On Recent Progress in Few-shot Classification

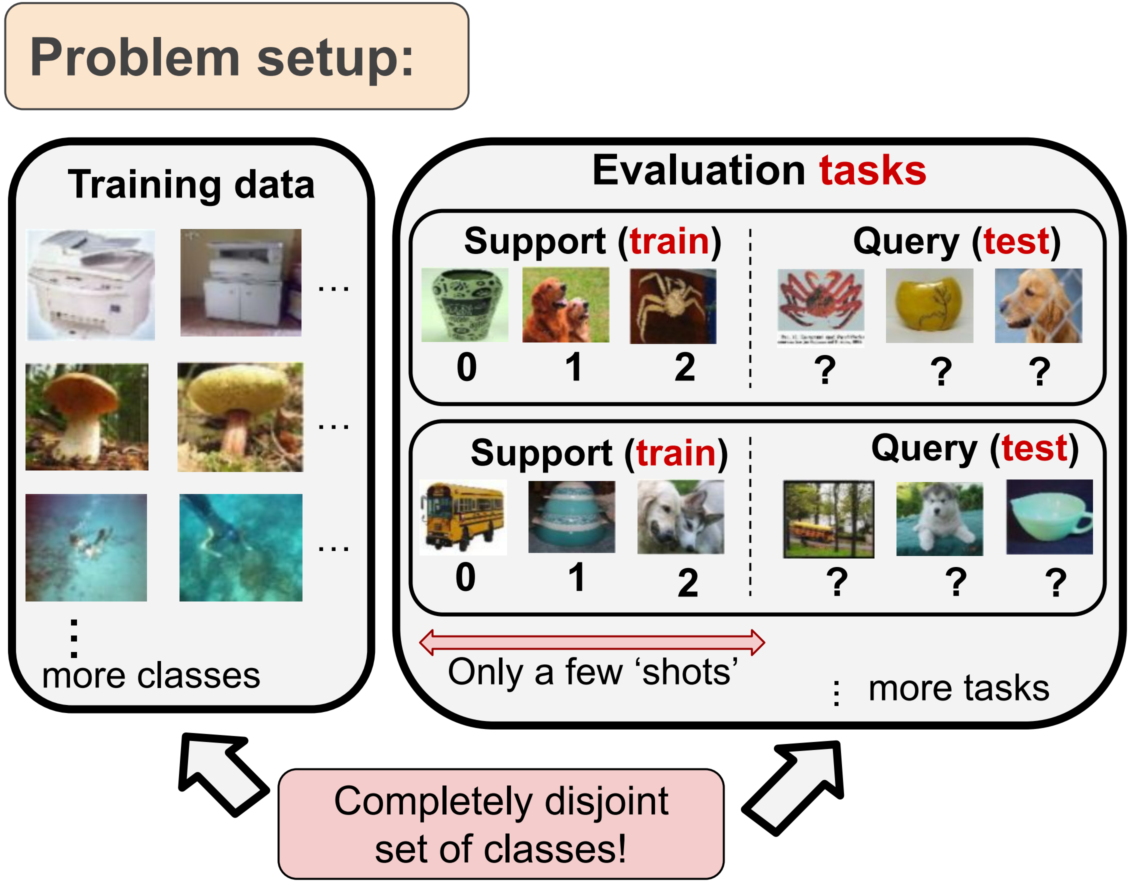
Eleni Triantafillou

Roadmap

- What is few-shot classification?
- Main families of approaches
- Representative meta-learning models
- Meta-Dataset: a more realistic and large-scale benchmark
- Open challenges and next steps

