On Few-Annotation Learning and Non-Linearity in Deep Neural Networks

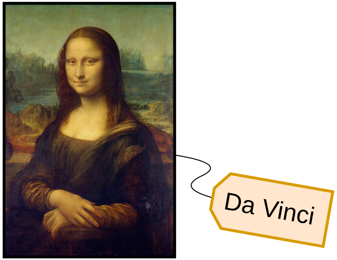
Quentin Bouniot

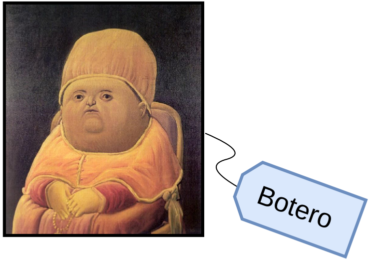
December 20, 2023

Outline

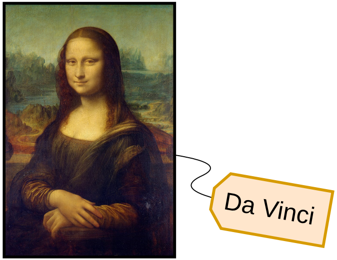
1. Introduction
2. Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning
3. Proposal-Contrastive Pretraining for Object Detection from Fewer Data
4. Understanding Deep Neural Networks Through the Lens of their Non-Linearity
5. Perspectives

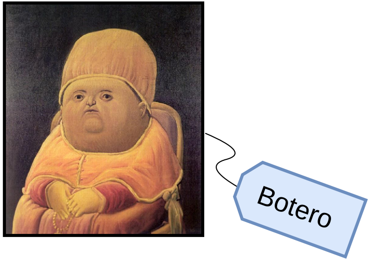
A Simple Problem ...

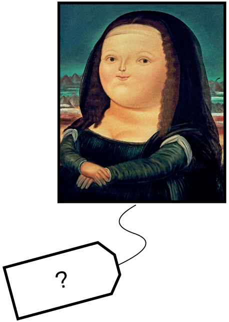




A Simple Problem ...

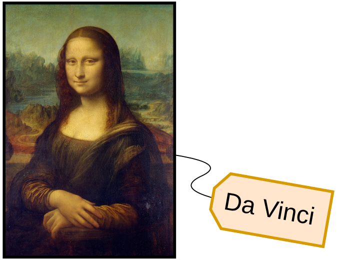


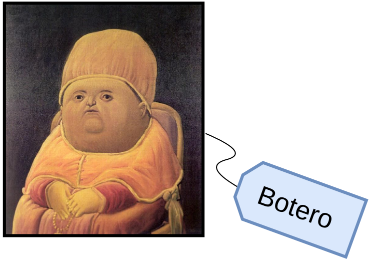


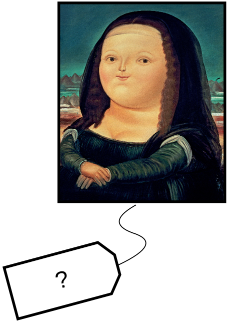


Who is the painter ?

A Simple Problem ... for a Human !



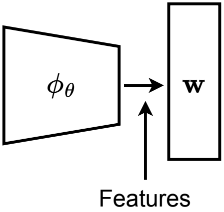


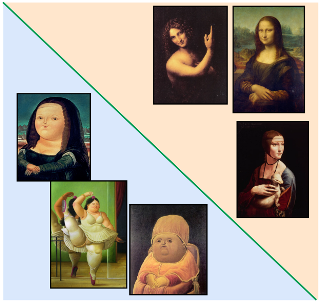


Who is the painter ?

- *Human* capacity to learn from few examples

Image Classification





- ϕ encoding function parametrized by θ
- Linear classifiers $\mathbf{w}$ (green line) separate each class

Learning from images

$$\mathcal{D}_{train} := \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\} \sim P(\mathbf{X}, \mathbf{Y})$$

Model parameters

Data points

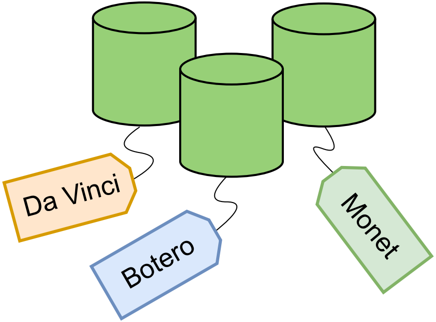
$$\boxed{\hat{\theta}, \hat{\mathbf{w}}} := \arg\min_{\theta, \mathbf{w}} \sum_{i=1}^{N} \boxed{\mathcal{L}} (\mathbf{y}_i , \mathbf{x}_i ; \theta, \mathbf{w})$$

Label

Loss function

- Learn parameters $\hat{\theta}$ and $\hat{\mathbf{w}}$ minimizing loss function $\mathcal{L}$ given data points $\mathbf{x}_i$ and labels $\mathbf{y}_i$.

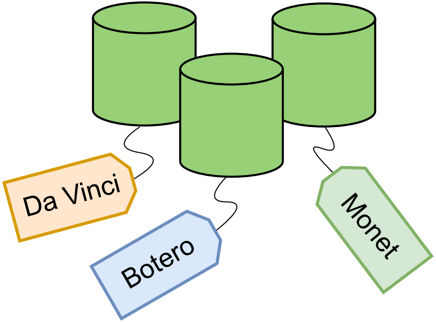
Practical Data Conditions



Expectations

- Many-Shot Learning: A lot of data and labels

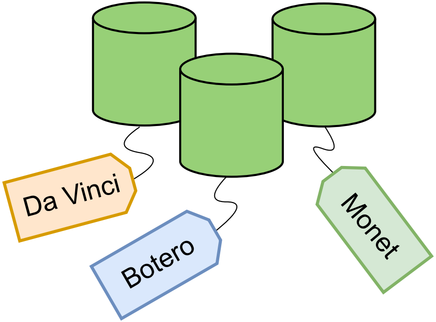
Practical Data Conditions



Expectations

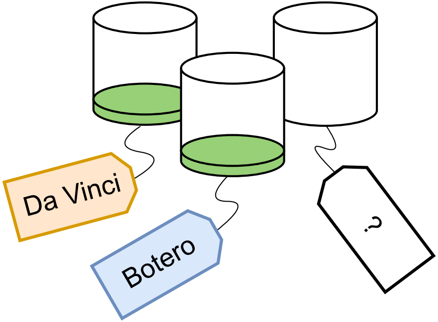
- Many-Shot Learning: A lot of data and labels
- But labeling data is costly !

Practical Data Conditions



Expectations

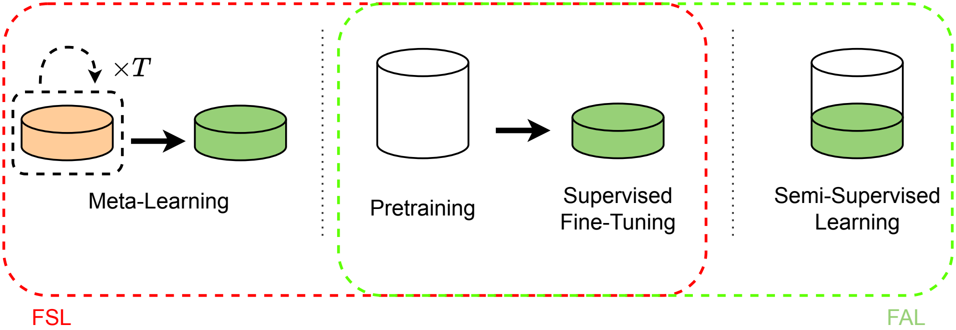
- Many-Shot Learning: A lot of data and labels
- But labeling data is costly !



Reality

- Few Annotation Learning (FAL): A lot of data and few labels
- Few Shot Learning (FSL): Few data and labels

General Frameworks



Outline

1 Introduction

2 Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning
- Meta-Learning 101
- Multi-Task Representation Learning Theory
- From Theory to Practice

3 Proposal-Contrastive Pretraining for Object Detection from Fewer Data

4 Understanding Deep Neural Networks Through the Lens of their Non-Linearity

5 Perspectives

Terminology

Meta-Learning 101

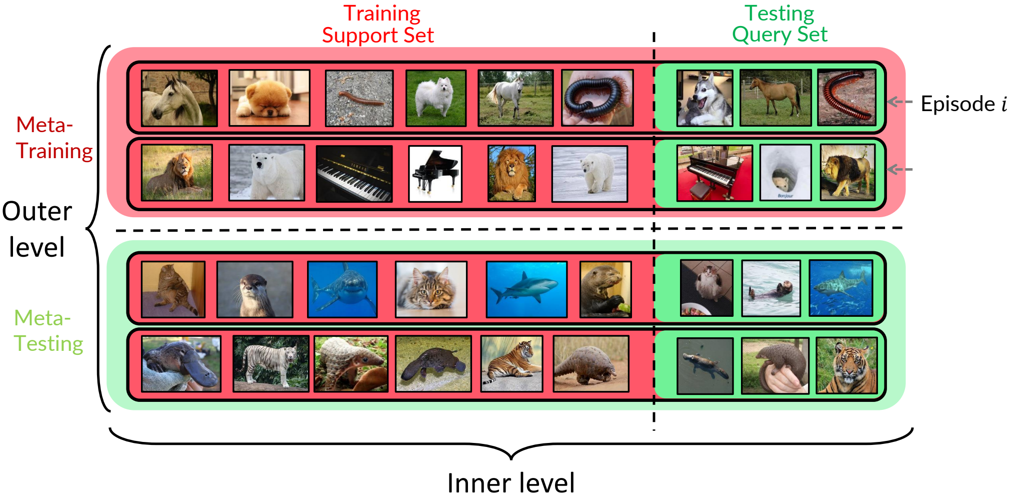
What is Meta-Learning ?

