UNIVERSITY OF WARSAW

WYDZIAŁ MATEMATYKI, INFORMATYKI I MECHANIKI

Self-supervised learning

Andrzej Janusz
Daniel Kałuża

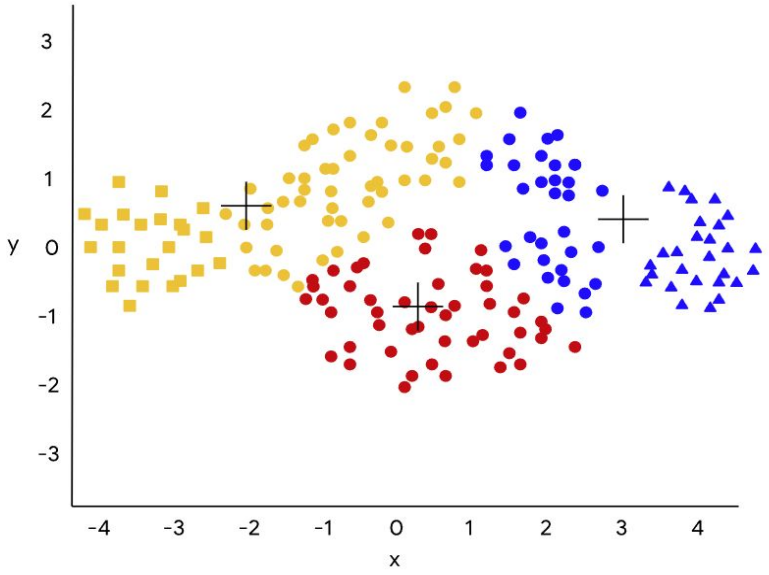
THE PLAN

- A recap from previous lectures.
- Motivation for semi-supervised learning.
- Weakly-supervised learning.
- Semi-supervised learning principles.
- Self-supervised learning.
- Examples of commonly used algorithms.
- Summary.

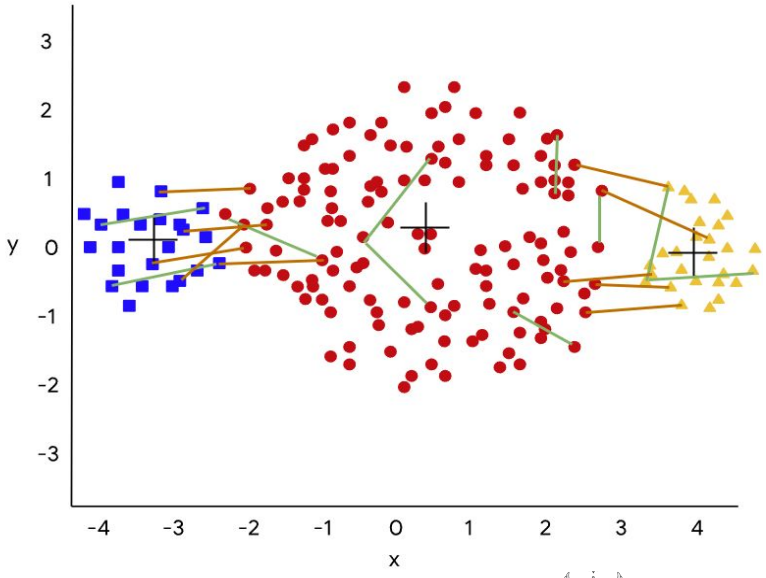
Previously - semi-(un)supervised learning

- Clustering algorithms may work with partially labeled data:
  - Must-links - two cases have to be placed in the same cluster.
  - Cannot-links - two cases must not be in the same cluster.

