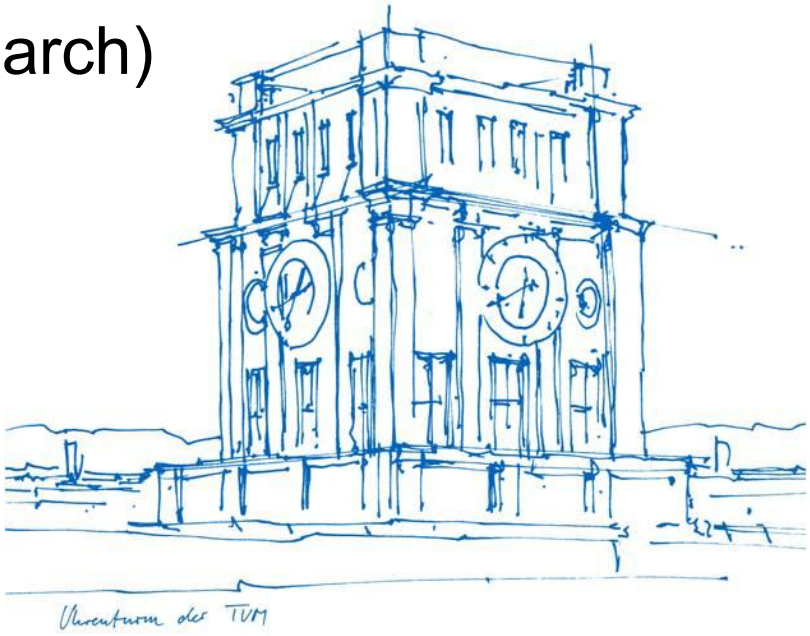
Learning Scene Representation with Knowledge Distillation from Consecutive Frames (Guided Research)

Author: Theodor Stoican

Supervisor: MSc. Shun-Cheng Wu

Technical University of Munich

Munich, 3rd November 2022

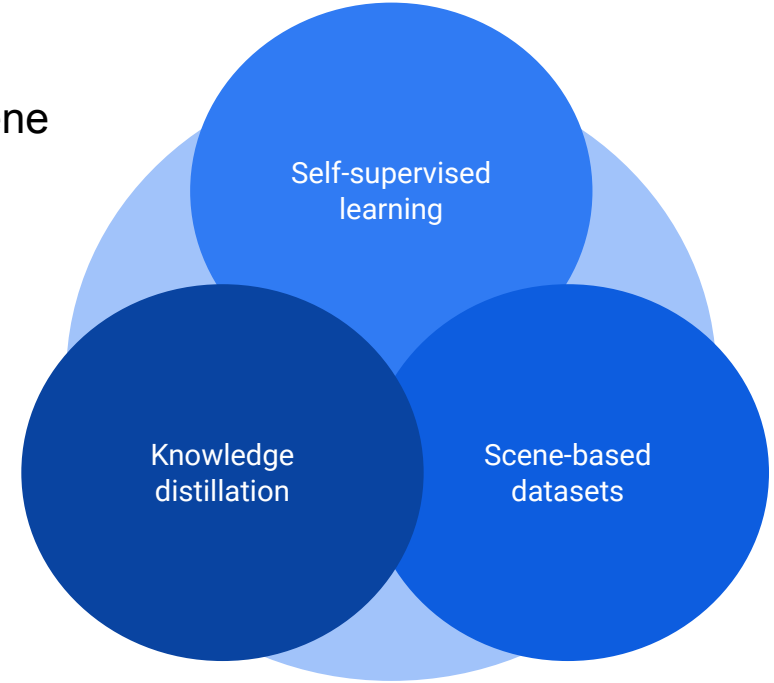


Outline

1. Problem Definition
2. Datasets
3. Contrastive Learning as Self-supervised Learning
4. Literature Review
5. Visualizing a bare ViT
6. DINO
    a. Fine-tuning DINO
7. ODIN
8. BYOL
9. Evaluation on a Single-object Dataset
10. Extention to Multi-object Datasets
11. Conclusion
12. References

Problem Definition

- Make use of the 3 on the right to generate scene representations
  - no supervision
  - some knowledge distillation to facilitate learning
  - (naturally) consecutive video frames of scenes wi th objects



Datasets

- CO3D (Common Objects in 3D)
  - multiple consecutive video frames of the same object
  - segmentation masks provides as g.t.

