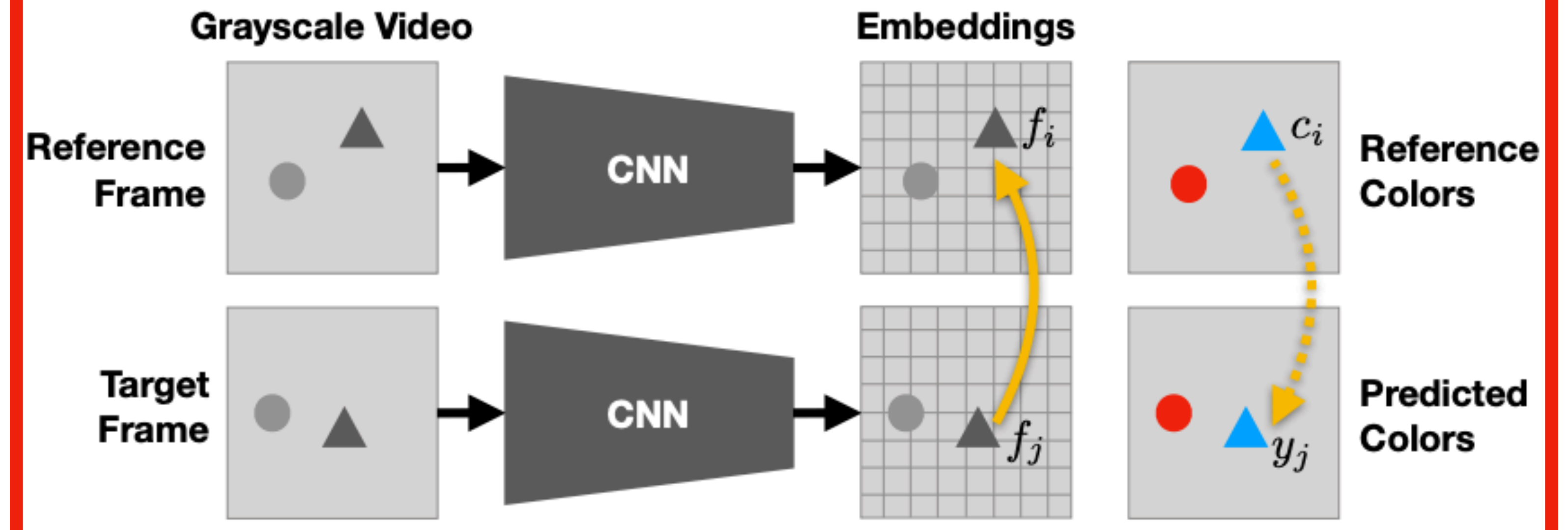
Key Idea

Use **colours** across unlabelled videos as a source of supervision.

