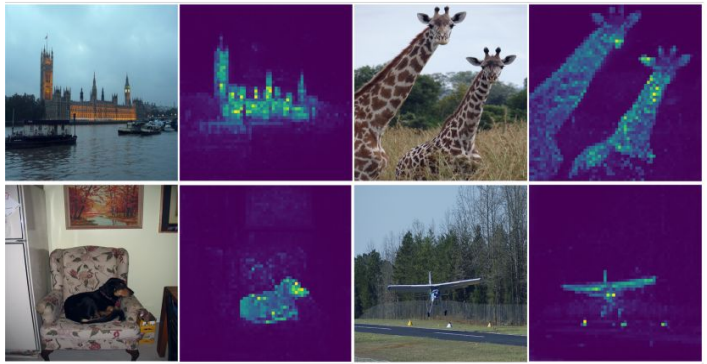
Distillation of the transformer into the CNN for fine-grained image recognition

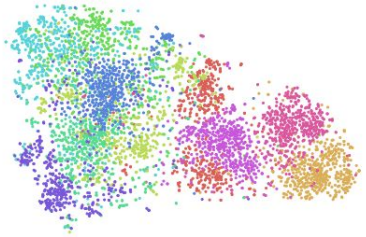
Mateusz Pach Michał Wronka

June 15, 2023

Motivation

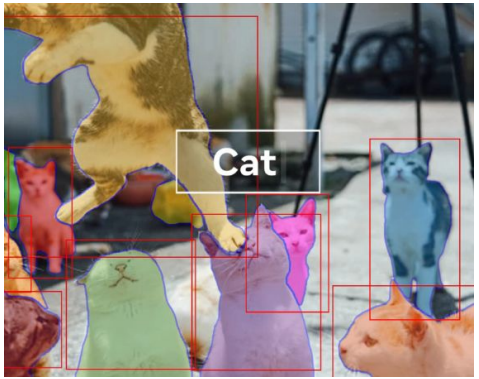
Transformers are powerful…





ResNet50 (DINO)

ViT (DINO)



Emerging Properties in Self-Supervised Vision Transformers (Caron et al., 2021)
Generalized Category Discovery (Vase et al., 2022)
Segment Anything (Kirillov et al., 2023)

Motivation

…but we like ResNets as they

can work with less parameters,

are based on the convolution,

and we know them better.

Motivation

What are the limits of the distillation which can be used in

Model compression

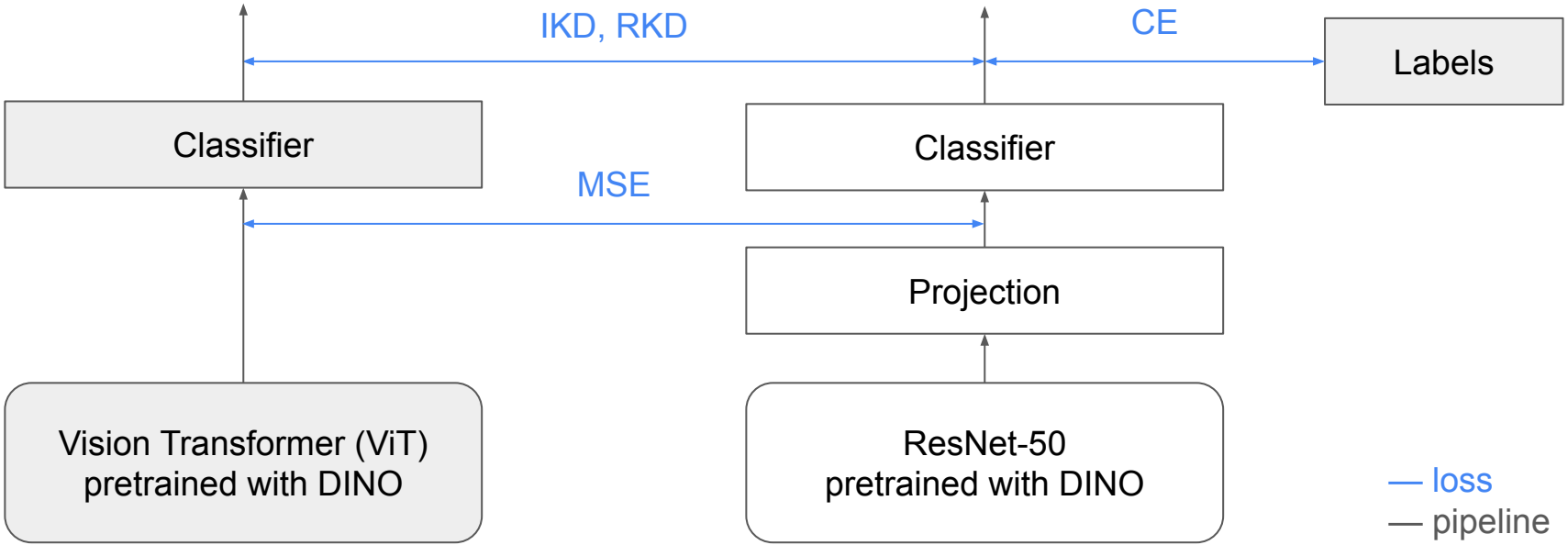
Few-shot learning

Continual learning

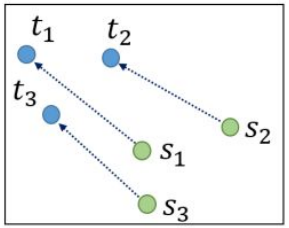
Transfer learning

and more.