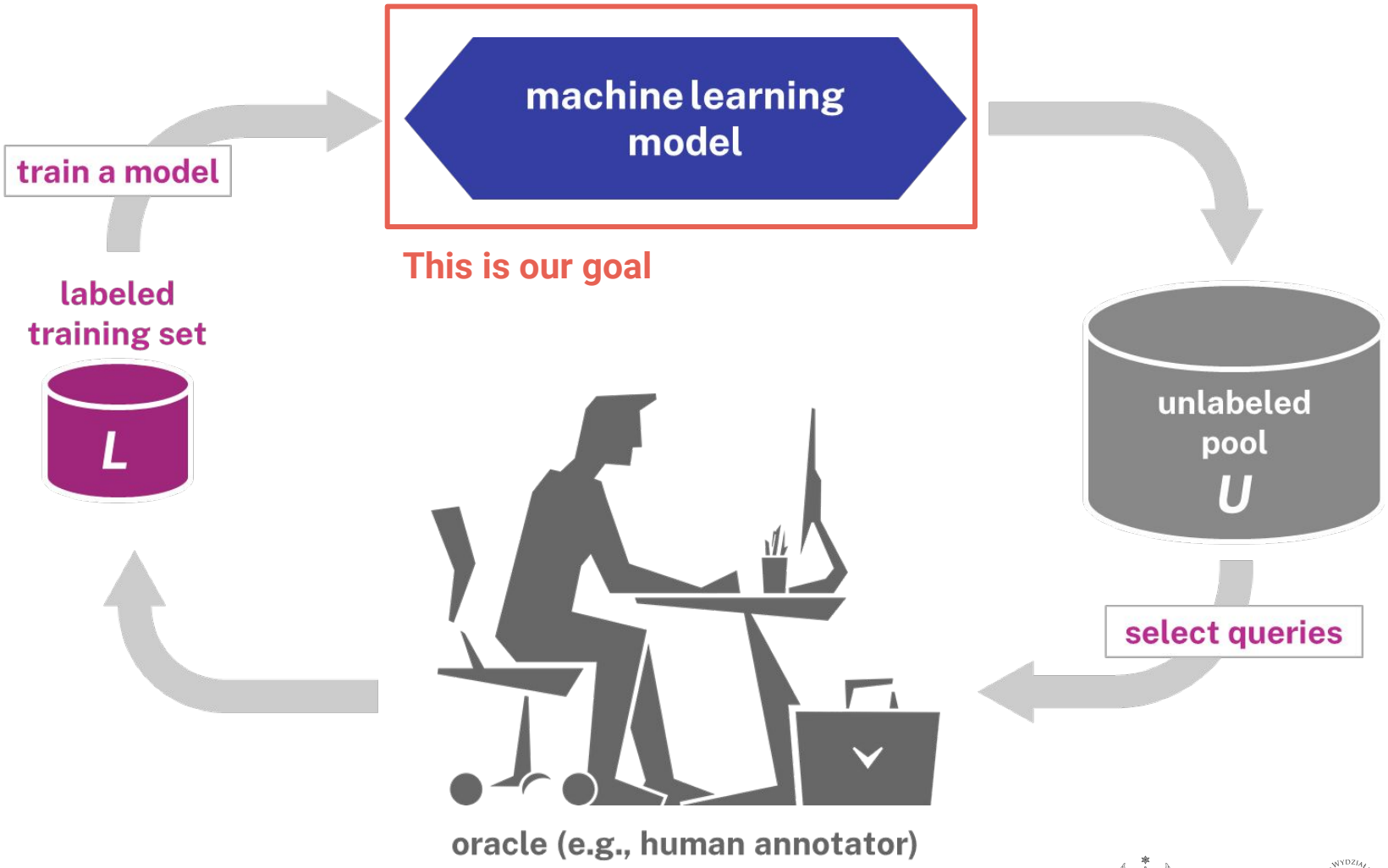
standard k-means



constrained k-means



Previously - the active learning cycle



Dealing with shortages in labeled data

An alternative or addition to active learning:

- The question: can we train a better prediction model without the access to additional labels or experts?
- We can try several “buzz word” strategies:
  - Weakly-supervised learning.
  - Semi-supervised learning.
  - Self-supervised learning.

Weakly-supervised learning

A machine learning paradigm that focuses on learning from data with imperfect labels:

- Noisy labels, possibly artificially generated (e.g. a part of the labels is wrong).
- Partial labels (e.g. we are given only one of true label in the multi-label classification task).
- Cross-task learning (e.g., learn image segmentation using information about image classification).
  - Some people claim that the weakly-supervised learning is an umbrella term that includes semi- and self-supervised techniques.

Semi-supervised learning

A set of machine learning techniques for partly labeled data sets (a small fraction of data is annotated).

The goal:

- Obtain the best possible model assuming only a fixed (and limited) set of labeled samples.

Key points:

- Incorporate the knowledge extracted from the unlabeled part of the data set.
- Utilize the shape of unlabeled data manifold.
- Use the unlabeled data for the representation learning.

Self-supervised learning

The main ideas:

- Pre-train your model using some artificially generated prediction task.
- Use the whole available data set to learn a good data representation (i.e. embeddings of data samples) and provide a warm start to your model.
- Use the learned embeddings to fine-tune your model for the target task with only a small amount of labels.
- Can be viewed as transfer learning from an artificially-defined prediction task.
- *Fine-tune the embeddings for the target task.

* - depends on the used model and approach

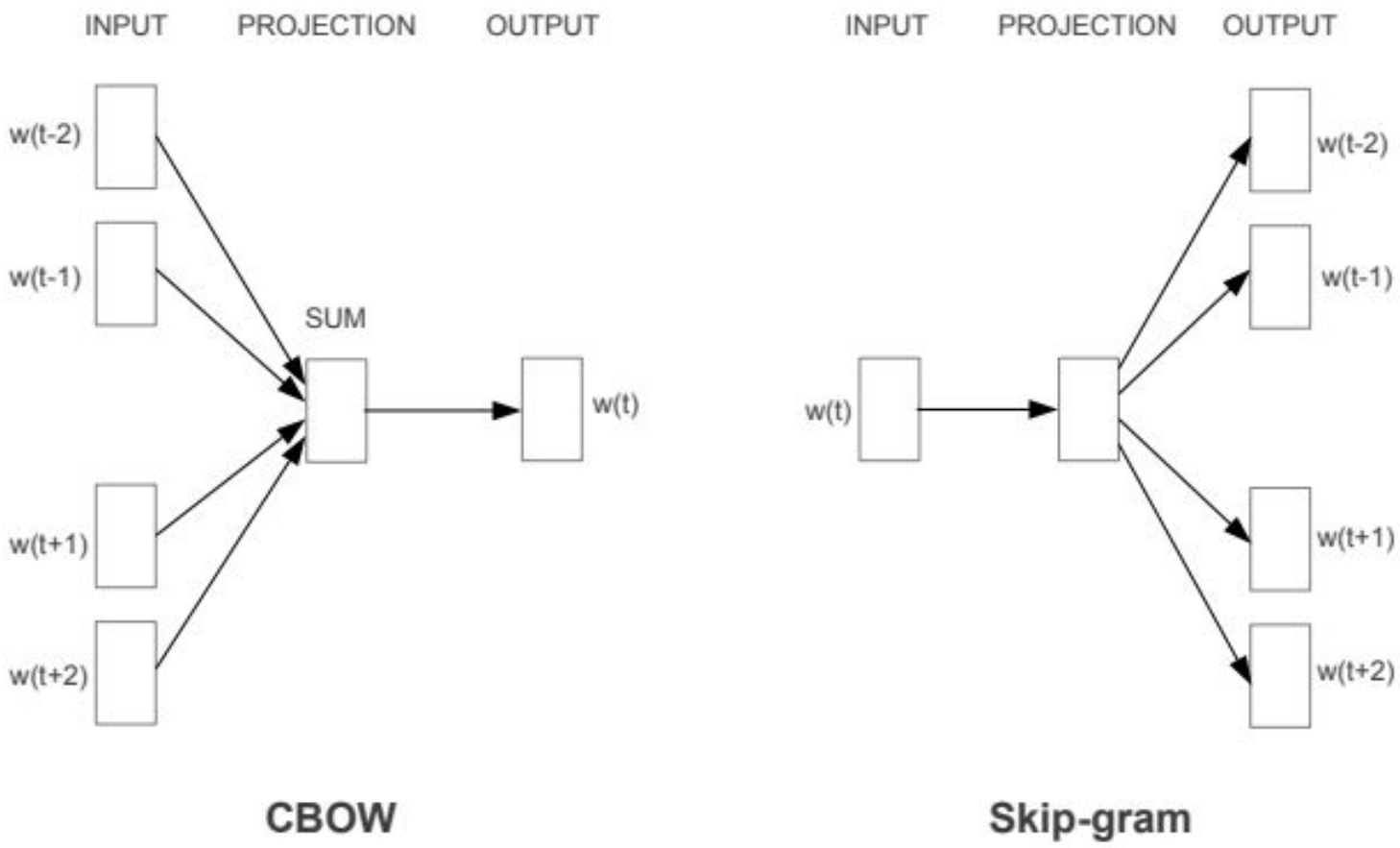
Self-supervised learning - main assumptions