- Meta-Training: solve a set of *source tasks*.
- Meta-Testing: use knowledge from meta-training to solve *previously unseen tasks* more efficiently.

How is it related to Few-Shot Learning ?

The Meta-learner *learns to learn* a new task with few shots.

Introducing episodes
Meta-Learning 101



N-way k-shot episode: task with N different classes and k images for each class.

Meta-Learning Problem Formulation

Meta-Learning 101

Data distributions:

Drawing N episodes

$$\forall t \in [1, \dots, N], \qquad \mathcal{T}_t \sim P(\mathcal{T}), \qquad \mathcal{T}_t := \mathcal{S}_t \cup \mathcal{Q}_t$$

Support sets

Query sets

Meta-Learning Problem Formulation

Meta-Learning 101

Data distributions:

Drawing N episodes

$$\forall t \in [1, \dots, N], \qquad \mathcal{T}_t \sim P(\mathcal{T}), \qquad \mathcal{T}_t := \mathcal{S}_t \cup \mathcal{Q}_t$$

Support sets

Query sets

Inner-level:

Inner loss function

$$\boxed{\hat{\theta}_t, \hat{\mathbf{w}}_t} = \arg\min_{\theta, \mathbf{w}} \sum_{(x,y) \in \mathcal{S}_t} \boxed{\mathcal{L}_{\text{inner}}} (x, y; \theta, \mathbf{w})$$

Parameters specialized to each episode

Meta-Learning Problem Formulation

Meta-Learning 101

Data distributions:

Drawing N episodes

$$\forall t \in [1, \ldots, N], \qquad \mathcal{T}_t \sim P(\mathcal{T}), \qquad \mathcal{T}_t := \mathcal{S}_t \cup \mathcal{Q}_t$$

Support sets

Query sets

Inner-level:

Inner loss function

$$\hat{\theta}_t, \hat{\mathbf{w}}_t = \arg\min_{\theta, \mathbf{w}} \sum_{(x,y) \in \mathcal{S}_t} \mathcal{L}_{\text{inner}}(x, y; \theta, \mathbf{w})$$

Parameters specialized to each episode

Outer-level:

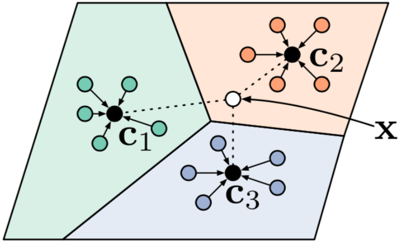
Initialization for new sets of episodes

Task-specific parameters learned

$$\hat{\theta}, \hat{\mathbf{w}} = \arg\min_{\theta, \mathbf{w}} \sum_{t=1}^{N} \sum_{(x,y) \in \mathcal{Q}_t} \mathcal{L}_{\text{outer}}(x, y; \hat{\theta}_t, \hat{\mathbf{w}}_t)$$

Outer loss function

Metric-based methods (ProtoNet [1])



- Support samples for each class i fused into **prototypes** $\mathbf{c}_i$.
- Probability distribution using **inverse of distances** to prototypes.

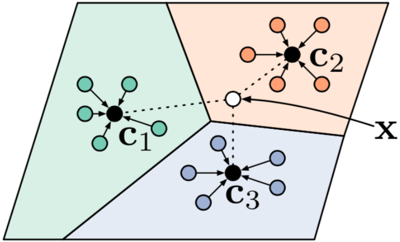
[1] Jake Snell, Kevin Swersky, and Richard S. Zemel. "Prototypical Networks for Few-shot Learning". In: *NeurIPS*. 2017

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *ICML*. 2017

Meta-Learning methods

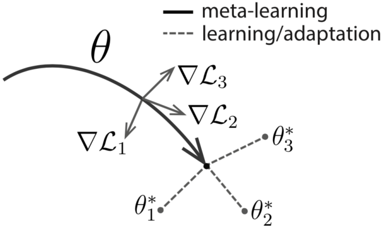
Meta-Learning 101

Metric-based methods (ProtoNet [1])



- Support samples for each class i fused into **prototypes** $\mathbf{c}_i$.
- Probability distribution using **inverse of distances** to prototypes.

Gradient-based methods (MAML [2])



- **End-to-end** bi-level optimization through **gradient descent**.

[1] Jake Snell, Kevin Swersky, and Richard S. Zemel. "Prototypical Networks for Few-shot Learning". In: *NeurIPS*. 2017

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *ICML*. 2017

Introduction to MTR

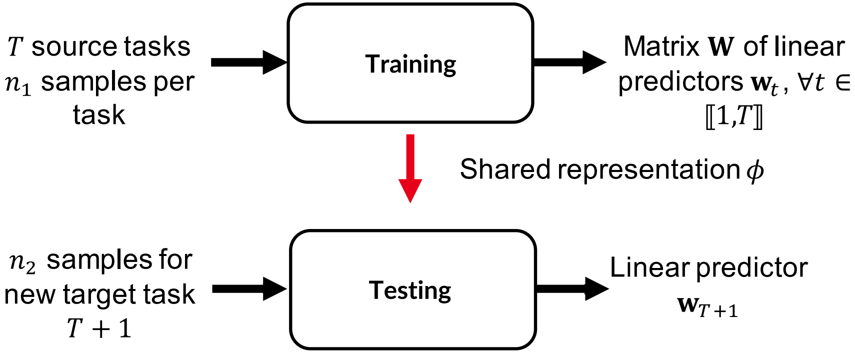
Multi-Task Representation Learning Theory

T source tasks
n_1 samples per task

→ **Training** →

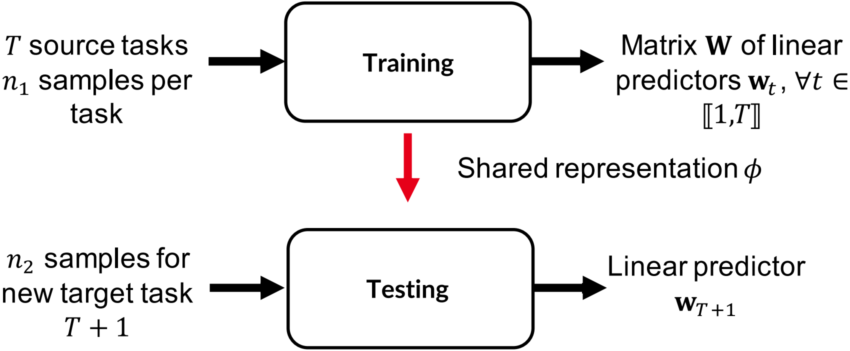
Matrix $\mathbf{W}$ of linear predictors $\mathbf{w}_t$, $\forall t \in [\![1, T]\!]$

Introduction to MTR
Multi-Task Representation Learning Theory



Introduction to MTR

Multi-Task Representation Learning Theory

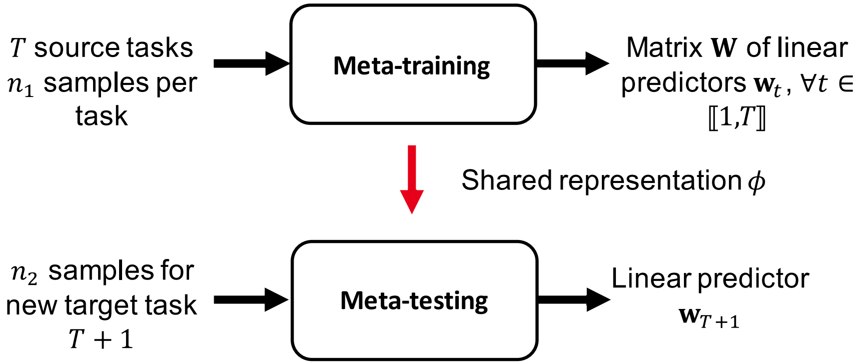


Goal: Minimize **excess risk** ER $= \mathcal{L}(\hat{\phi}, \hat{\mathbf{w}}_{T+1}) - \mathcal{L}(\phi^*, \mathbf{w}^*_{T+1})$,

- **True risk $\mathcal{L}$**
- **Optimal representation ϕ^***
- $\mathbf{w}^*_{T+1}$ **ideal target linear predictor.**

Link with Meta-Learning

Multi-Task Representation Learning Theory



Goal: Minimize **excess risk** ER $= \mathcal{L}(\hat{\phi}, \hat{\mathbf{w}}_{T+1}) - \mathcal{L}(\phi^*, \mathbf{w}^*_{T+1})$,

- **True risk $\mathcal{L}$**
- **Optimal representation ϕ^***
- $\mathbf{w}^*_{T+1}$ **ideal target linear predictor.**

Few-Shot Multi-Task Learning Theory
Multi-Task Representation Learning Theory

Few-Shot Learning bound[3]

If assumptions are satisfied:

$$\mathrm{ER}(\phi, \mathbf{w}_{T+1}) \leq O\left(\frac{1}{n_1 T} + \frac{1}{n_2}\right)$$

Number of samples for target task

Number of samples per source tasks

Number of source tasks

- ✓ All *source* and *target* data are useful to decrease the bound of *excess risk*.
- ✓ Increasing **either** T or n_1 have an effect on the bound.