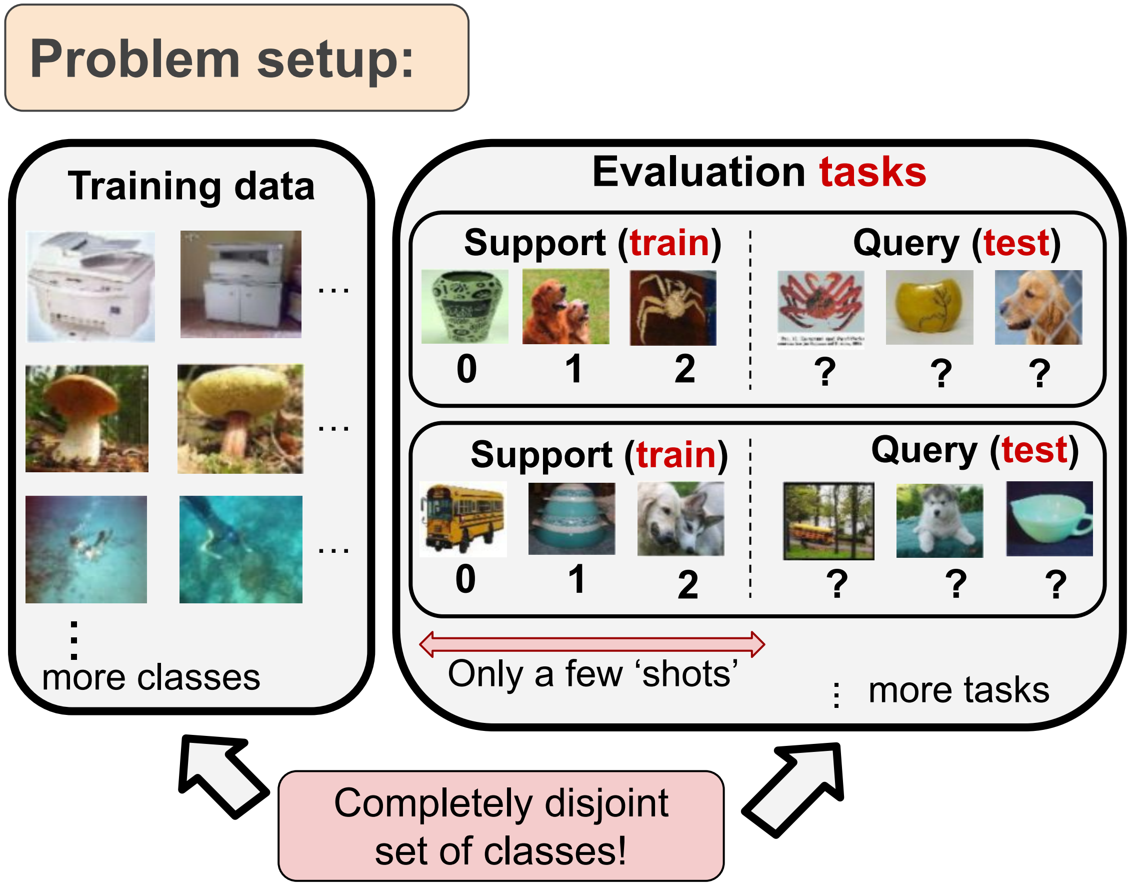
Few-shot Classification

- Learn new classes from few examples
- Practically important and scientifically interesting



Few-shot Classification

- Learn new classes from few examples
- Practically important and scientifically interesting
- Two challenges:
  1. How to use the training data to learn a model that supports rapid learning? ('*training approach*')
  2. How to use the support set of each task to adapt? ('*inference algorithm*')



Families of approaches to few-shot classification

Different approaches differ by their choice of training approach and inference algorithm:

- Generative modeling
- Metric learning
- Transfer learning
- Meta-learning

Disclaimer: this presentation does not thoroughly cover all approaches. Happy to chat offline and point to related papers / more information about any particular approach.

Teaser: Early Generative Approaches (Fei-Fei et al., 2006)

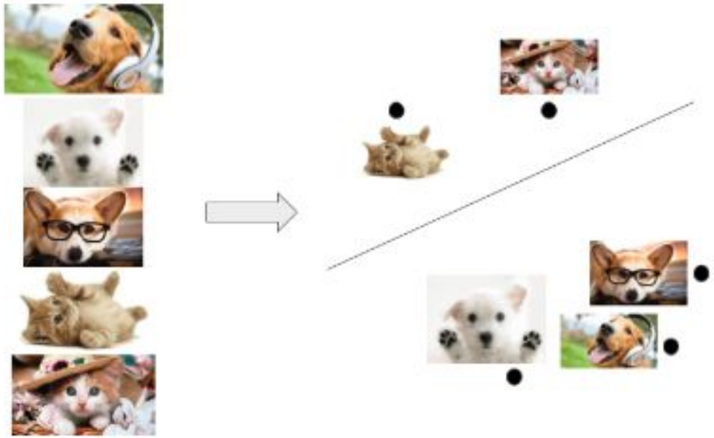
- At test time, we need to estimate parameters θ_c for $p(x|\theta_c)$ for the *test* class c using its **support set** $\mathcal{S}_c$
  - Bayes' rule to the rescue: $p(\theta_c|\mathcal{S}_c) \propto p(\theta_c)p(\mathcal{S}_c|\theta_c)$
  - The prior $p(\theta_c)$ reflects across-class general knowledge
  - A query point x^* then is classified as class c with probability:

$$p(x^*|\mathcal{S}_c) = \int p(x^*|\theta_c)p(\theta_c|\mathcal{S}_c)d\theta_c$$

- **Training approach**: Learn a prior probability density over models of classes
- **Inference algorithm**: Compute the posterior predictive probability of each query example x^* as shown above.

Metric Learning

- Learn an **embedding space** where examples cluster according to class labels



- **Training approach**: learn a *similarity function*.
- **Inference algorithm**: classify examples of new classes based on their *similarity* to the few support examples.