Method

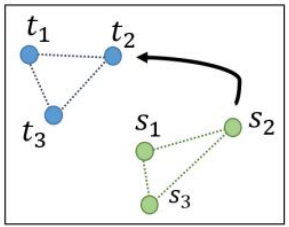

IKD, RKD
CE
Labels
Classifier
Classifier
MSE
Projection
Vision Transformer (ViT)
pretrained with DINO
ResNet-50
pretrained with DINO
loss
pipeline

Method: knowledge distillation



Point to Point

$$\mathcal{L}_{\text{IKD}} = \sum_{x_i \in \mathcal{X}} \text{KL}\Big(\text{softmax}\big(\frac{f_T(x_i)}{\tau}\big), \text{softmax}\big(\frac{f_S(x_i)}{\tau}\big)\Big)$$



Structure to Structure

$$\mathcal{L}_{\text{RKD}} = \sum_{(x_1,..,x_n) \in \mathcal{X}^N} l\big(\psi(t_1,..,t_n), \psi(s_1,..,s_n)\big)$$

Relational Knowledge Distillation (Park et al., 2019)

Datasets

Distillation: CUB-200-2011

Evaluation: CUB-200-2011 and CIFAR-10

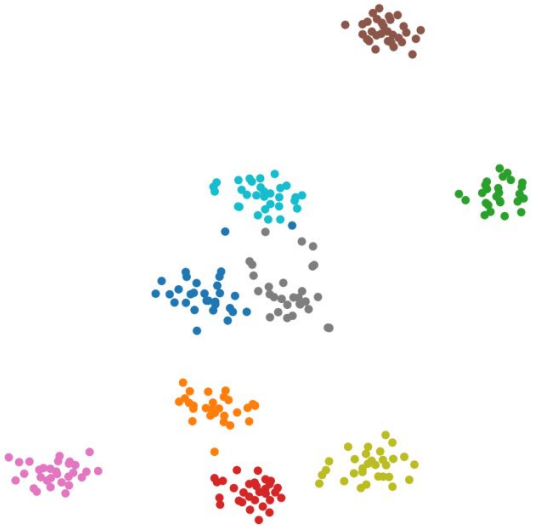
The Caltech-UCSD Birds-200-2011 Dataset (Wah et al., 2011)
Learning multiple layers of features from tiny images (Krizhevsky, 2009)

Results: accuracy

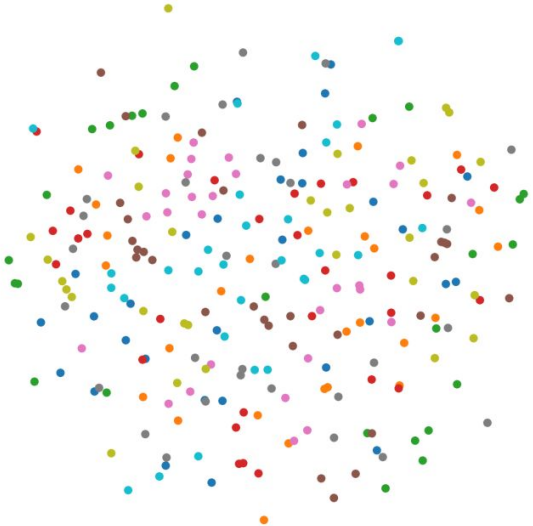
	Labels CE	Embedding MSE	IKD	RKD	Accuracy in %
A		✓		✓	0.8
B			✓	✓	69.9
C		✓	✓	✓	67.8
D	✓				77.7
E	✓	✓	✓		71.9
F	✓	✓		✓	74.0
G	✓		✓	✓	74.4
H	✓	✓	✓	✓	70.9
O	Teacher ViT				65.2

Table 1: Accuracy on CUB-200-2011.

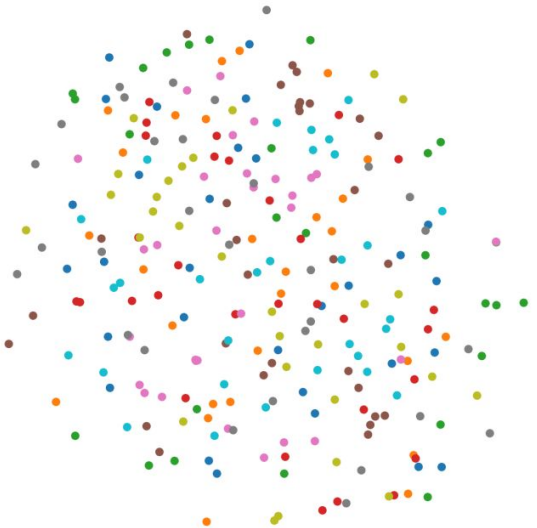
Results: clustering of CUB-200-2011



O (Teacher ViT)



D (Labels CE)



B (IKD+RKD)

Results: clustering of CIFAR-10

O (Teacher ViT)

D (Labels CE)

B (IKD+RKD)

Conclusions

- Distillation of the transformer’s knowledge into the CNN is explored.
- It is shown that the CNN can be trained with success using previously trained ViT and no labels.
- Experiments in multiple setups fail to transfer the embedding structure suggesting it may not be trivial with distillation or it is an architecture attribute.

Questions