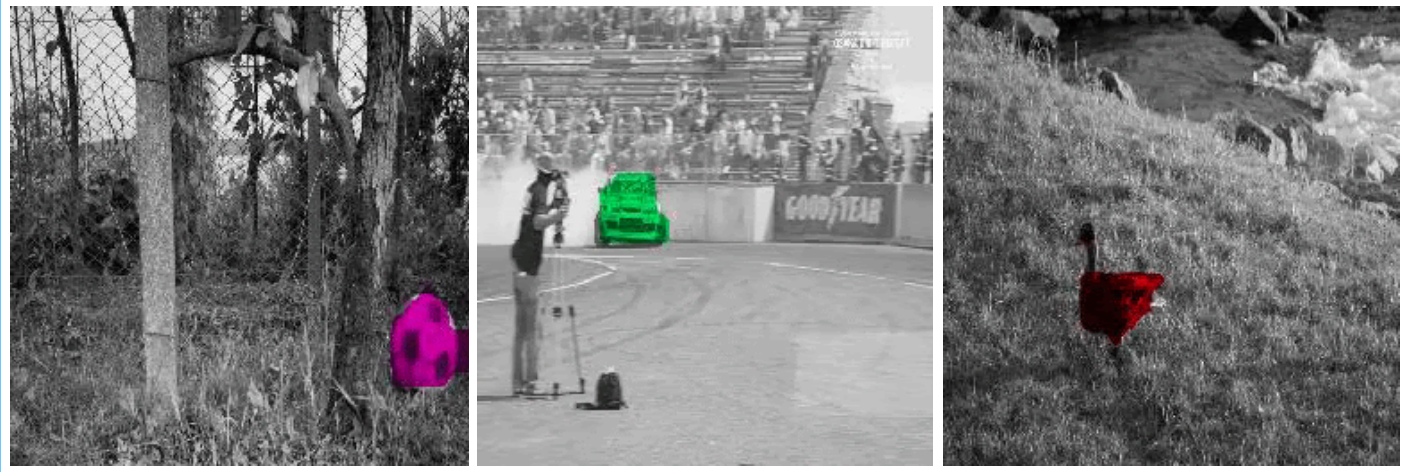


Method

Compute low dimensional embeddings at each location:



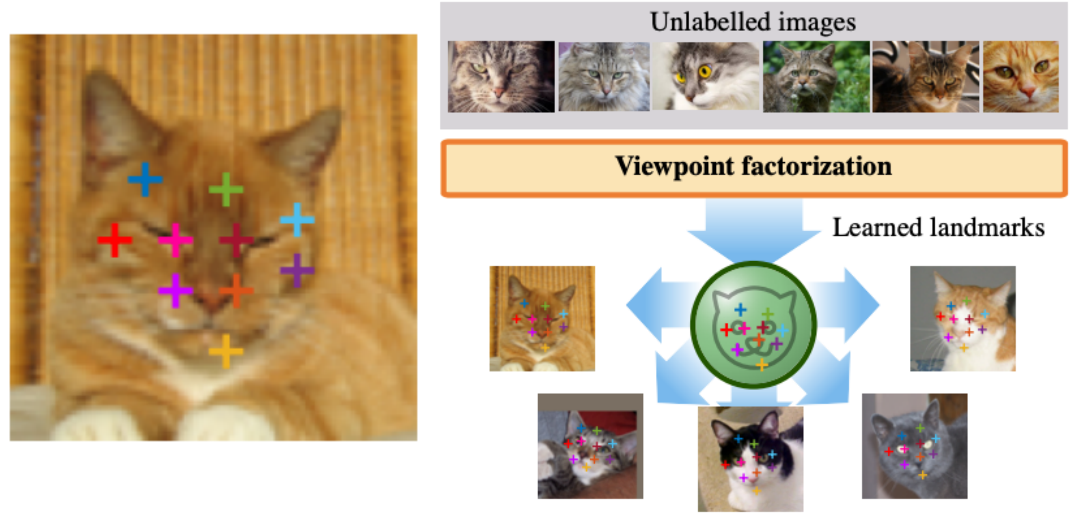
Results



Learning object keypoints

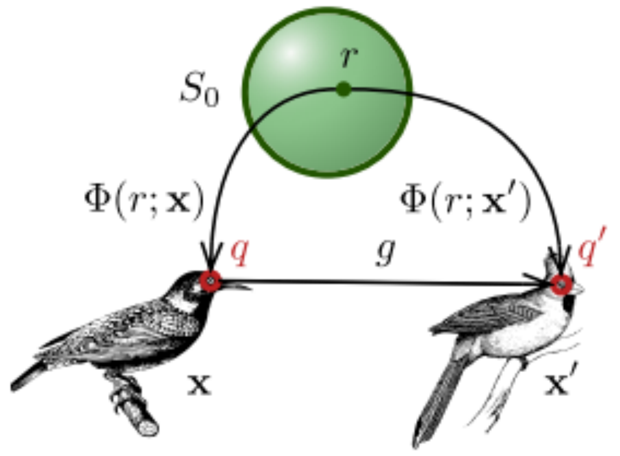
Key Idea

Learn **keypoints** for consistent locations.