Sample image from CO3D

- TUM RGB-D SLAM dataset
  - multiple consecutive video frames of multiple objects
  - obtained using a Kinect sensor

Contrastive Learning as Self-supervised Learning

- A technique for having two networks learn from each other
  - typically, the input is augmented twice
  - each model sees only one version
  - the features output by the models should be similar
- Requires 2 (or more) networks

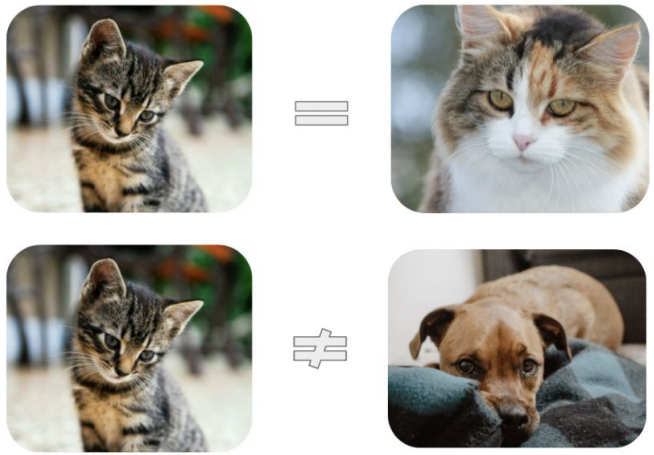


Figure from:
©https://towardsdatascience.com/understanding-contrastive-learning-d5b19fd96607

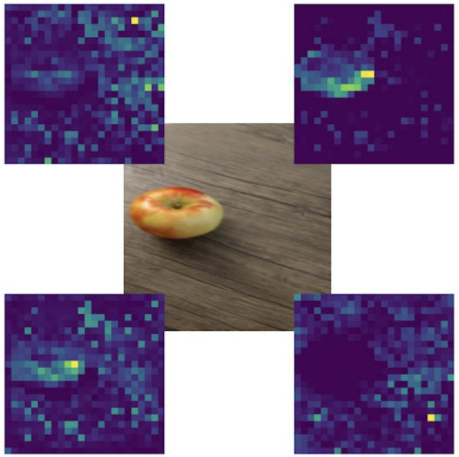
Literature Review

- 3 (applicable to our use-case) models were found
  - ODIN (**O**bject **di**scovery and representation **n**etworks) [3]
  - DINO (Self-**di**stillation with **no** labels) [1]
  - BYOL (**B**ootstrap **y**our **o**wn **l**atent) [4]

- All of them based on contrastive learning
  - made of multiple (2 or more) networks
  - input is transformed differently for each network
  - (implicit **knowledge distillation**)

Visualizing a Bare ViT

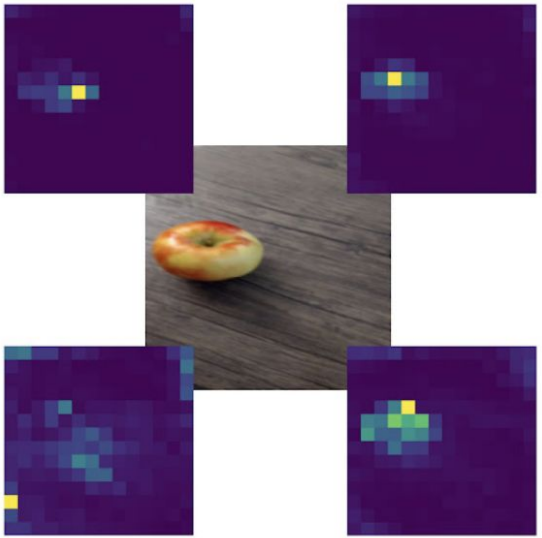
- To understand the potential of these models
  - attempt to visualize the last self-attention layer (corresponding to the [CLS] token) of a bare ViT - [2] - (pretrained on ImageNet) - **feature map** extraction
  - the contours of the objects are visible
- Limitation
  - very noisy



4 (from the 12) self-attention heads of the last layer on a CO3D sample

DINO

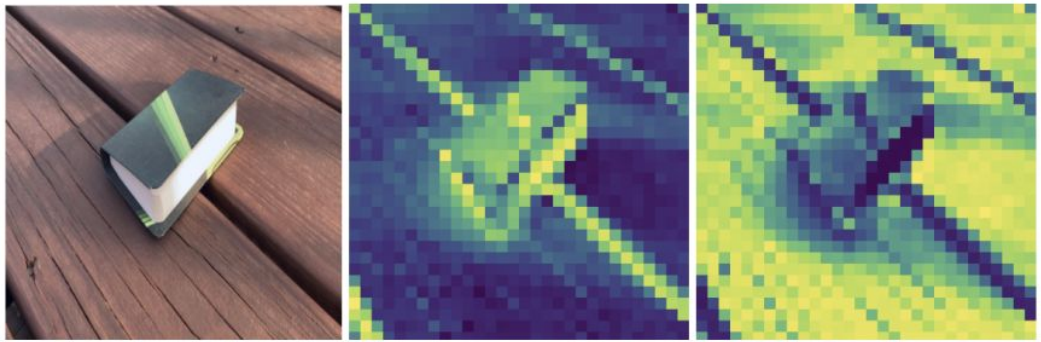
- Took a pretrained model (on ImageNet) from HuggingFace
- Extracted the **feature map** identically
- Ran it directly on CO3D
  - clear object contours
  - less noisy than a bare ViT