Siamese Networks (Koch et al., 2015)

- Metric learning method: use the training clases to learn a similarity metric.
- Siamese network: two identical branches for predicting similarity of pairs
- Koch et al. (2015) showed that a siamese network is capable of one-shot classification

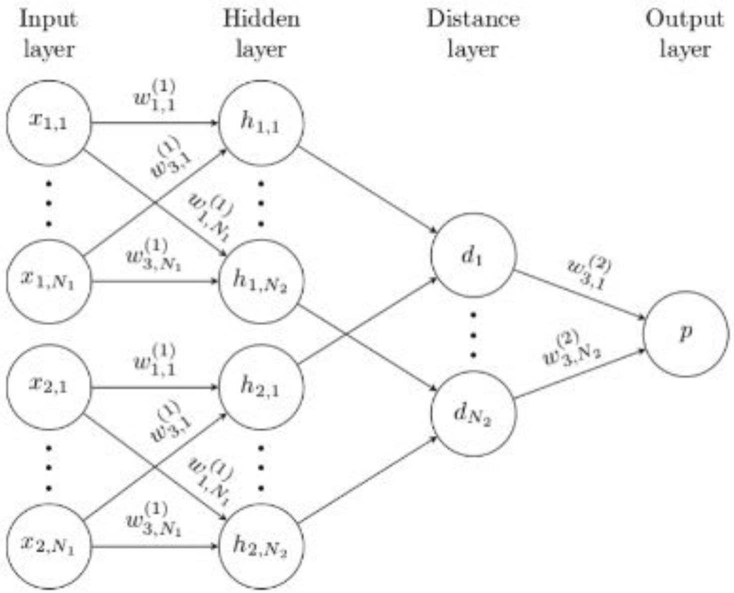
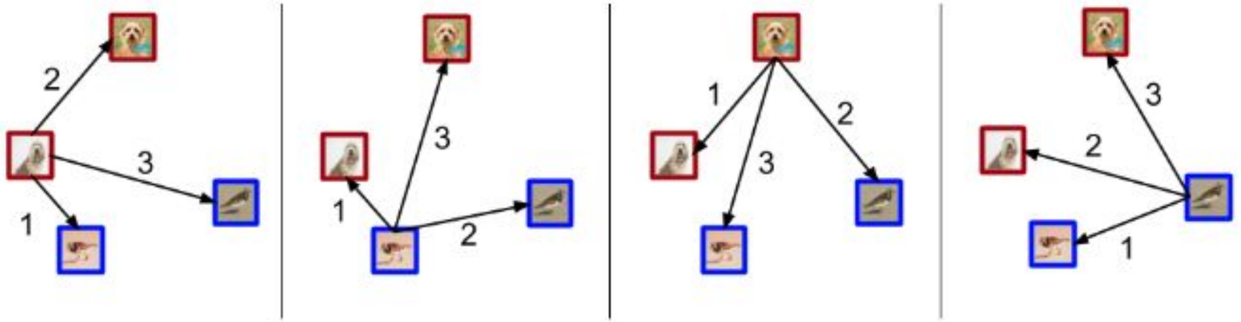


Figure: from (Koch et al., 2015)

Optimizing mean Average Precision (Triantafillou et al., 2017)

- Can we learn a better metric via a more informative objective?
- What do we *want* to hold in the embedding space? **Each point should be closer to *all* similar ones than to *any* dissimilar one.**
- We enforce this via a **structured objective**:
  - Compute ranks of predicted similarity between all examples
  - Maximize the mean Average Precision (mAP) of these rankings



Transfer learning vs Meta-learning

- Arguably the two most popular approaches recently

Transfer learning vs Meta-learning

- Arguably the two most popular approaches recently
- Influential work advocated for meta-learning

Matching Networks for One Shot Learning

Oriol Vinyals
Google DeepMind
vinyals@google.com

Charles Blundell
Google DeepMind
cblundell@google.com

Timothy Lillicrap
Google DeepMind
countzero@google.com

Koray Kavukcuoglu
Google DeepMind
korayk@google.com

Daan Wierstra
Google DeepMind
wierstra@google.com

Prototypical Networks for Few-shot Learning

Jake Snell
University of Toronto*

Kevin Swersky
Twitter

Richard S. Zemel
University of Toronto, Vector Institute

Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks

Chelsea Finn [1] **Pieter Abbeel** [1 2] **Sergey Levine** [1]

Transfer learning vs Meta-learning

- Influential work advocated for meta-learning
- Recently, strong transfer learning ‘baselines’ emerged