- “A good representation should transfer with a little training.”

  Yann LeCun

- Data representations can be reused in multiple tasks.
- A data representation can be further fine-tuned for a specific task.
- Multiple pre-training tasks can be “fused”.
- Pre-training doesn’t require human annotations.
- Pre-training can be done once beforehand.

Word embeddings - CBOWs vs. skip-grams

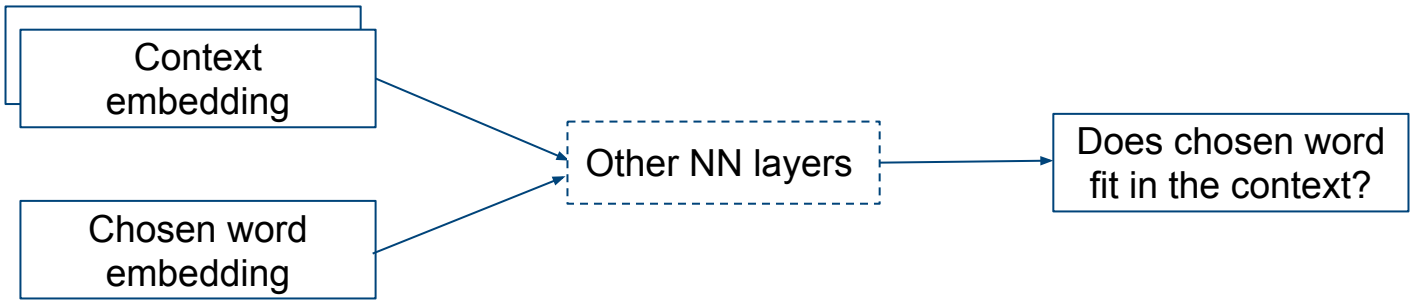


source: Mikolov et al., Efficient Estimation of Word Representations in Vector Space

Word embeddings - negative sampling

Alice has a cat. Cat has Alice.

Alice has a _____. Cat has Alice.