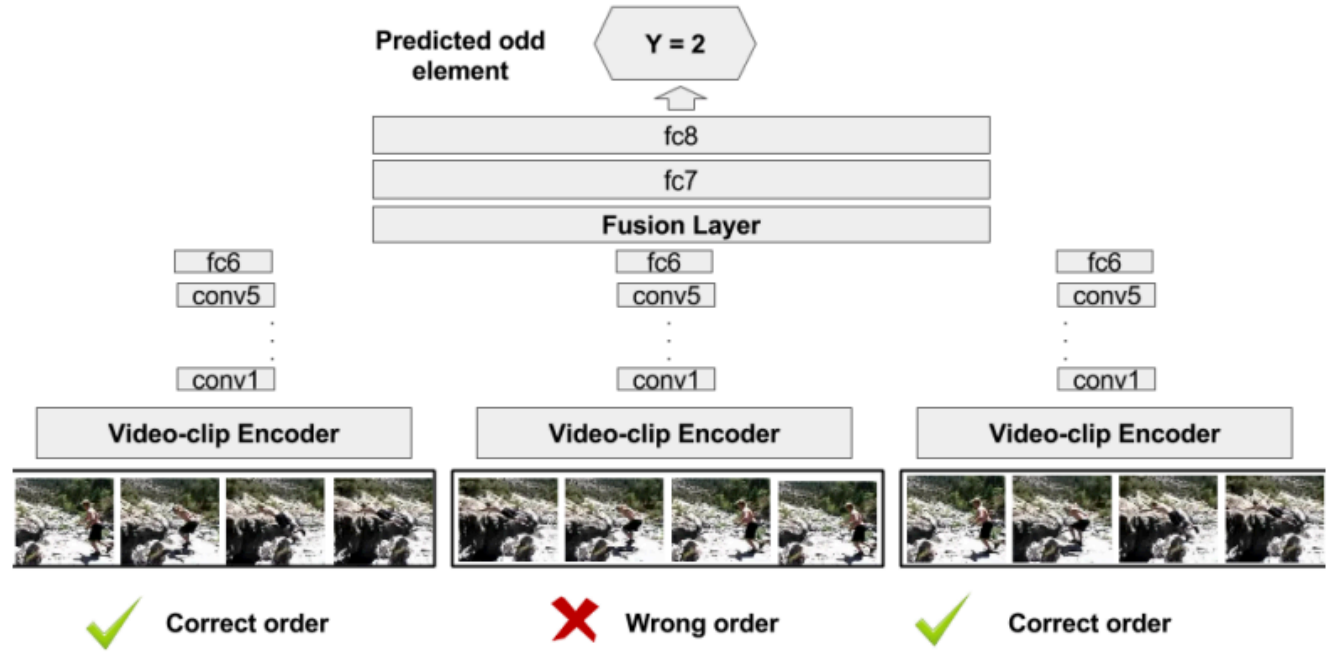
Method

Enforce **equivariance** with respect to geometric image transformations.



Learning video representations for action recognition

Key idea: pass a model several video clips and task it with predicting which clip was shuffled.



References/Image credits

C. Vondrick et al., "Tracking Emerges by Colorizing Videos", *ECCV* (2018)

J. Thewlis et al., "Unsupervised Learning of Object Landmarks by Factorized Spatial Embeddings", ICCV (2017)

B. Fernando et al., "Self-Supervised Video Representation Learning with Odd-One-Out Networks", CVPR 2017

Outline

- Self-supervised learning
- Pseudo-labelling

Semi-supervised learning and pseudo-labelling

Semi-supervised learning

Semi-supervised learning considers the situation in which the learner has access to both labelled data (typically small in scale) and unlabelled data (typically large in scale).