[3] Simon S. Du et al. "Few-Shot Learning via Learning the Representation, Provably". In: *ICLR*. 2021; Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. "Provable Meta-Learning of Linear Representations". In: *arXiv*. 2020.

Important Assumptions

Multi-Task Representation Learning Theory

Assumption 1: Diversity of the source tasks[4]

$$\boxed{\text{Condition Number } \kappa(\mathbf{W}^*) = \frac{\sigma_{\text{max}}(\mathbf{W}^*)}{\sigma_{\text{min}}(\mathbf{W}^*)} \textit{ should not increase} \text{ with } T.}$$

- Optimal predictors $\mathbf{W}^* = [\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*]$ **cover** all the **directions evenly**

[4] Simon S. Du et al. "Few-Shot Learning via Learning the Representation, Provably". In: *ICLR*. 2021; Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. "Provable Meta-Learning of Linear Representations". In: *arXiv*. 2020.

Important Assumptions
Multi-Task Representation Learning Theory

Assumption 1: Diversity of the source tasks[4]

Condition Number $\kappa(\mathbf{W}^*) = \frac{\sigma_{\text{max}}(\mathbf{W}^*)}{\sigma_{\text{min}}(\mathbf{W}^*)}$ *should not increase* with T.

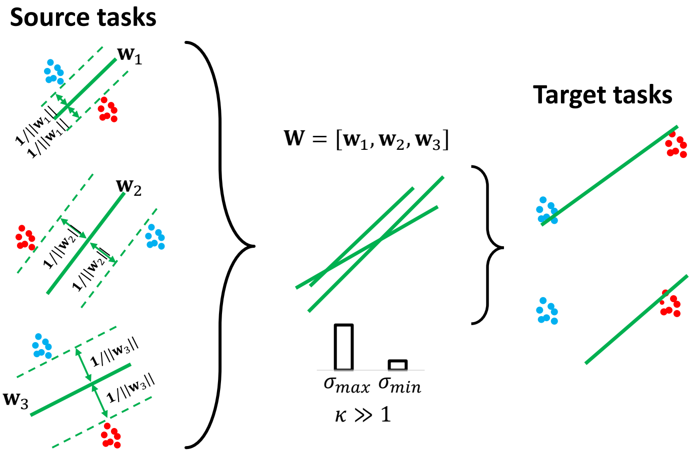
- Optimal predictors $\mathbf{W}^* = [\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*]$ **cover** all the **directions evenly**

Assumption 2: Constant classification margin[4]

Norm of predictors $\|\mathbf{w}_t^*\|_{t \in [\![1,T]\!]}$ *should not increase* with T

[4] Simon S. Du et al. "Few-Shot Learning via Learning the Representation, Provably". In: *ICLR*. 2021; Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. "Provable Meta-Learning of Linear Representations". In: *arXiv*. 2020.

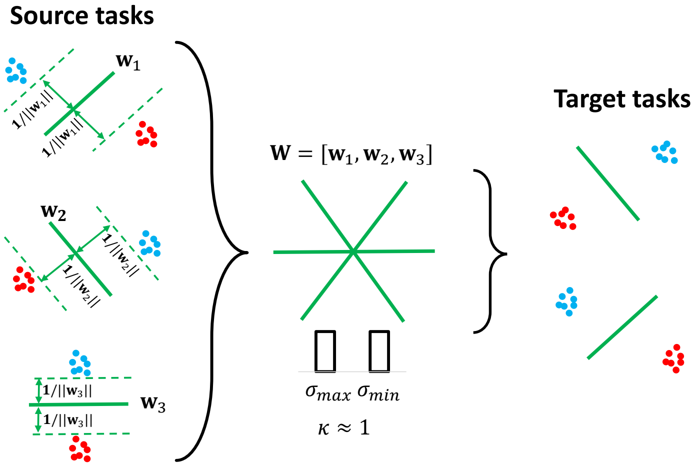
Illustration: Violated Assumptions
Multi-Task Representation Learning Theory



× Linear predictors cover **only part of the space** or **over-specialize** to the tasks

Illustration: Satisfied Assumptions

Multi-Task Representation Learning Theory



✓ **Assumption 1** makes sure that **linear predictors are complementary**

✓ **Assumption 2 avoids under- or over-specialization** to the tasks

What Happens in Practice ?
From Theory to Practice

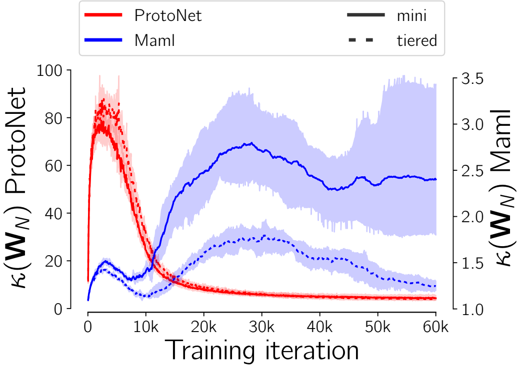
Idea:

- Verify *assumptions 1 and 2* for meta-learning algorithms.

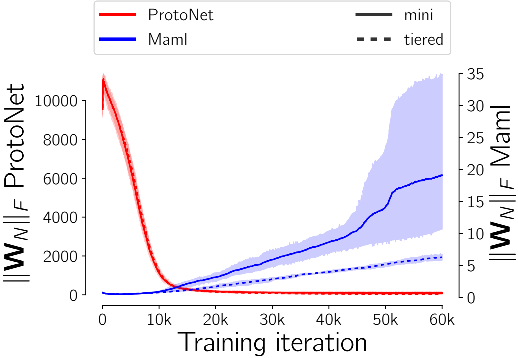
How ?

- Monitor *condition number* $\kappa(\mathbf{W}_N)$ and *norm of the predictors* $\|\mathbf{W}_N\|_F$ for the last N tasks

What Happens in Practice ?
From Theory to Practice



Monitoring the *condition number*



Monitoring the *norm*

✓ ProtoNet *naturally verifies* the assumptions

× MAML *does not verify* the assumptions

Why Does it Happen ?

From Theory to Practice

Case of ProtoNet:

- Theorem (informal)

> If all prototypes are normalized,
> then all ProtoNet encoders verify Assumption 1.

✓ Norm minimization is *enough* to obtain well-behaved condition number for ProtoNet.

Why Does it Happen ?
From Theory to Practice

Case of MAML:

- Theorem (informal)

At iteration i, if $\sigma_{\text{min}} = 0$ for last two tasks, then $\kappa(\hat{\mathbf{W}}_2^{i+1}) \geq \kappa(\hat{\mathbf{W}}_2^{i})$.

✓ The condition number for MAML can **increase** between iterations.

What can we do ?
From Theory to Practice

Ensuring Assumption 1: **Spectral** regularization

$$\boxed{\kappa(\mathbf{W}_N) = \frac{\sigma_{\text{max}}(\mathbf{W}_N)}{\sigma_{\text{min}}(\mathbf{W}_N)}}$$

✓ Regularizing with $\kappa(\mathbf{W}_N)$ leads to a better coverage of the searched space

What can we do ?
From Theory to Practice

Ensuring Assumption 1: **Spectral** regularization

$$\boxed{\kappa(\mathbf{W}_N) = \frac{\sigma_{\text{max}}(\mathbf{W}_N)}{\sigma_{\text{min}}(\mathbf{W}_N)}}$$

✓ Regularizing with $\kappa(\mathbf{W}_N)$ leads to a better coverage of the searched space

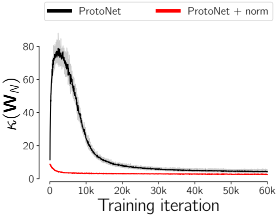
Ensuring Assumption 2: **Norm regularization or normalization** for linear predictors

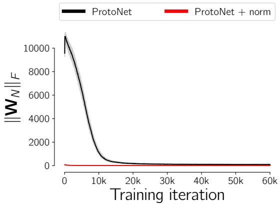
✓ **Normalizing predictors** ensure **constant margin** that **does not change** with T