A CLOSER LOOK AT FEW-SHOT CLASSIFICATION

Wei-Yu Chen
Carnegie Mellon University
weiyuc@andrew.cmu.edu

Yen-Cheng Liu & Zsolt Kira
Georgia Tech
{ycliu,zkira}@gatech.edu

Yu-Chiang Frank Wang
National Taiwan University
ycwang@ntu.edu.tw

Jia-Bin Huang
Virginia Tech
jbhuang@vt.edu

A BASELINE FOR FEW-SHOT IMAGE CLASSIFICATION

Guneet S. Dhillon[1], **Pratik Chaudhari**[2*], **Avinash Ravichandran**[1], **Stefano Soatto**[1,3]
[1]Amazon Web Services, [2]University of Pennsylvania, [3]University of California, Los Angeles
{guneetsd, ravinash, soattos}@amazon.com, pratikac@seas.upenn.edu

Transfer learning vs Meta-learning

- Influential work advocated for meta-learning
- Recently, strong transfer learning 'baselines' emerged
- More recently, hybrid 'pre-training' and meta-learning models achieved better results

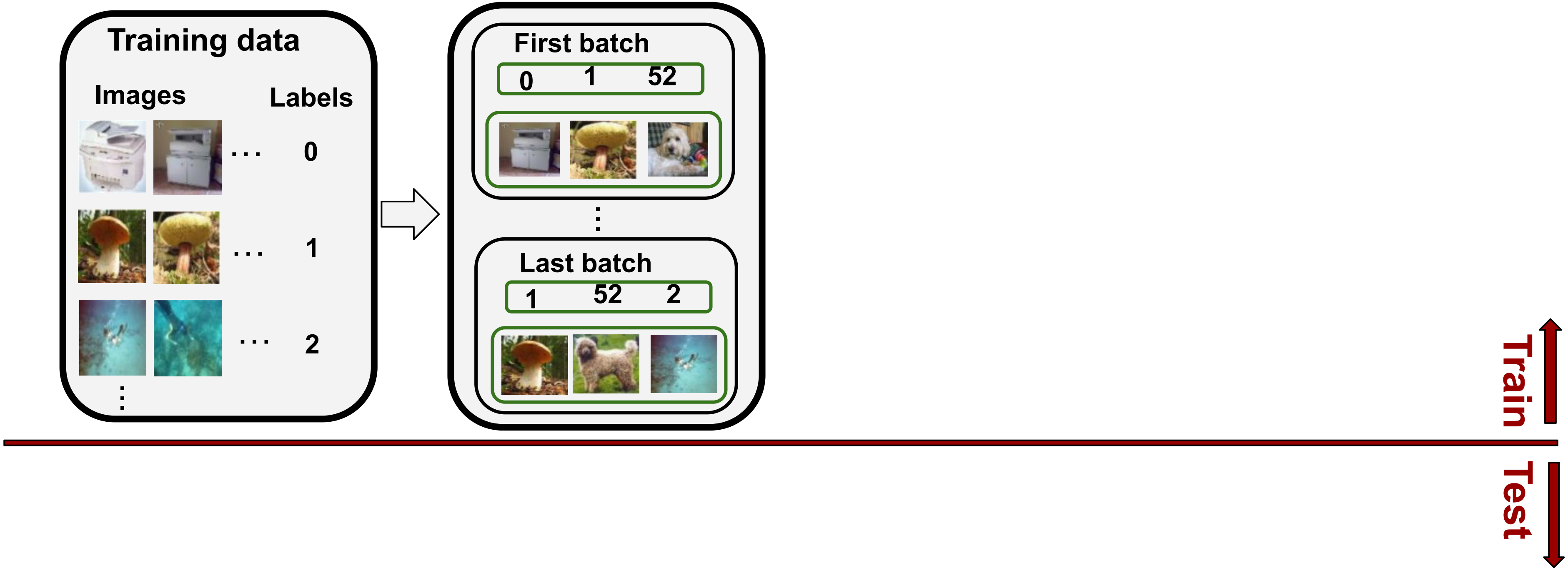
META-DATASET: A DATASET OF DATASETS FOR LEARNING TO LEARN FROM FEW EXAMPLES

Eleni Triantafillou[*†], Tyler Zhu[†], Vincent Dumoulin[†], Pascal Lamblin[†], Utku Evci[†], Kelvin Xu[‡†], Ross Goroshin[†], Carles Gelada[†], Kevin Swersky[†], Pierre-Antoine Manzagol[†] & Hugo Larochelle[†]
[*]University of Toronto and Vector Institute, [†]Google AI, [‡]University of California, Berkeley
Correspondence to: eleni@cs.toronto.edu