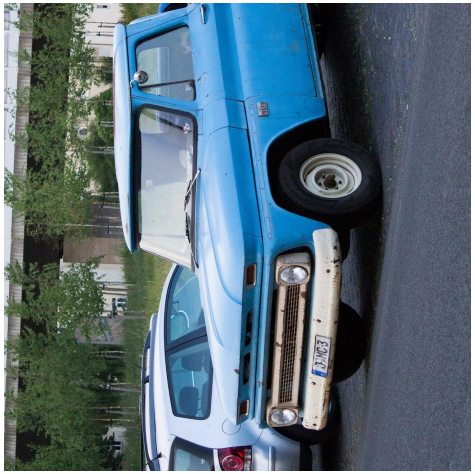


RotNet - Image Rotation prediction

0°

90°

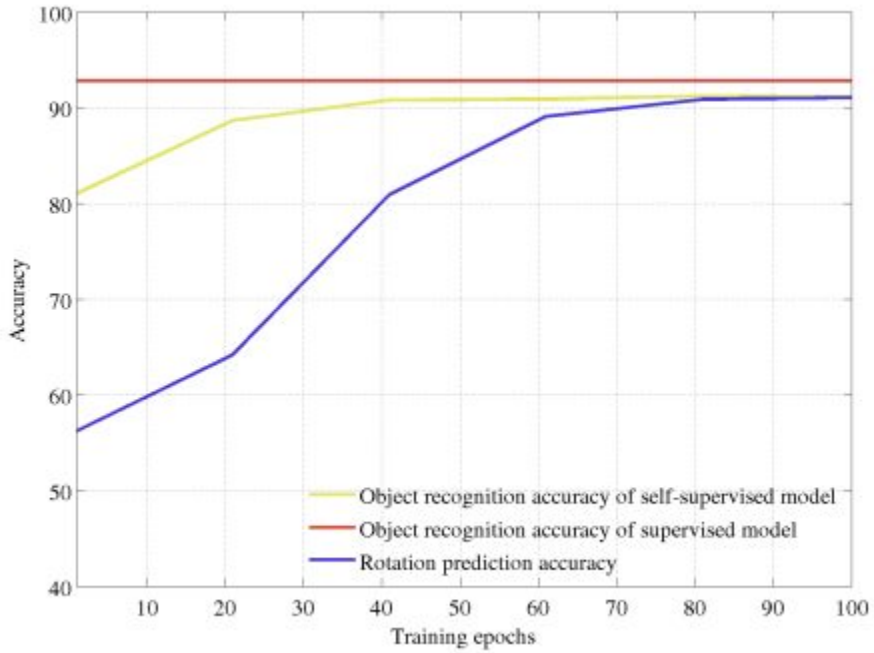
180°

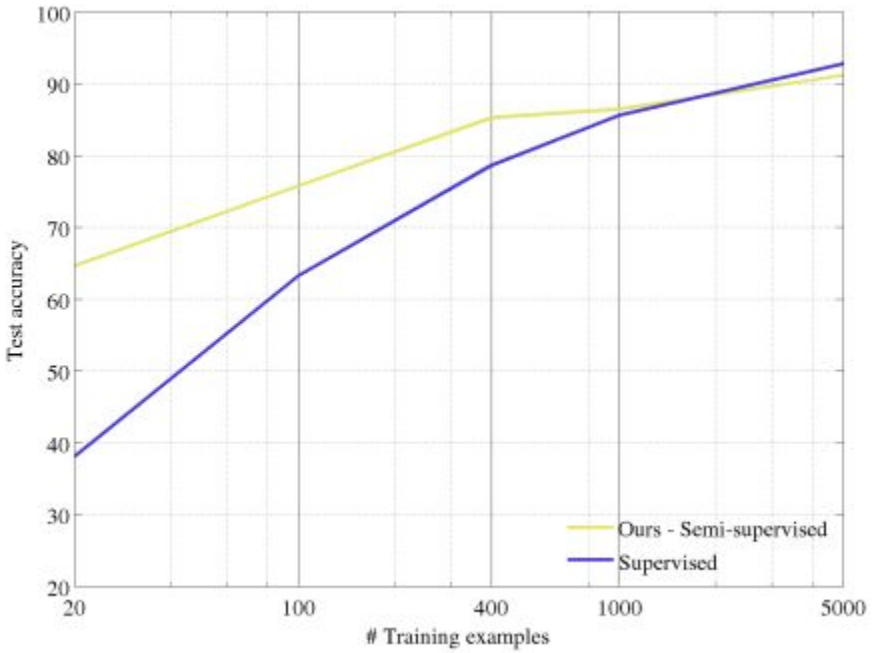


270°

UNIVERSITY OF WARSAW

RotNet - CIFAR10 Results

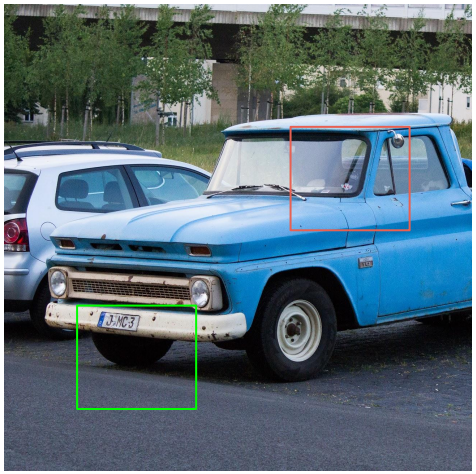


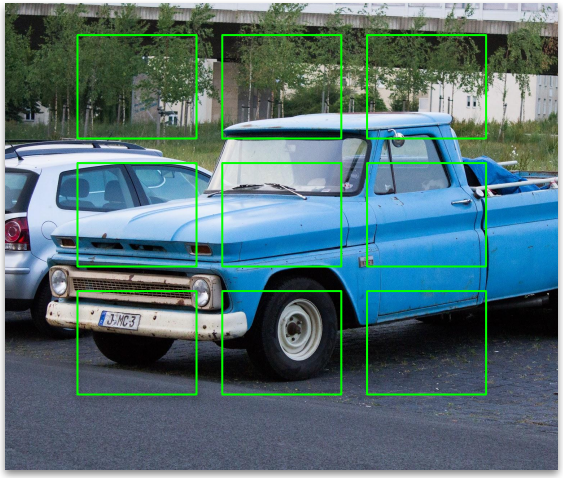


source: https://openreview.net/pdf?id=S1v4N2l0-