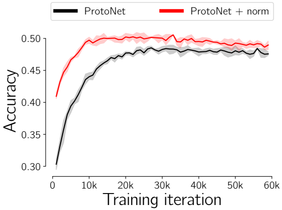
Experimental Results

From Theory to Practice

Experiments on mini-ImageNet 5-way 1-shot



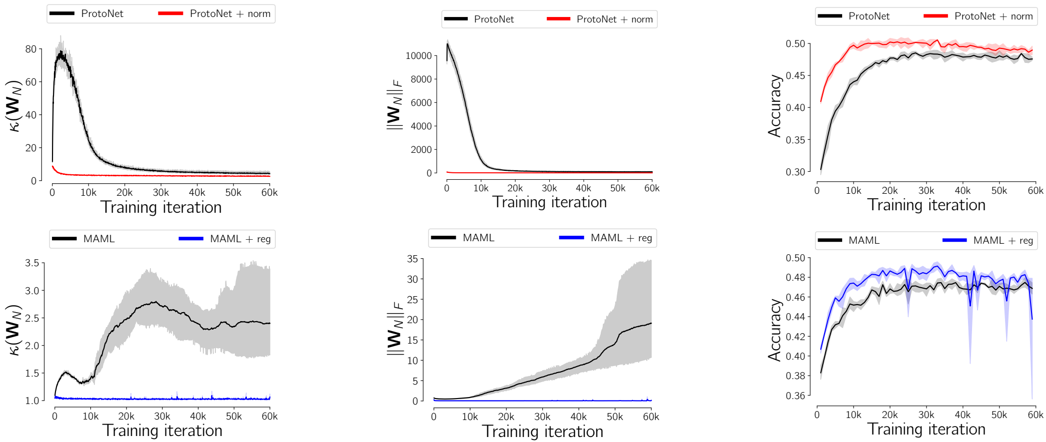




Experimental Results

From Theory to Practice

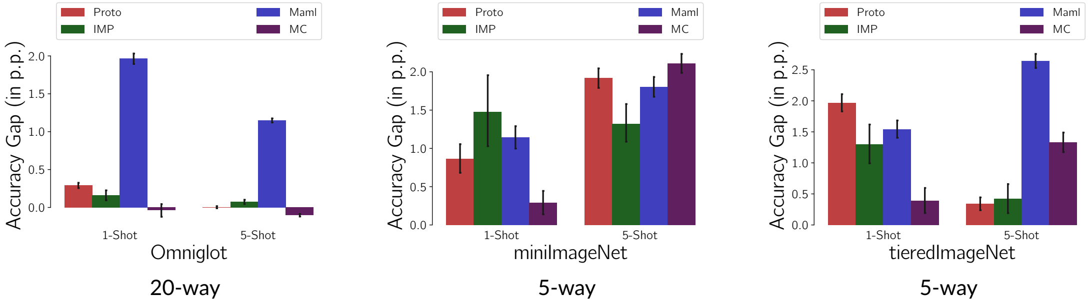
Experiments on mini-ImageNet 5-way 1-shot



✓ Our **regularization** and **normalization** have the intended effects.

Experimental Results

From Theory to Practice



- ✓ *Statistically significant* improvements with our regularization and normalization.
- ✓ *Better generalization* when the assumptions are not verified naturally.

Take Home Message I

Improving Few-Shot Learning Through Multi-Task Representation Learning Theory[5]

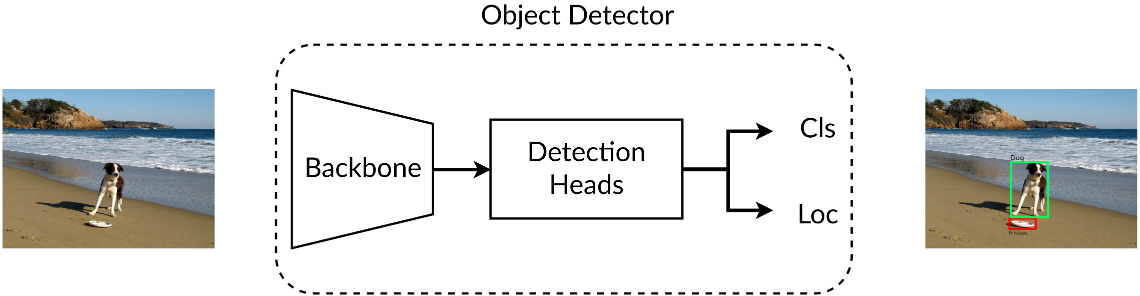
- ✓ **Connection** between Meta-Learning and Multi-Task Representation Learning Theory
- ✓ Explaining why some meta-learning methods **naturally fulfill** theoretical assumptions of the best learning bounds.
- ✓ We prove that it is possible to enforce the assumptions and propose **practical ways** which leads to **significant** performance improvements.

[5] Quentin Bouniot, Ievgen Redko, Romaric Audigier, et al. "Improving Few-Shot Learning Through Multi-task Representation Learning Theory". In: *ECCV*. 2022.

Outline

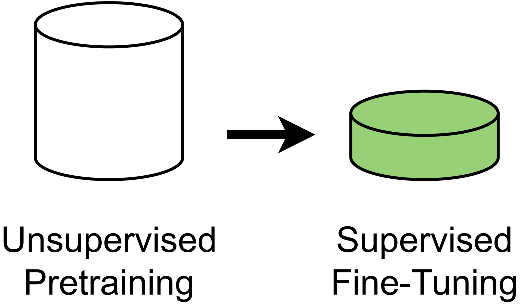
1. Introduction
2. Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning
3. Proposal-Contrastive Pretraining for Object Detection from Fewer Data
   - Motivations and Background
   - Proposal Selection Contrast (ProSeCo)
   - Experimental Results
4. Understanding Deep Neural Networks Through the Lens of their Non-Linearity
5. Perspectives

Object Detectors in a Nutshell
Motivations and Background



- Detectors composed of **backbone model** and **detection-specific heads**.
- Predict **class (Cls)** and **location (Loc)** for each objects in an image.

Setting considered
Motivations and Background

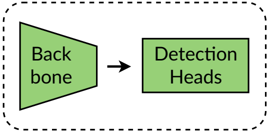


Pretraining in Object Detection

Motivations and Background

Overall Pretraining

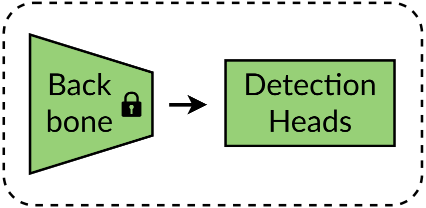
(a)[6]



✓ Consistency

✗ Costly

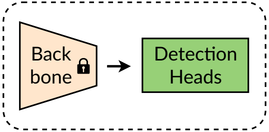
(b) Our Approach



✓ Consistency

✓ Less costly

(c)[7]



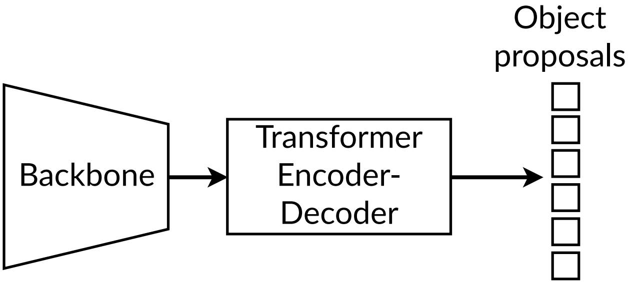
✗ Discrepancy

✓ Less costly

[6] Fangyun Wei et al. "Aligning pretraining for detection via object-level contrastive learning". In: *NeurIPS*. 2021

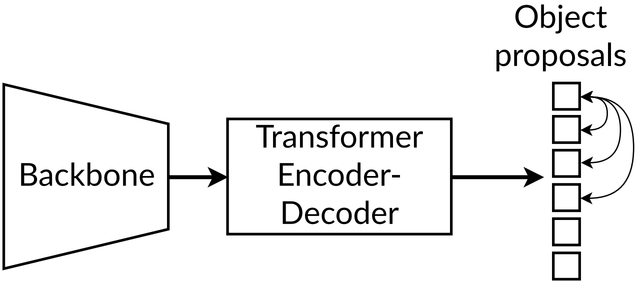
[7] Zhigang Dai et al. "Up-DETR: Unsupervised pre-training for object detection with transformers". In: *CVPR*. 2021; Amir Bar et al. "Detreg: Unsupervised pretraining with region priors for object detection". In: *CVPR*. 2022

Transformer-based Detectors
Motivations and Background



- ▶ Transformer-based detectors generates N proposals $\gg k$ objects in images.

Transformer-based Detectors
Motivations and Background

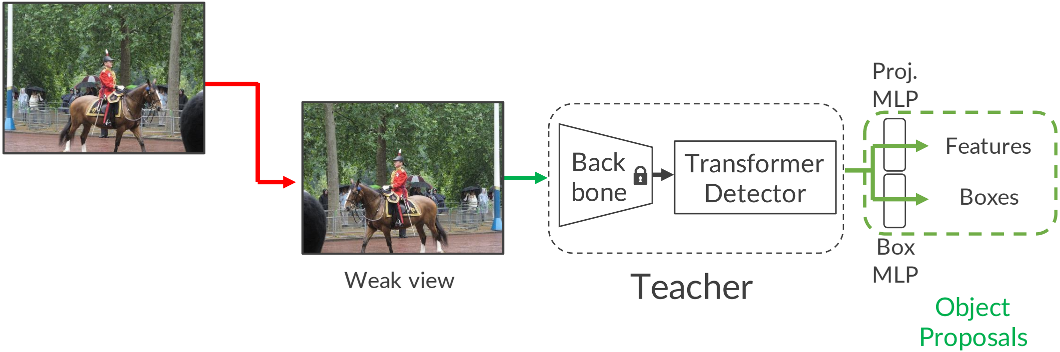


- Transformer-based detectors generates N proposals $\gg k$ objects in images.