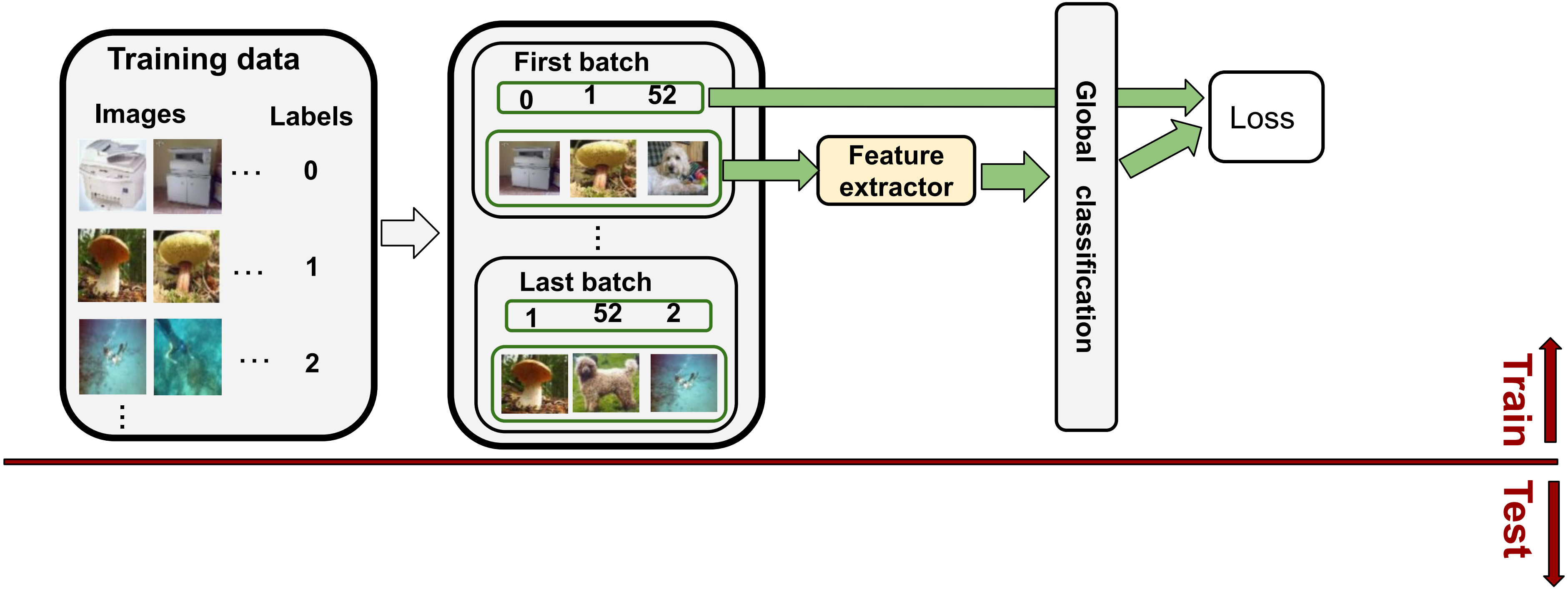
A New Meta-Baseline for Few-Shot Learning

Yinbo Chen[1] Xiaolong Wang[2] Zhuang Liu[2] Huijuan Xu[2] Trevor Darrell[2]

Transfer learning ‘baseline’