Pseudo-labelling

Pseudo-labelling (sometimes called "self-training" or "self-labelling") refers to variations of a simple algorithm:

- Train a classifier on the labelled data
- Predict the labels of the unlabelled data (the resulting predictions are "pseudo-labels")
- Retrain the model on the pseudo-labels
- [Optional] re-generate the pseudo-labels, and repeat.

Example: Word sense disambiguation - Yarowsky, 1995

Task: Perform word sense disambiguation across a corpus (in this case, for the word "plant").

1. Obtain an initial small collection of labelled samples, and use them to train a classifier
2. Predict labels for unlabelled instances, retaining those with high confidence (optionally filtering/ expanding the labelled set via automatic heuristics)
3. Repeat until convergence to a final state

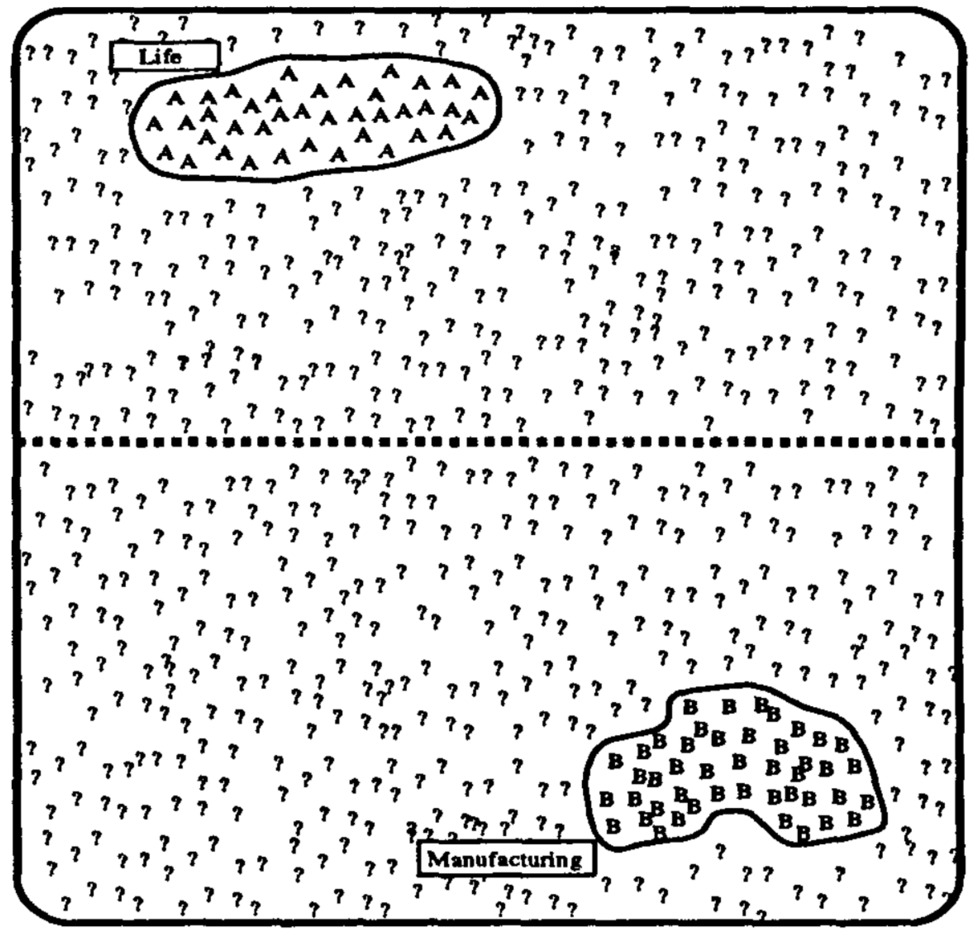


Figure 1: Sample Initial State

A = SENSE-A training example
B = SENSE-B training example
? = currently unclassified training example
$\boxed{\text{Life}}$ = Set of training examples containing the collocation "life".