Contribution: Contrastive learning between proposals.

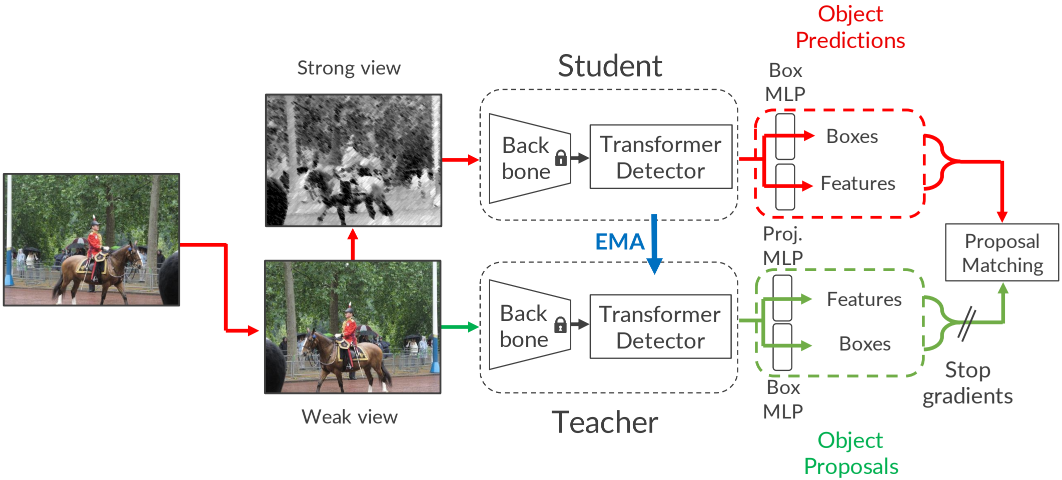
Proposal-Contrastive Learning

Proposal Selection Contrast (ProSeCo)



Proposal-Contrastive Learning

Proposal Selection Contrast (ProSeCo)



- Object Proposals from Teacher are matched with Predictions from Student.

Proposal-Contrastive Learning
Proposal Selection Contrast (ProSeCo)

Unsupervised Proposal Matching

$$\hat{\sigma}_i^{\text{prop}} = \underbrace{\arg\min_{\sigma \in \mathfrak{S}_N}}_{\text{Permutations of } N \text{ elements}} \sum_{j=1}^{N} \mathcal{L}_{\text{prop\_match}}\big(\underbrace{\mathbf{y}_{(i,j)}}_{\text{Object Proposals}}, \underbrace{\hat{\mathbf{y}}_{(i,\sigma(j))}}_{\text{Object Predictions}}\big)$$

- Proposal j found by the teacher associated to prediction $\hat{\sigma}_i^{\text{prop}}(j)$ of the student.

Proposal-Contrastive Learning

Proposal Selection Contrast (ProSeCo)

Unsupervised Proposal Matching

Object Proposals

$$\hat{\sigma}_i^{\text{prop}} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{j=1}^{N} \mathcal{L}_{\text{prop\_match}}\left(\mathbf{y}_{(i,j)}, \hat{\mathbf{y}}_{(i,\sigma(j))}\right)$$

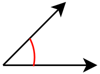
Permutations of N elements

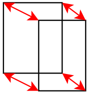
Object Predictions

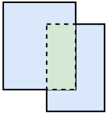
- Proposal j found by the teacher associated to prediction $\hat{\sigma}_i^{\text{prop}}(j)$ of the student.

Matching Cost $\mathcal{L}_{\text{prop\_match}}$ depends on:

- features similarity
- L_1 loss of box coordinates
- generalized IoU loss

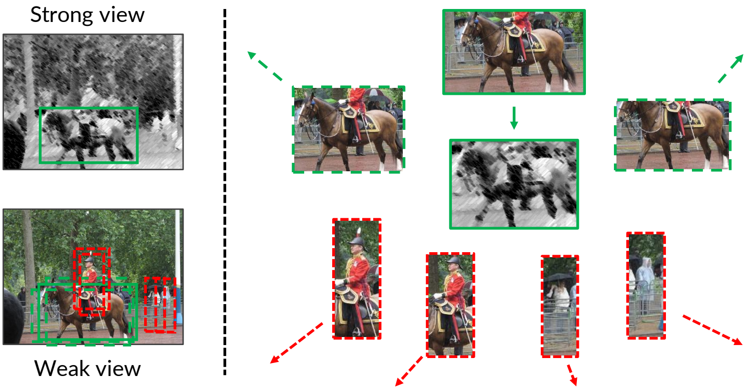






Proposal-Contrastive Learning
Proposal Selection Contrast (ProSeCo)

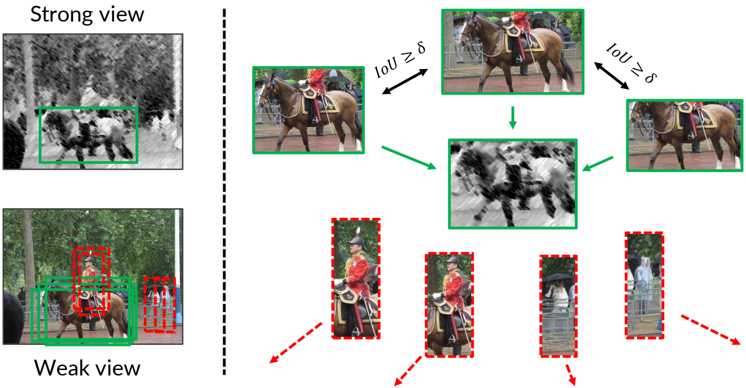
Naive way



× Close proposals considered as negative examples.

Proposal-Contrastive Learning
Proposal Selection Contrast (ProSeCo)

Localization-aware Contrastive loss



✓ Overlapping proposals are considered as positive examples.

Proposal-Contrastive Learning

Proposal Selection Contrast (ProSeCo)

Soft Contrastive Estimation (SCE) loss function[8]

Relations between proposals

Temperature

$$\boxed{p'_{(in,jm)}} = \frac{\mathbb{1}_{i \neq n} \mathbb{1}_{j \neq m} \exp(\mathbf{z}_{(i,j)} \cdot \mathbf{z}_{(n,m)} / \tau_t)}{\sum_{k=1}^{N_b} \sum_{l=1}^{N} \mathbb{1}_{i \neq k} \mathbb{1}_{j \neq l} \exp(\mathbf{z}_{(i,j)} \cdot \mathbf{z}_{(k,l)} / \tau_t)}$$

Features of Object Proposals

[8] Julien Denize et al. "Similarity contrastive estimation for self-supervised soft contrastive learning". In: *WACV*. 2023.

Proposal-Contrastive Learning

Proposal Selection Contrast (ProSeCo)

Soft Contrastive Estimation (SCE) loss function[8]

Relations between proposals

Temperature

$$p'_{(in,jm)} = \frac{\mathbb{1}_{i \neq n} \mathbb{1}_{j \neq m} \exp(\mathbf{z}_{(i,j)} \cdot \mathbf{z}_{(n,m)} / \tau_t)}{\sum_{k=1}^{N_b} \sum_{l=1}^{N} \mathbb{1}_{i \neq k} \mathbb{1}_{j \neq l} \exp(\mathbf{z}_{(i,j)} \cdot \mathbf{z}_{(k,l)} / \tau_t)}$$

Features of Object Proposals

Features of Object Predictions

$$p''_{(in,jm)} = \frac{\exp(\mathbf{z}_{(i,j)} \cdot \hat{\mathbf{z}}_{(n,m)} / \tau)}{\sum_{k=1}^{N_b} \sum_{l=1}^{N} \exp(\mathbf{z}_{(i,j)} \cdot \hat{\mathbf{z}}_{(k,l)} / \tau)}$$

Contrastive aspect between predictions and proposals

[8] Julien Denize et al. “Similarity contrastive estimation for self-supervised soft contrastive learning”. In: *WACV*. 2023.