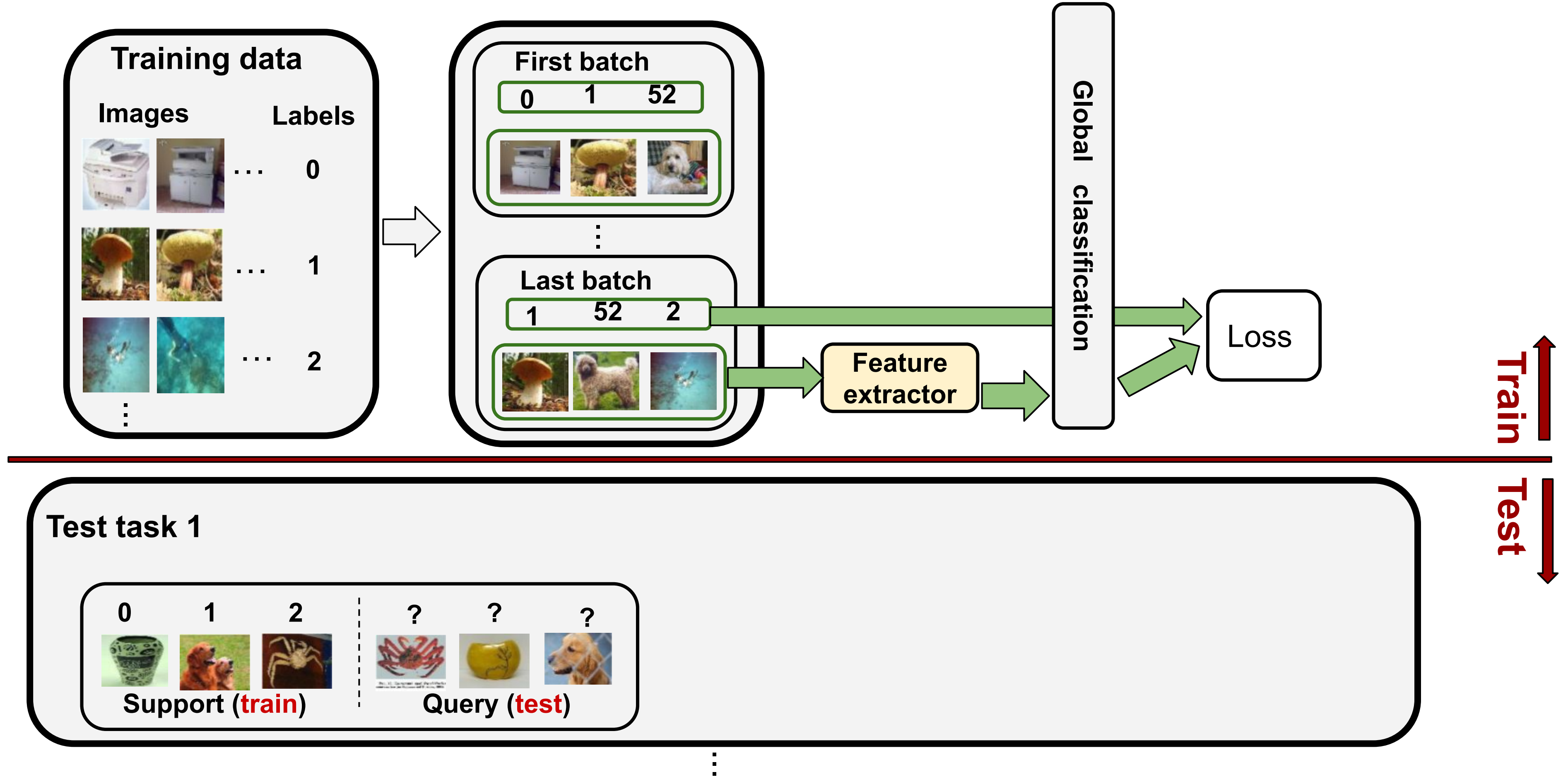


Transfer learning ‘baseline’