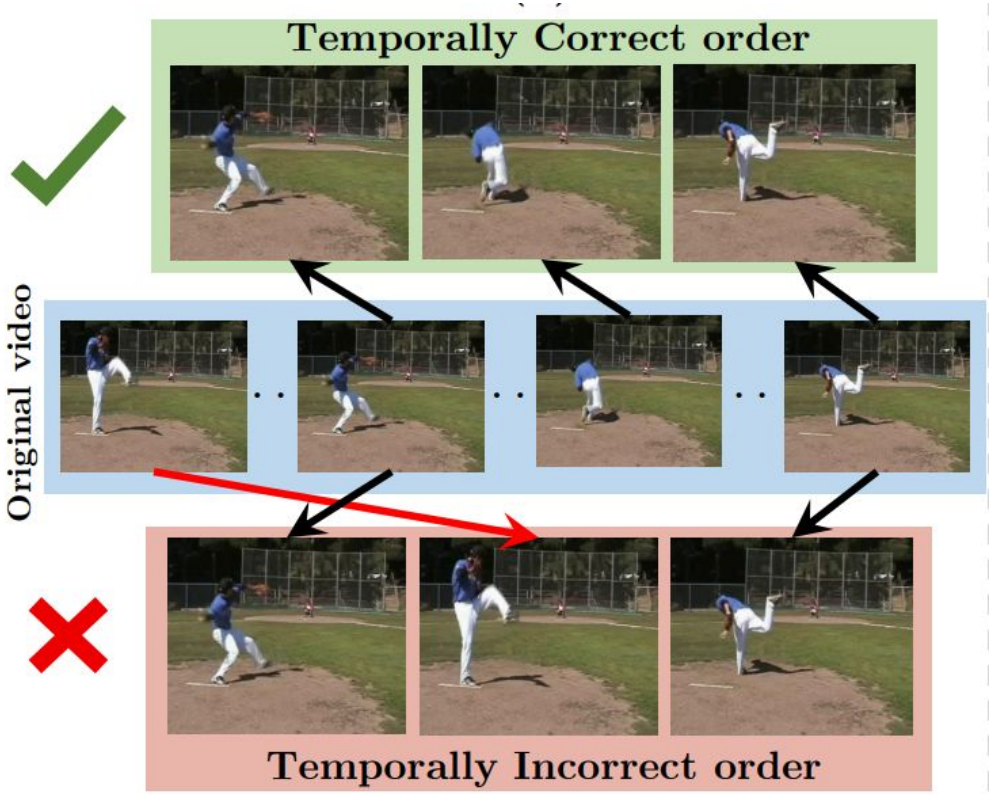
Other examples of image pre-training tasks

- Jigsaw
- Colorization
- Position of 2 patches
- Fill in the blanks



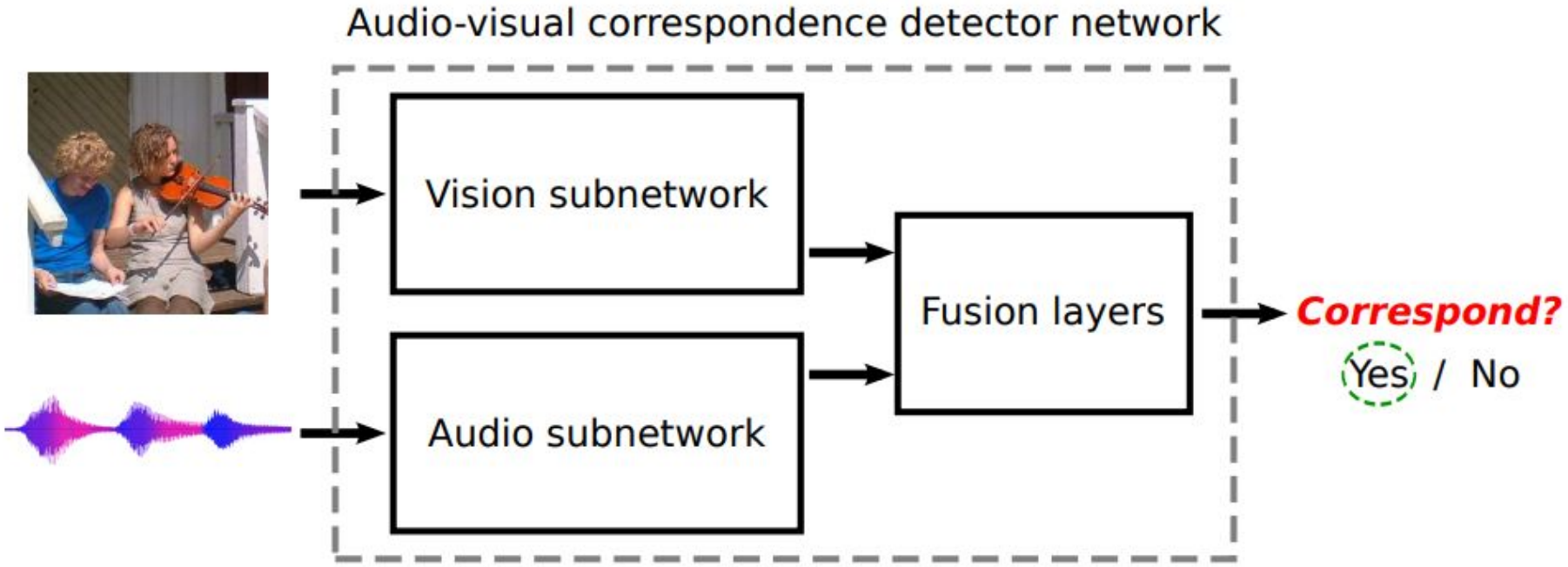


Videos: Shuffle and Learn



Misra et al, Shuffle and Learn: Unsupervised Learning using Temporal Order Verification, ECCV 2016