Proposal-Contrastive Learning
Proposal Selection Contrast (ProSeCo)

Localization-aware similarity distribution

$$w^{\text{Loc}}_{(in,jm)} = \lambda_{\text{SCE}} \cdot \mathbb{1}_{i=n} \mathbb{1}_{IoU_i(j,m)\geq\delta} + (1-\lambda_{\text{SCE}}) \cdot p'_{(in,jm)}$$

IoU between proposals in same image above threshold δ

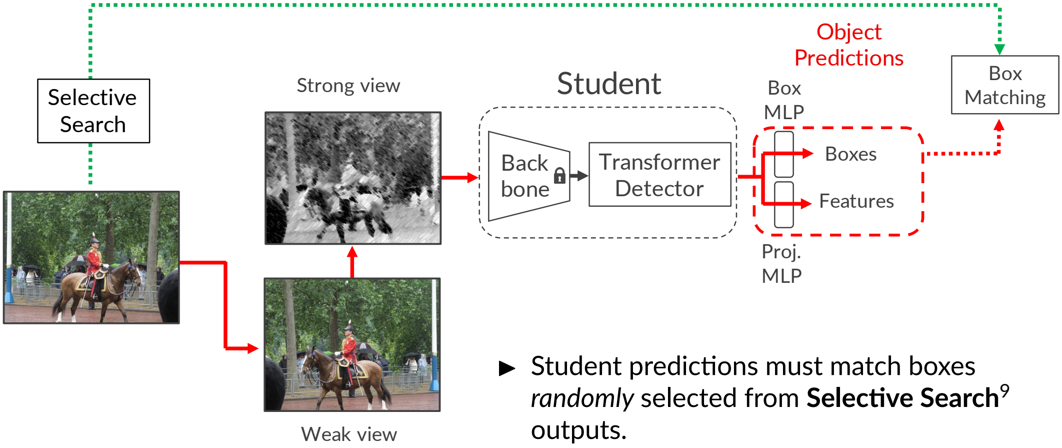
Localized SCE (LocSCE) function

$$\mathcal{L}_{\text{LocSCE}}(\mathbf{y}, \hat{\mathbf{y}}, \hat{\sigma}^{\text{prop}}) = -\frac{1}{N_b N} \sum_{i=1}^{N_b} \sum_{n=1}^{N_b} \sum_{j=1}^{N} \sum_{m=1}^{N} w^{\text{Loc}}_{(in,jm)} \log(p''_{(in,j\hat{\sigma}^{\text{prop}}_n(m))})$$

Effective batch size

Avoiding Collapse

Proposal Selection Contrast (ProSeCo)

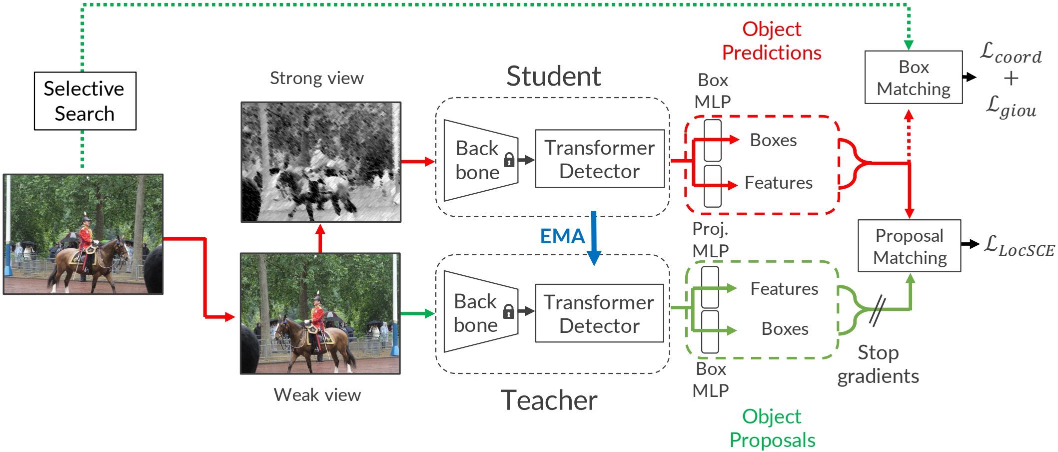


- Student predictions must match boxes *randomly* selected from **Selective Search**[9] outputs.

[9] Jasper RR Uijlings et al. "Selective search for object recognition". In: *IJCV*. 2013.

Full pretraining procedure

Proposal Selection Contrast (ProSeCo)



- Full pretraining procedure with both contrastive and localization learning.

Pretraining on ImageNet, finetuning on Mini-COCO
Experimental Results

Pretraining	Arch.	Mini-COCO		
		1% (1.2k)	5% (5.9k)	10% (11.8k)
Supervised	Trans.	13.0	23.6	28.6
SwAV[10]	Trans.	13.3	24.5	29.5
SCRL[11]	Trans.	16.4	26.2	30.6
DETReg[12]	Trans.	15.9	26.1	30.9
Supervised	Conv.	–	19.4	24.7
SoCo*[13]	Conv.	–	26.8	31.1
ProSeCo (Ours)	Trans.	**18.0**	**28.8**	**32.8**

[10] Mathilde Caron et al. "Unsupervised learning of visual features by contrasting cluster assignments". In: *NeurIPS*. 2020.

[11] Byungseok Roh et al. "Spatially consistent representation learning". In: *CVPR*. 2021.

[12] Amir Bar et al. "Detreg: Unsupervised pretraining with region priors for object detection". In: *CVPR*. 2022.

[13] Fangyun Wei et al. "Aligning pretraining for detection via object-level contrastive learning". In: *NeurIPS*. 2021.

Finetuning on other datasets

Experimental Results

Pretraining	FSOD-test	FSOD-train	PASCAL VOC	Mini-VOC	
	100% (11k)	100% (42k)	100% (16k)	5% (0.8k)	10% (1.6k)
Supervised	39.3	42.6	59.5	33.9	40.8
DETReg[14]	43.2	43.3	63.5	43.1	48.2
ProSeCo (Ours)	**46.6**	**47.2**	**65.1**	**46.1**	**51.3**

✓ Improvements of about **2 points over SOTA** on all datasets considered.

[14] Amir Bar et al. "Detreg: Unsupervised pretraining with region priors for object detection". In: *CVPR*. 2022.

Take Home Message II

We propose ProSeCo, a Proposal-Contrastive Pretraining strategy for Object Detection with Transformers.[15]

- ✓ Leverage high number of Object Proposals for **Proposal-Contrastive Learning**.
- ✓ Our **ProSeCo improves performance** when training with limited labeled data.
- ✓ **Consistency** with object-level features is important for Object Detection.
- ✓ **Location information** helps for Proposal-Contrastive learning.

[15] Quentin Bouniot, Romaric Audigier, et al. "Proposal-Contrastive Pretraining for Object Detection from Fewer Data". In: *ICLR*. 2023.

Outline

1. Introduction
2. Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning
3. Proposal-Contrastive Pretraining for Object Detection from Fewer Data
4. Understanding Deep Neural Networks Through the Lens of their Non-Linearity
   - Quantifying Non-linearity
   - Journey through DNNs History
   - Additional Results
5. Perspectives

Motivations

Non-linearity is at the heart of DNNs

- *Universal function approximators* thanks to non-linearity.
- Mainly introduced through *activation functions*.

No such notion of quantifying non-linearity exists in the literature.

- Research mainly focus on quantifying expressive power of DNNs.

Goal: Measure non-linearity *from data distribution*

Quantifying Non-Linearity
General idea

Measure non-linearity as lack of linearity through Optimal Transport (OT)

- We know the closed-form solution of the OT problem for random variables following normal distributions.
- For any $\mathbf{X}$ and $\mathbf{Y}$, if $\mathbf{Y} = T\mathbf{X}$ with T PSD, then *the solution of OT problem is exactly the one of their normal approximations*.
- We obtain a bound on the difference of the two OT problems.
- We can define the *affinity score* using this bound.

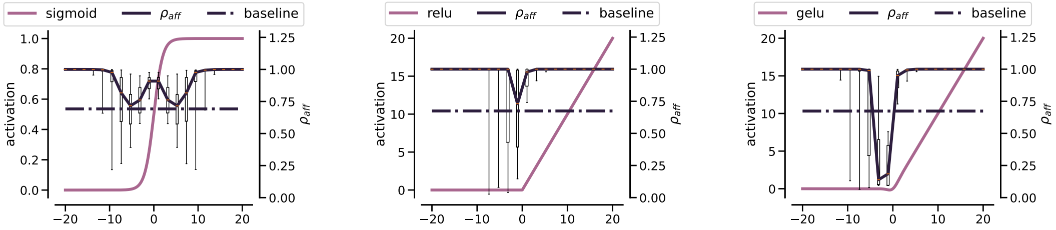
Quantifying Non-Linearity
Affinity Score

$$\boxed{\rho_{\mathrm{aff}}(\mathbf{X}, \mathbf{Y}) = 1 - \frac{\overbrace{W_2}^{\text{2-Wasserstein distance}}(\overbrace{T_{\mathrm{aff}}}^{\text{OT map between normal approximations}}\mathbf{X}, \mathbf{Y})}{\sqrt{2}\,\mathrm{Tr}[\underbrace{\Sigma(\mathbf{Y})}_{\text{Covariance of Y}}]^{\frac{1}{2}}}}$$

- ρ_{aff} describes how much Y differs from being a *PSD affine transformation of X*.
- $0 \leq \rho_{\mathrm{aff}}(X, Y) \leq 1$, and $\rho_{\mathrm{aff}}(X, Y) = 1 \Leftrightarrow Y = T_{\mathrm{aff}} X$.

Quantifying Non-Linearity
First Examples

Affinity scores over input domain of activation functions



- $\mathbf{X} \sim \mathcal{N}(\mu, \sigma)$, with μ sliding over the domain and multiple σ for each μ.
- $\rho_{\text{aff}}(\mathbf{X}, f(\mathbf{X}))$ for popular activation functions f.
- Activation functions can be characterized by *the lowest score achieved* and *the range of non-linearity*.

Non-linearity signature

Journey through DNNs History

Notations

- Define a neural network N as a *composition of layers* F_i:
  $N = F_L \odot ... \odot F_i ... \odot F_1 = \bigodot_{k=1,\dots,L} F_k$ where $\odot$ stands for a composition.

- Each layer F_i is a function $F_i : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^{h\times w\times c}$ whose outputs $F_i(\mathbf{X}_i)$ are inputs of the following layer F_{i+1}. Usual F_i include convolution, feedforward, pooling or activation functions.