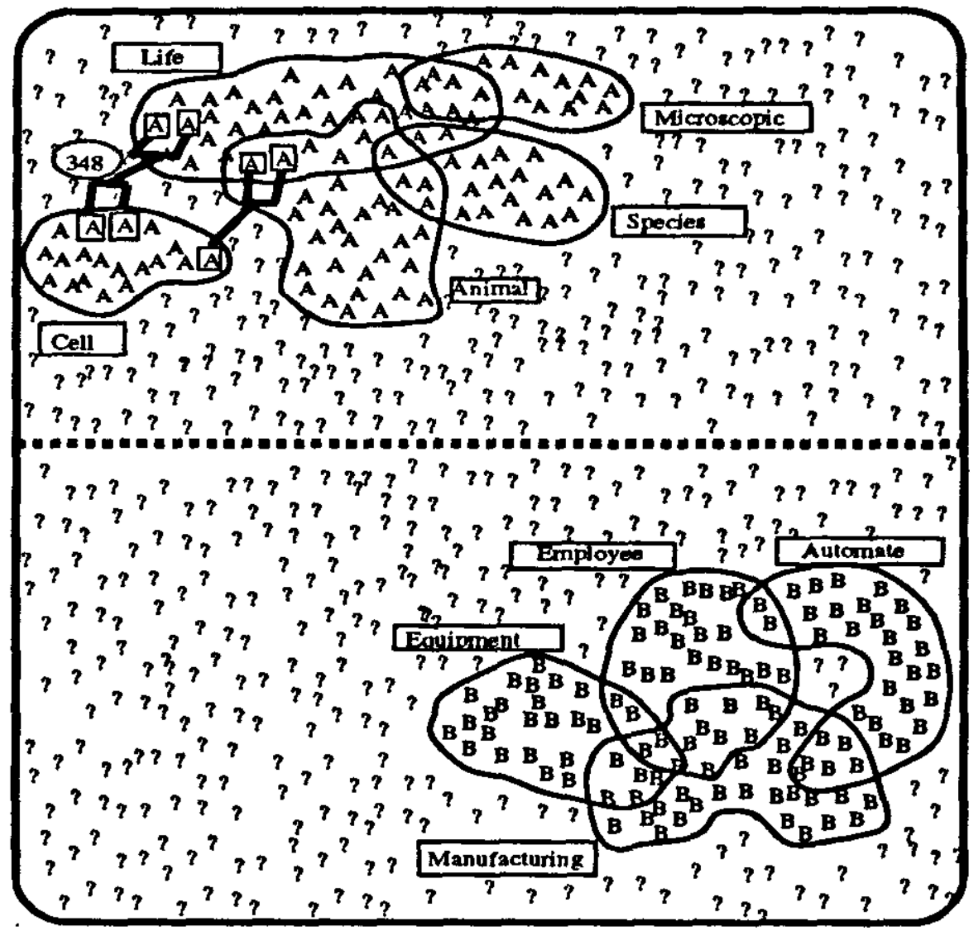


Figure 2: Sample Intermediate State

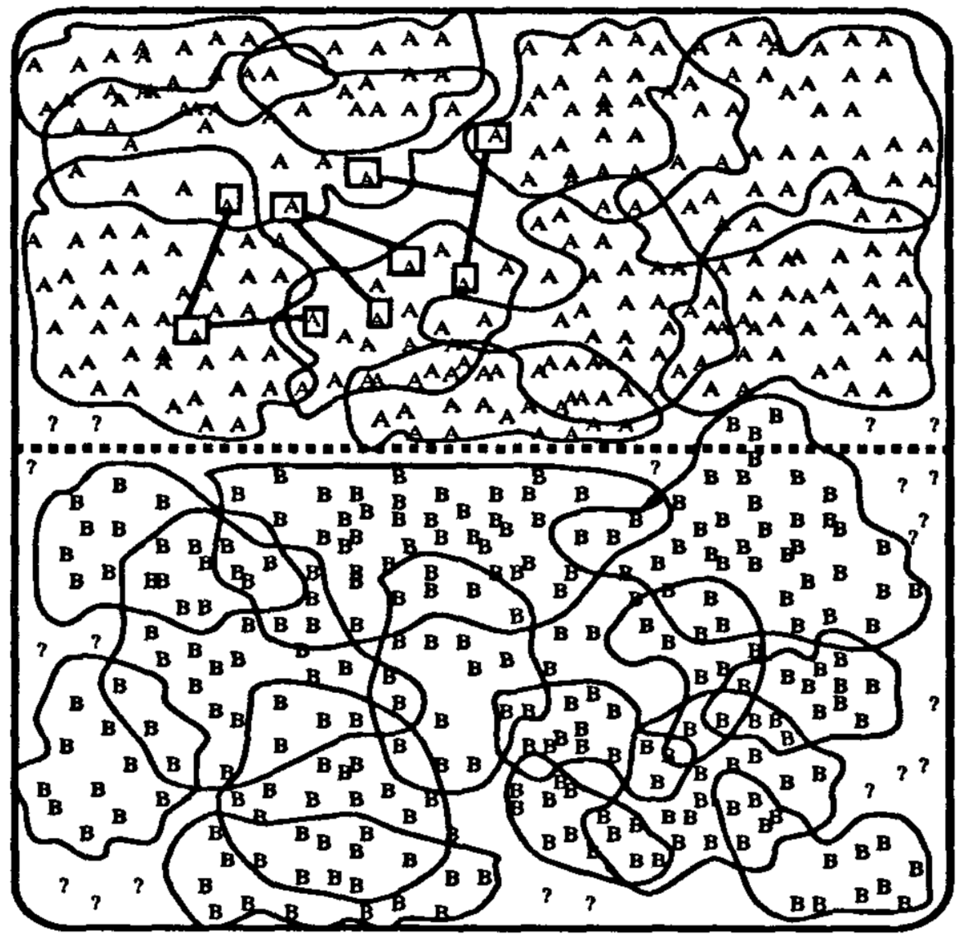


Figure 3: Sample Final State

"It thrives on raw, unannotated monolingual corpora - the more the merrier", Yarowsky

References/Image credits

H. J. Scudder, "Probability of error of some adaptive pattern-recognition machines", IEEE Trans. Inf. Theory (1965)

D. Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods", ACL (1995)