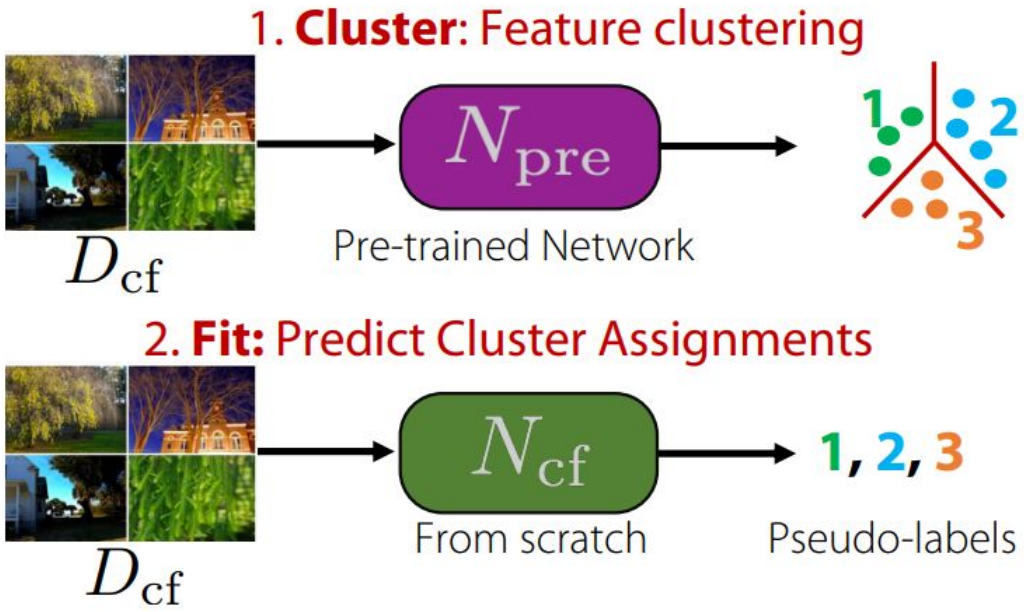
Multimodal representation learning



Arandjelović et al, Look, Listen and Learn, ICCV 2017

The ClusterFit algorithm

1. Take pre-trained NN, run inference to obtain embeddings for downstream task data.
2. Cluster the embeddings using the k-means algorithm.
3. Use predicted clusters as pseudo-labels for the new NN.
4. Use the obtained representation and model weights to initialize the final model.

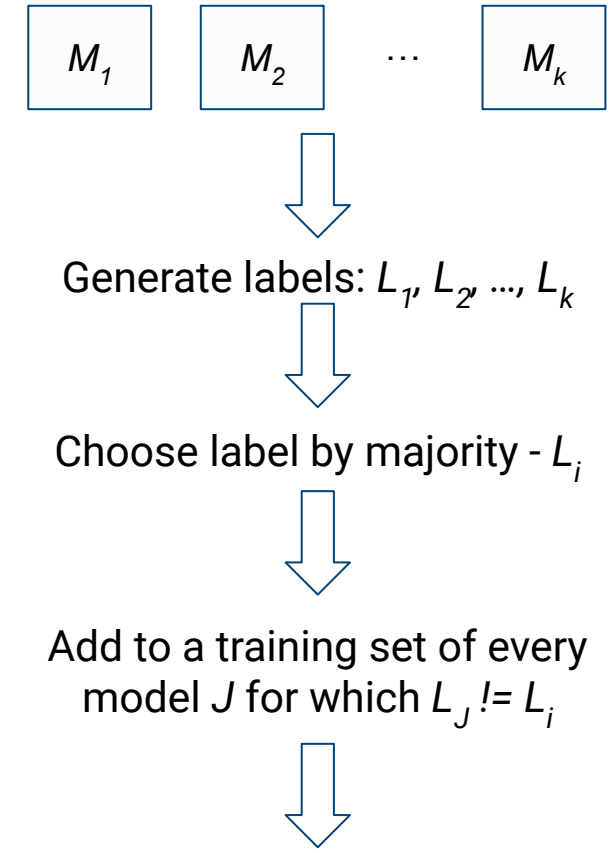


Source: Yen et al., ClusterFit: Improving Generalization of Visual Representations

ClusterFit - why does it work?

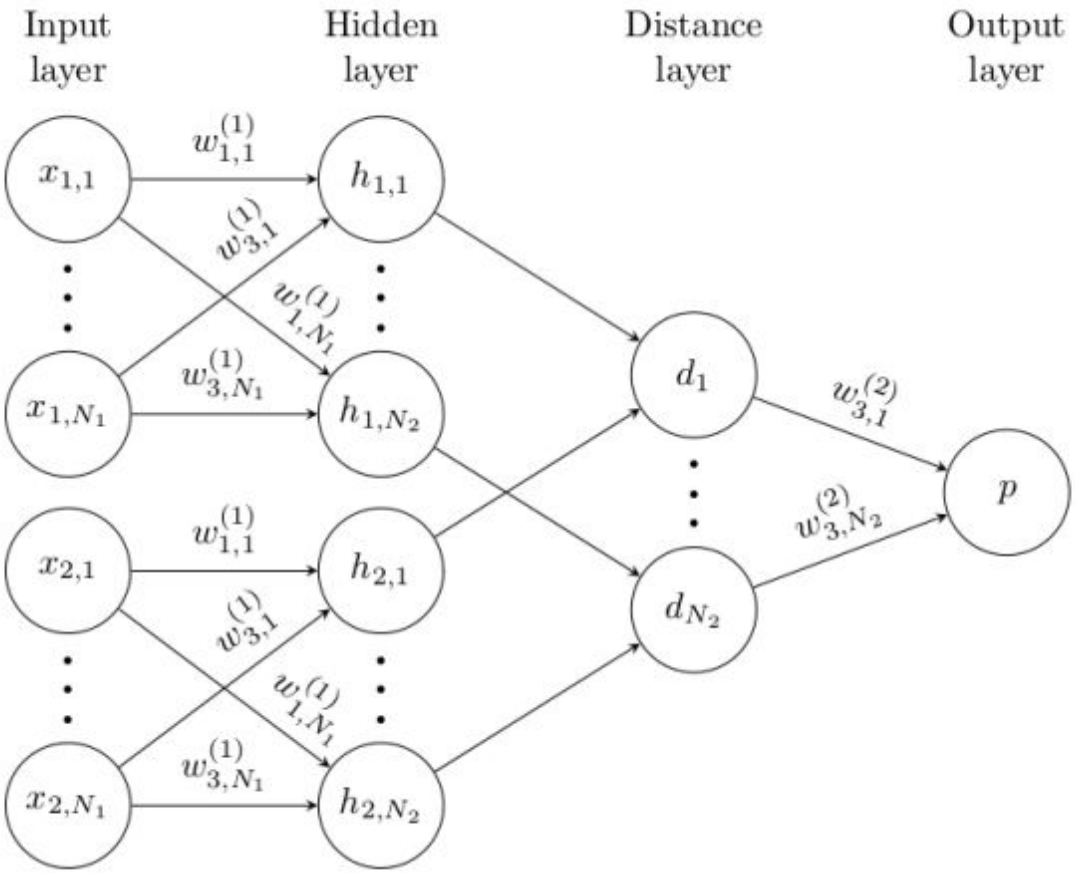
- Transfers some of the knowledge from other pre-trained NN, but may “weaken” its pre-trained objective-specific bias.
- As we learn the representation in an unsupervised manner, we try to capture patterns in the whole data set.
- The representation obtained from the clustering tends to be robust to noise in the labeled data set.