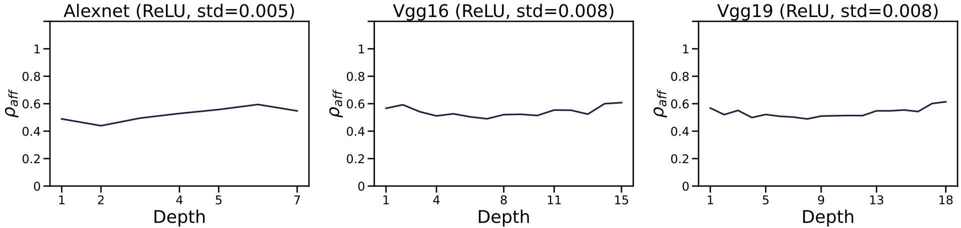
- Define a *finite set of common activation functions* $\mathcal{A} := \{\sigma | \sigma : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^{h\times w\times c}\}$

- Let r be a *dimensionality reduction function* such that $r : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^c$

Non-linearity signature of N given X:

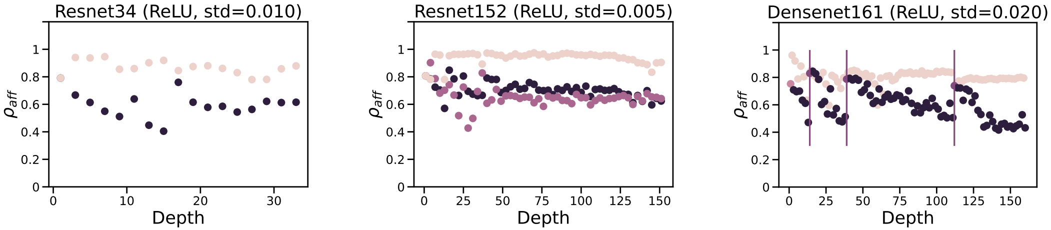
$$\boxed{\rho_{\text{aff}}(N;\mathbf{X}) = \{\rho_{\text{aff}}(r(\mathbf{X}_i), \sigma(r(\mathbf{X}_i))), \forall \sigma \in F_i \cap \mathcal{A}, i \in \{1,\dots,L\}\}}$$

Early Convnets
Journey through DNNs History



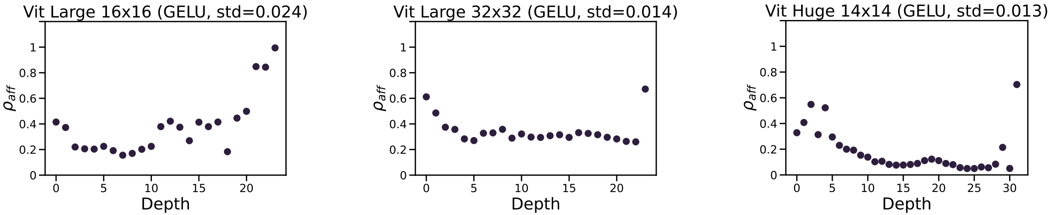
- Early convnets had **tiny variations** in non-linearity propagation.

Deeper Networks
Journey through DNNs History



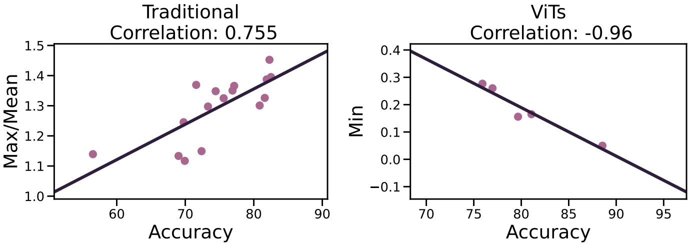
- Different color codes stand for *distinct* activation functions appearing *repeatedly* in the architecture (*e.g.* every first ReLU in residual blocks for ResNet).
- Deeper networks with *residual connections* have a **shaking effect** in their non-linearity signatures.

Vision Transformers
Journey through DNNs History



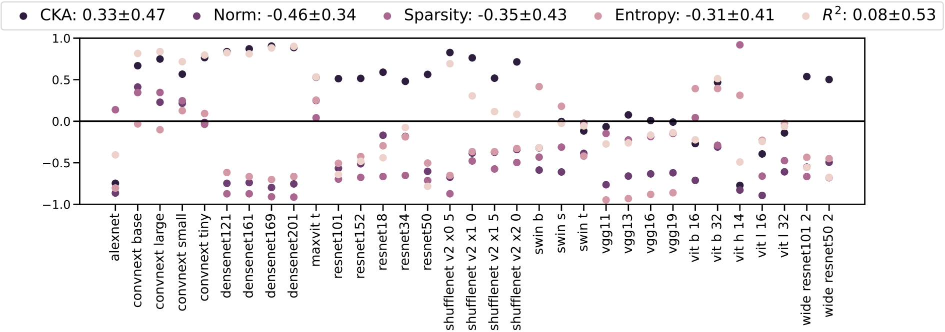
- Activation functions only present in their MLP blocks.
- **Highly non-linear** compared to convnets.

Correlation with Accuracy
Additional Results



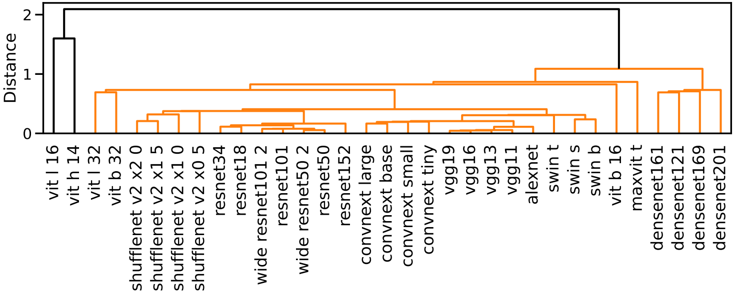
- We separate architectures into semantically meaningful groups: **Traditional architectures** (Alexnet, VGGs, ResNets and DenseNets) and **ViTs**.
- Confirms **shaking effect** for traditional models.
- Clear trend toward **more non-linearity in ViTs**.

Unique Measure
Additional Results



- **No other criterion consistently correlates with the affinity score** across 33 architectures used in our test.

Clustering of architectures
Additional Results



- Clustering of the architectures using **the pairwise DTW distances** between non-linearity signatures.

Take-Home Message III

Understanding Deep Neural Networks Through the Lens of their Non-Linearity[16]

- ✓ First theoretical sound tool to measure non-linearity in DNNs
- ✓ Different developments in Deep Learning can be understood through the prism of non-linearity
- ✓ Variety of potential applications

[16] Quentin Bouniot, Ievgen Redko, Anton Mallasto, et al. "Understanding deep neural networks through the lens of their non-linearity". In: *arXiv preprint arXiv:2310.11439* (2023).

Outline

1. Introduction
2. Improving Few-Shot Classification with Meta-Learning through Multi-Task Learning
3. Proposal-Contrastive Pretraining for Object Detection from Fewer Data
4. Understanding Deep Neural Networks Through the Lens of their Non-Linearity
5. Perspectives

Perspectives

Towards bridging the gap between MTR theory and Meta-learning in practice.

- Take into account similarity between source and test tasks for *cross-domain generalization*.

Perspectives

Towards bridging the gap between MTR theory and Meta-learning in practice.

- Take into account similarity between source and test tasks for *cross-domain generalization*.

Towards leveraging unlabeled data for Object Detection using Transformers.

- Improvements from self- and semi-supervision are less significant than for convolutional methods. Consider *more suited* unsupervised tasks ?

Perspectives

Towards bridging the gap between MTR theory and Meta-learning in practice.

- Take into account similarity between source and test tasks for *cross-domain generalization*.

Towards leveraging unlabeled data for Object Detection using Transformers.

- Improvements from self- and semi-supervision are less significant than for convolutional methods. Consider *more suited* unsupervised tasks ?

Towards efficient adaptation through non-linearity analysis

- Comparing datasets through distance between non-linearity signatures
- Regularization of non-linearity signatures during training.

Thank you for listening !

Do not hesitate to contact me if you have questions.

Contributions

- Quentin Bouniot, Ievgen Redko, Romaric Audigier, et al. "Improving Few-Shot Learning Through Multi-task Representation Learning Theory". In: *ECCV*. 2022.
- Quentin Bouniot, Romaric Audigier, et al. "Proposal-Contrastive Pretraining for Object Detection from Fewer Data". In: *ICLR*. 2023.
- Quentin Bouniot, Ievgen Redko, Anton Mallasto, et al. "Understanding deep neural networks through the lens of their non-linearity". In: *arXiv preprint arXiv:2310.11439* (2023).

References I

Jake Snell, Kevin Swersky, and Richard S. Zemel. “Prototypical Networks for Few-shot Learning”. In: *NeurIPS*. 2017.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”. In: *ICML*. 2017.

Simon S. Du et al. “Few-Shot Learning via Learning the Representation, Provably”. In: *ICLR*. 2021.

Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. “Provable Meta-Learning of Linear Representations”. In: *arXiv*. 2020.

Fangyun Wei et al. “Aligning pretraining for detection via object-level contrastive learning”. In: *NeurIPS*. 2021.