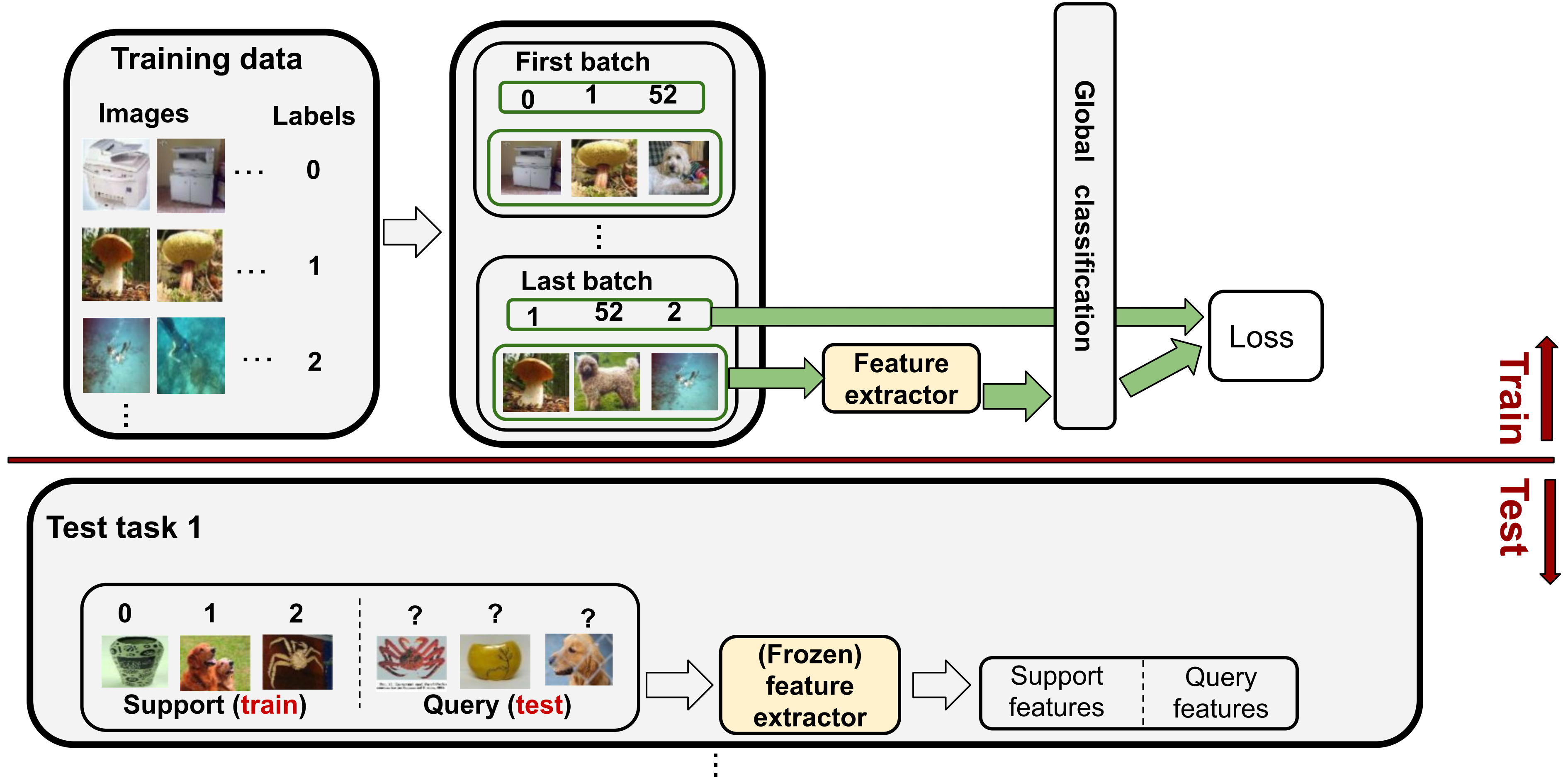


Transfer learning ‘baseline’

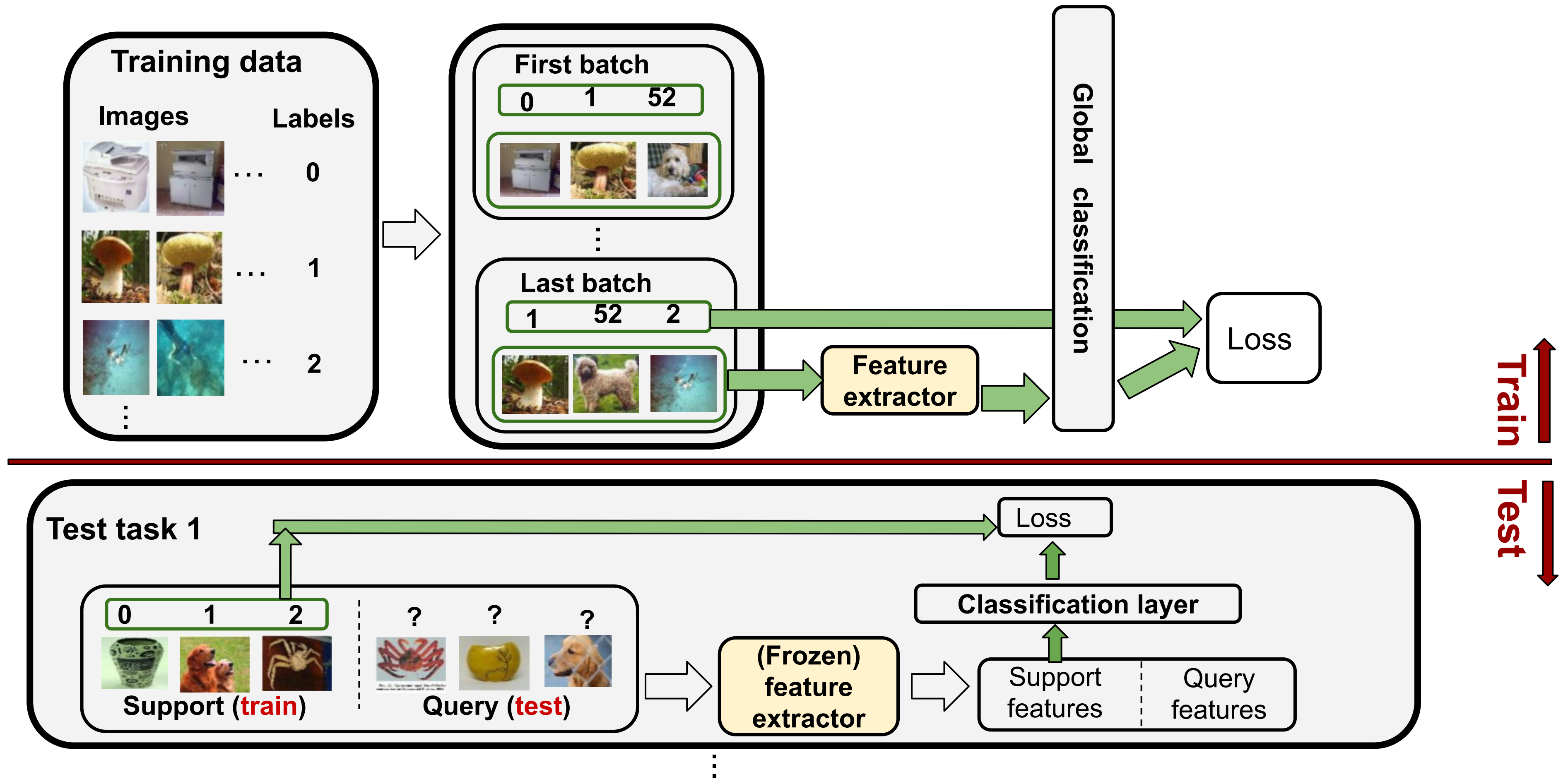
Transfer learning ‘baseline’