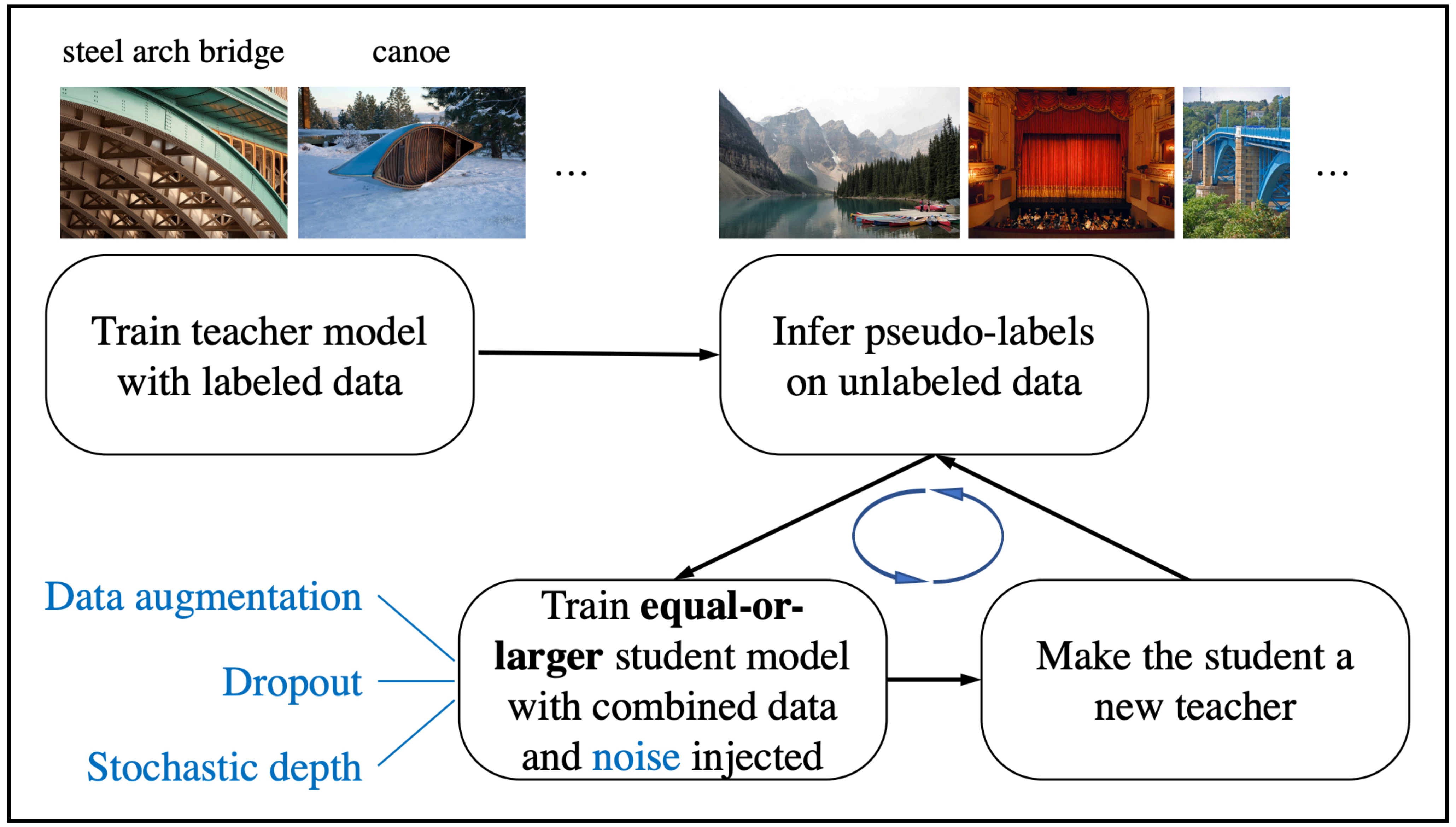
Pseudo-labelling

Self-Training with Noisy Student - Xie 2020

Pseudo-labelling was recently applied to large-scale image classification using:

ImageNet (1M labelled images)

JFT-300M (303M unlabelled images)



This approach achieved significant performance gains over ImageNet-only training.

Pseudo-labelling may become increasingly valuable in future as sensory data grows faster than annotation

Reference/Image credits

Q. Xie et al., "Self-Training With Noisy Student Improves ImageNet Classification", CVPR (2020)

Bonus quote

We are particles of dust on the surface of our planet, which is itself scarcely a grain of sand in the infinite space of the universe. We are the youngest species among the living things of the earth, hardly out of the cradle according to the time reckoning of geology, still in the learning stage, hardly half-grown, said to be mature only through mutual agreement. Nevertheless, because of the mighty stimulus of the law of causality, we have already grown beyond our fellow creatures and are overcoming them in the struggle for existence. We truly have reason to be proud that it has been given to us to understand, slowly and through hard work, the incomprehensibly great scheme of things. Surely we need not feel in the least ashamed if we have not achieved this understanding upon the first flight of an Icarus.

The Facts of Perception, Hermann Helmholtz, 1878