Zhigang Dai et al. “Up-DETR: Unsupervised pre-training for object detection with transformers”. In: *CVPR*. 2021.

Amir Bar et al. “Detreg: Unsupervised pretraining with region priors for object detection”. In: *CVPR*. 2022.

References II

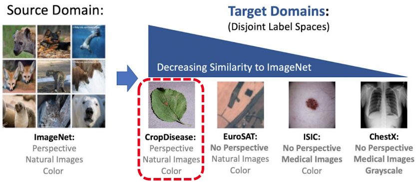
- Julien Denize et al. "Similarity contrastive estimation for self-supervised soft contrastive learning". In: *WACV*. 2023.
- Jasper RR Uijlings et al. "Selective search for object recognition". In: *IJCV*. 2013.
- Mathilde Caron et al. "Unsupervised learning of visual features by contrasting cluster assignments". In: *NeurIPS*. 2020.
- Byungseok Roh et al. "Spatially consistent representation learning". In: *CVPR*. 2021.
- Yunhui Guo et al. "A Broader Study of Cross-Domain Few-Shot Learning". In: *ECCV*. 2020.
- Zhi Tian et al. "Fcos: Fully convolutional one-stage object detection". In: *ICCV*. 2019.
- Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *NeurIPS*. 2015.
- Tsung-Yi Lin et al. "Feature pyramid networks for object detection". In: *CVPR*. 2017.

References III

Xizhou Zhu et al. "Deformable DETR: Deformable Transformers for End-to-End Object Detection". In: *ICLR*. 2021.

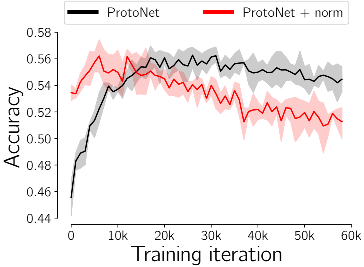
Nicolas Carion et al. "End-to-end object detection with transformers". In: *ECCV*. 2020.

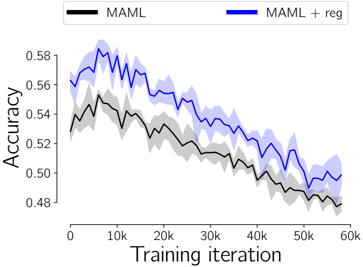
Experimental Results



5-way
1-shot

Guo et al., “A Broader Study of Cross-Domain Few-Shot Learning”





× Improvement does not translate to cross-domain for *metric-based methods*.

✓ *Gradient-based methods* keep their accuracy gains.

Few-Shot Learning Setting
Background in Object Detection

How do object detectors handle data scarcity ?

Method	Arch.	Mini-COCO			
		0.5% (590)	1% (1.2k)	5% (5.9k)	10% (11.8k)
FCOS[17]	Conv.	5.42 ± 0.01	8.43 ± 0.03	17.01 ± 0.01	20.98 ± 0.01
FRCNN + FPN[18]	Conv.	6.83 ± 0.15	9.05 ± 0.16	18.47 ± 0.22	23.86 ± 0.81
Def. DETR[19]	Trans.	$\mathbf{8.95 \pm 0.51}$	$\mathbf{12.96 \pm 0.08}$	$\mathbf{23.59 \pm 0.21}$	$\mathbf{28.55 \pm 0.08}$

- Performance on COCO with different **percentages** of labeled training data.
- **Def. DETR** stronger than FRCNN + FPN and FCOS **with fewer labeled data.**

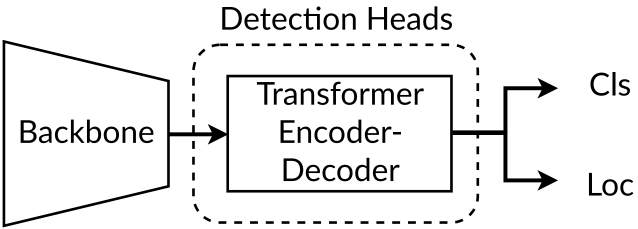
[17] Zhi Tian et al. "Fcos: Fully convolutional one-stage object detection". In: *ICCV*. 2019.

[18] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *NeurIPS*. 2015; Tsung-Yi Lin et al. "Feature pyramid networks for object detection". In: *CVPR*. 2017.

[19] Xizhou Zhu et al. "Deformable DETR: Deformable Transformers for End-to-End Object Detection". In: *ICLR*. 2021.

Object Detection 101
Background in Object Detection

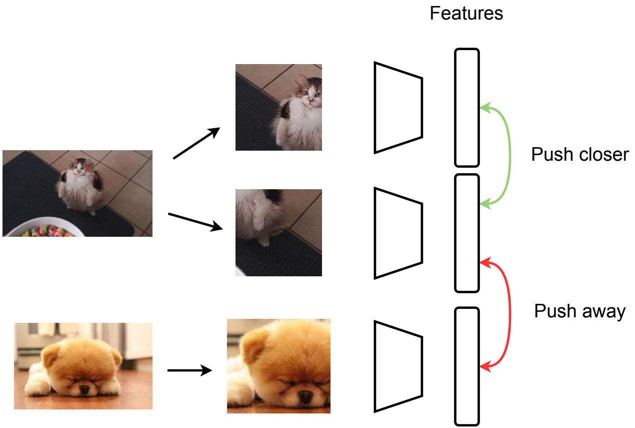
Transformer-based methods (e.g., DETR[20])



- **Simpler** overall architecture, without **hand-crafted heuristics**.
- Increasingly popular architecture and **strong performance with few data**.

[20] Nicolas Carion et al. "End-to-end object detection with transformers". In: *ECCV*. 2020.

Classical Contrastive Learning
Unsupervised Pretraining for Object Detection with Fewer Annotation



- ▶ Push closer positive examples and push away negative examples.

Ablation Studies

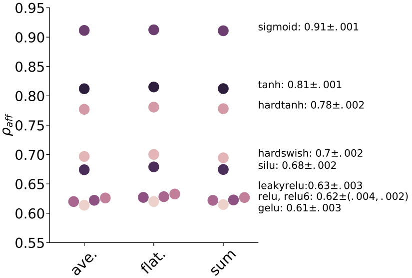
Pretraining	Dataset	mAP
ProSeCo w/ SwAV	COCO	27.4
ProSeCo w/ SwAV	IN	27.8
DETReg w/ SCRL	IN	28.0
ProSeCo w/ SCRL	IN	**28.8**

Loss	δ	mAP
SCE	1.0	26.1
LocSCE (Ours)	0.2	27.0
LocSCE (Ours)	0.7	27.1
LocSCE (Ours)	0.5	**27.8**

- ▶ **Dataset diversity** more important than closeness to downstream task
- ✓ **Consistency** in the features improves performance
- ✓ **Location of proposals** helps for introducing **easy positives** for contrastive learning

Quantifying Non-Linearity
Dimensionality reduction

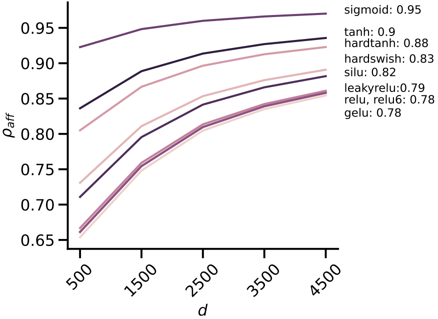
Affinity scores are robust to dimensionality reduction

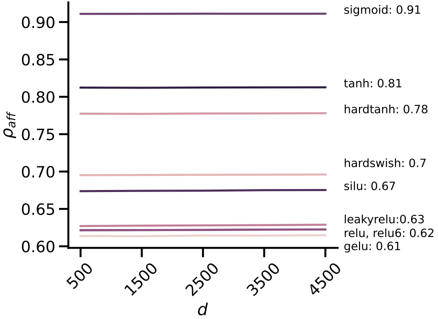


- ▶ Manipulating 4-order tensor is computationally expensive
- ▶ Averaging over a dimension preserve affinity scores

Quantifying Non-Linearity
Covariance estimation

Shrinkage of the covariance makes it robust to sample size

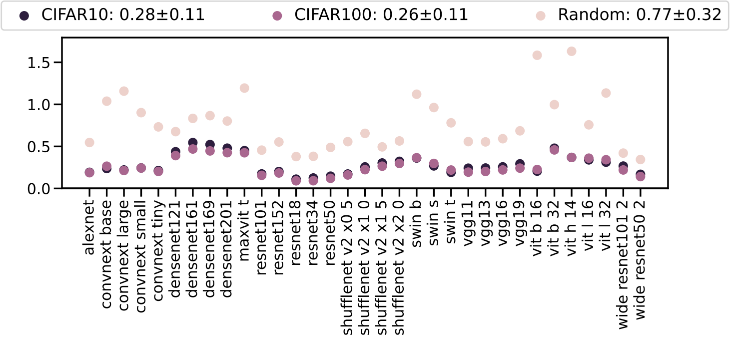




- *Ledoit-Wolfe shrinkage* of the covariance gives stable results for affinity scores.

Deviations between datasets

Additional Results



- **Deviations to ImageNet of different datasets** (CIFAR10, CIFAR100, random data), for each architecture.