Democratic Co-learning



- Learners should differ by architecture.
- Several criteria to add an artificially labeled sample:
  - The majority of learners agree.
  - Total confidence of the majority group is greater than the confidence of the remaining models.
- Repeat until the convergence.

Commonly used architecture



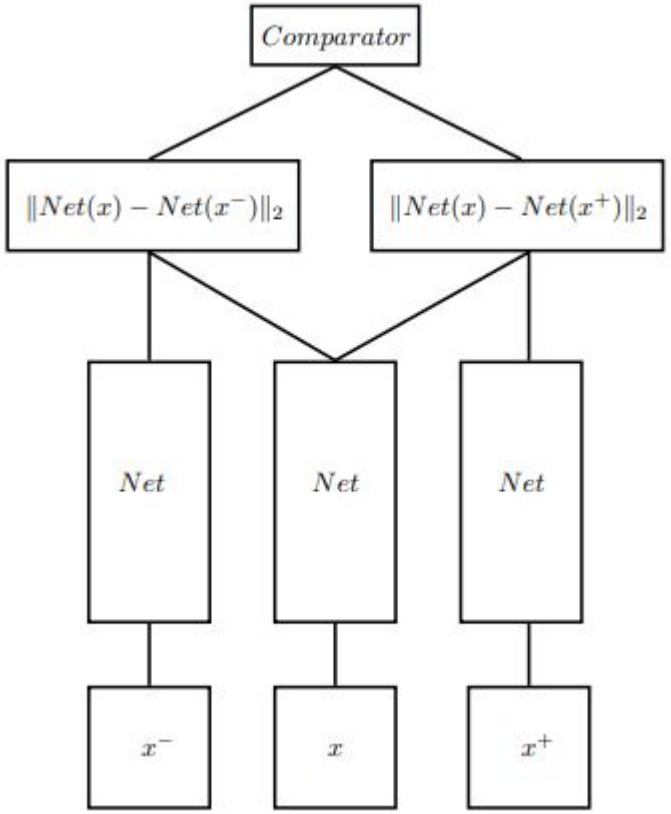
Source: Siamese Neural Networks for One-shot Image Recognition

- Shared weights across the first few blocks of the network.
- Symmetrical groups of disjoint neurons.
- Introduces the same method of embedding for multiple inputs.

Metric learning

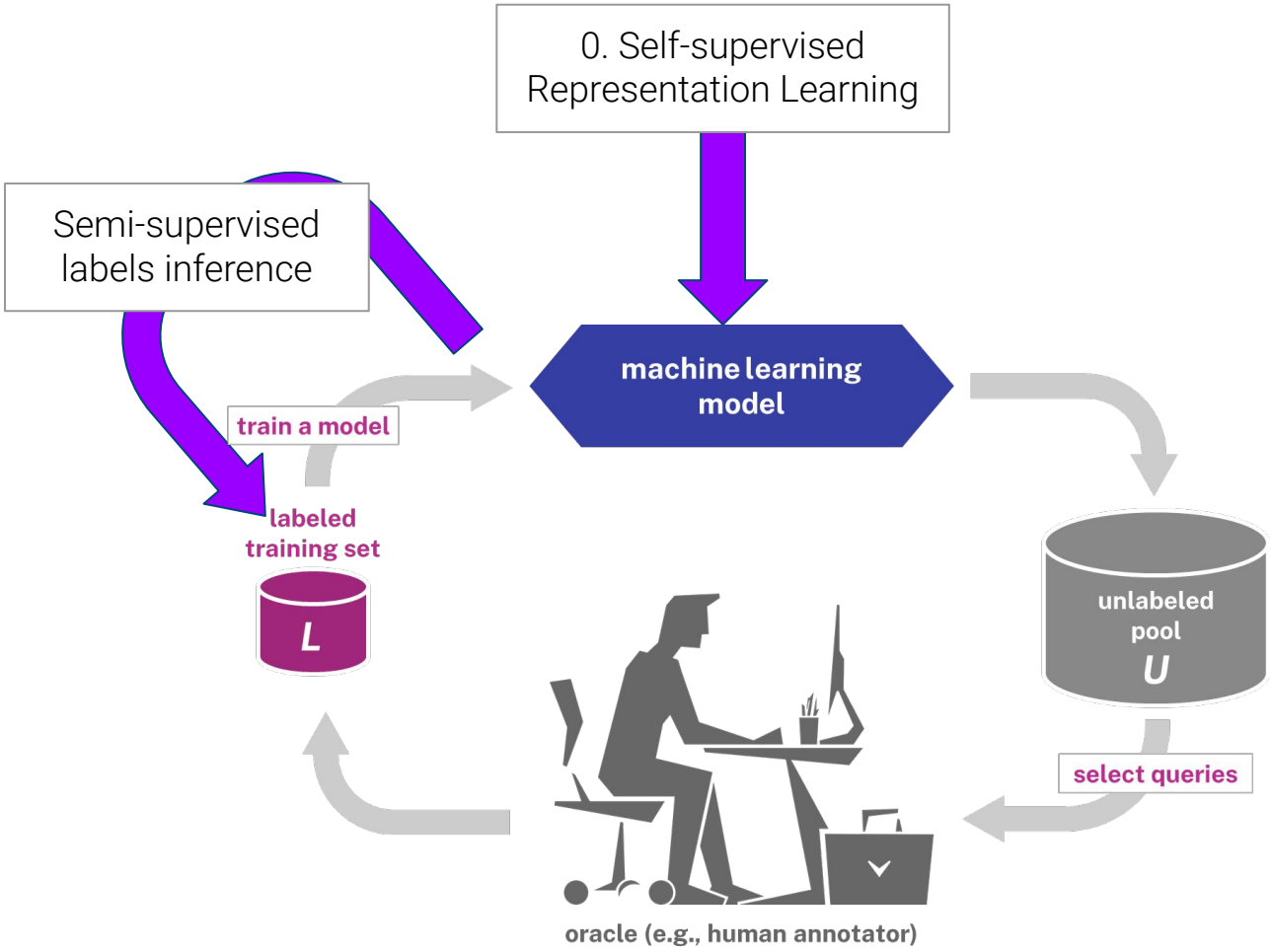
- Siamese like network - shared weights
- x^+ denotes example from the same class as x, we want the distance of embeddings x and x^+ to be small
- x^- is an example from other class than x, we want to maximize the distance
- Loss is proportional to d_+^2 where:

$$d_+ = \frac{e^{||Net(x)-Net(x^+)))||_2}}{e^{||Net(x)-Net(x^+)))||_2} + e^{||Net(x)-Net(x^-)))||_2}}$$



Source: Hoffer et al., Deep metric learning using triplet network

Applications in a combination with AL

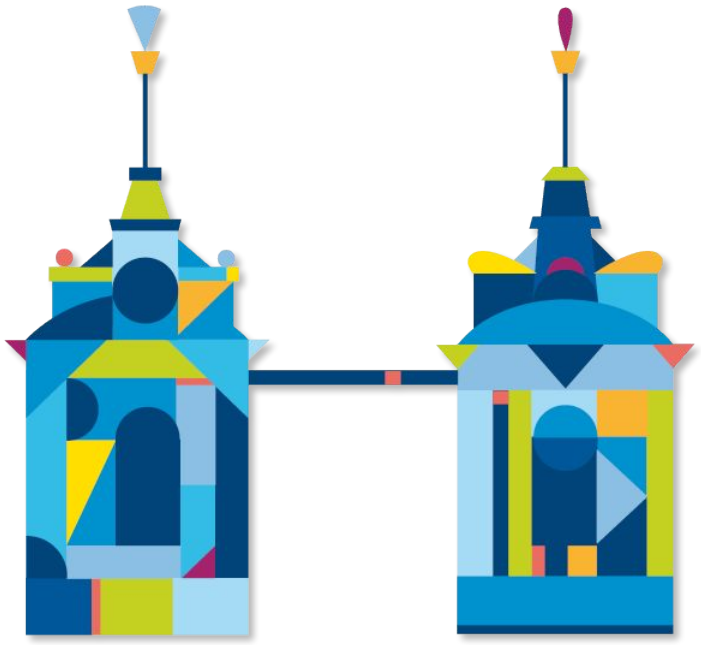


Summary

- We discussed similarities and differences between weakly-, semi-, and self-supervised learning methods.
- We considered several examples of embeddings learning methods that can be used in semi-supervised learning scenarios.
- We talked about exemplary “artificial” tasks for self-supervised learning using different data modalities (i.e., texts, images, video).
- We talked about a popular approach to prediction model co-training.

Literature:

1. P. Singh, N. Komodakis: Improving Recognition of Complex Aerial Scenes Using a Deep Weakly Supervised Learning Paradigm. IEEE Geosci. Remote. Sens. Lett. 15(12): 1932-1936, (2018).
2. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean: Distributed Representations of Words and Phrases and Their Compositionality. In: Proceedings of NIPS 2013, Volume 2, pp. 3111–3119, (2013).
3. L. Jing and Y. Tian: Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey,. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 11, pp. 4037-4058, (2021).
4. I. Misra, C.L. Zitnick, M. Hebert: Shuffle and learn: unsupervised learning using temporal order verification. European Conference on Computer Vision (ECCV), 527-544, (2016).
5. R. Arandjelović, A. Zisserman: Look, Listen and Learn. IEEE International Conference on Computer Vision (ICCV), (2017).
6. X. Yan, I. Misra, A. Gupta, D. Ghadiyaram, D. Mahajan: ClusterFit: Improving Generalization of Visual Representations. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6508-6517, (2020).
7. J.E. van Engelen, H.H. Hoos: A survey on semi-supervised learning. Mach. Learn. 109, 373–440. (2020).
8. Y. Zhou, S. Goldman: Democratic co-learning. 16th IEEE International Conference on Tools with Artificial Intelligence, pp. 594-602, (2004).
9. E. Hoffer, N. Ailon: Deep Metric Learning Using Triplet Network. In: Feragen, A., Pelillo, M., Loog, M. (eds.) Similarity-Based Pattern Recognition. SIMBAD 2015. LNCS, vol. 9370 (2015).
10. I. Misra, L. van der Maaten: Self-Supervised Learning of Pretext-Invariant Representations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6707-6717, (2020).



QUESTIONS OR COMMENTS?

a.janusz@mimuw.edu.pl

or

d.kaluza@mimuw.edu.pl