4 (from the 12) self-attention heads of the last layer on a CO3D sample

Fine-tuning DINO

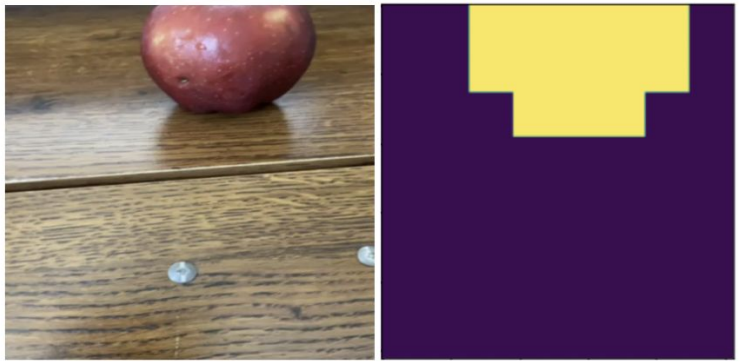
- Took this a step further
  - (ideally) each attention head should focus on an object
  - force first two heads focus on the **foreground** and **background**
  - use the segmentation masks as ground truth



Left: Sample from CO3D. **Middle**: Foreground object. **Right**: Background.

ODIN

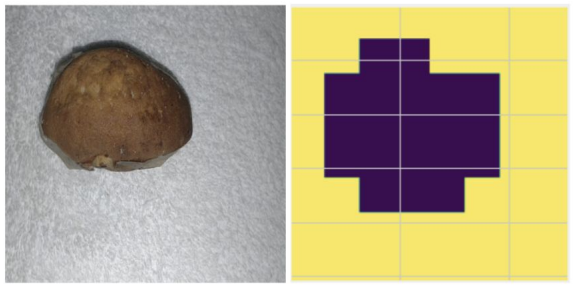
- Reimplemented ODIN from scratch
  - switched from ViT to ResNet-18 as backbone
  - **feature map:** apply K-Means on top of the last layer
  - obtained best results by training on images of res. 256 x 256



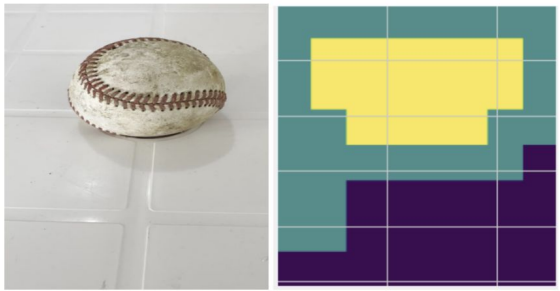
Sample from ODIN, trained with images of resolution 256 x 256

BYOL

- An ancestor of both BYOL and DINO
- Implementation is public, used it to fine-tune on CO3D
  - **feature map:** apply K-Means on top of the last layer
  - resolution of images: 512 x 512
- **Limitation** - K-Means with k=2 does not always detect the objects
  - used K-Means with knee locator to fix this



K-Means (k=2) on top of BYOL representation



K-Means (with knee locator, k=3), on top of BYOL representation

Evaluation on a Single-object Dataset

- A comparison of the 3 models on the CO3D dataset has been made
  - metric: IoU (intersection over union)
  - trained on inputs with different resolutions
  - PCA applied on top of some models' representations

	IoU score on CO3D
ODIN 256	78.6
ODIN 512	59.08
BYOL 256	66.58
BYOL 256 + PCA	66.58
BYOL 512	78.01
BYOL 512 + PCA	78.01

A table with a comparison of all models, with various configurations.

Extension to Multi-object Datasets

- Some models may be difficult to adapt to multi-object scenarios
  - **DINO** is too computationally expensive (uses ViTs)
  - **ODIN** needs a significant adjustment of implementation
- Out of the box, only **BYOL**
  - after fine-tuning, promising results
  - no formal evaluation though

K-Means (with knee locator) on top of BYOL representation.

Conclusion

- Self-supervised SOTA models are very powerful
  - scene representation is learned as a side effect
- Fine-tuning on single-object datasets
  - yields very good results even with less-powerful models
- Fine-tuning on multi-object datasets
  - simpler models (BYOL) provide promising results
  - further implementation work for more complex models (ODIN)

References

[1] Mathilde Caron et al. “Emerging properties in self-supervised vision transformers”. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, pp. 9650–9660.

[2] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: arXiv preprint arXiv:2010.11929 (2020).

[3] Olivier J H´enaff et al. “Object discovery and representation networks”. In: arXiv preprint arXiv:2203.08777 (2022).

[4] Jean-Bastien Grill et al. “Bootstrap your own latent-a new approach to self-supervised learning”. In: Advances in neural information processing systems 33 (2020), pp. 21271– 21284