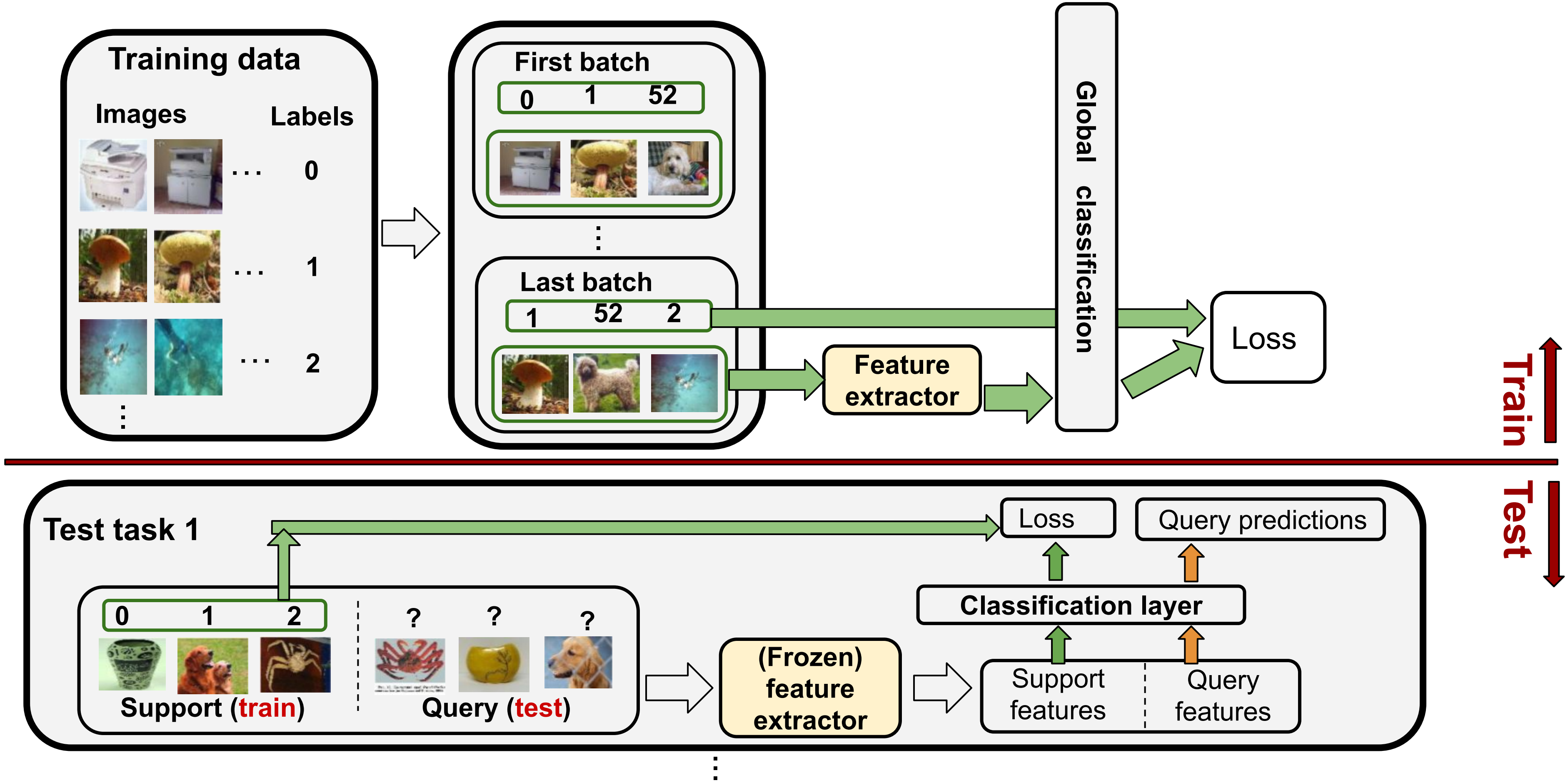
Transfer learning 'baseline'



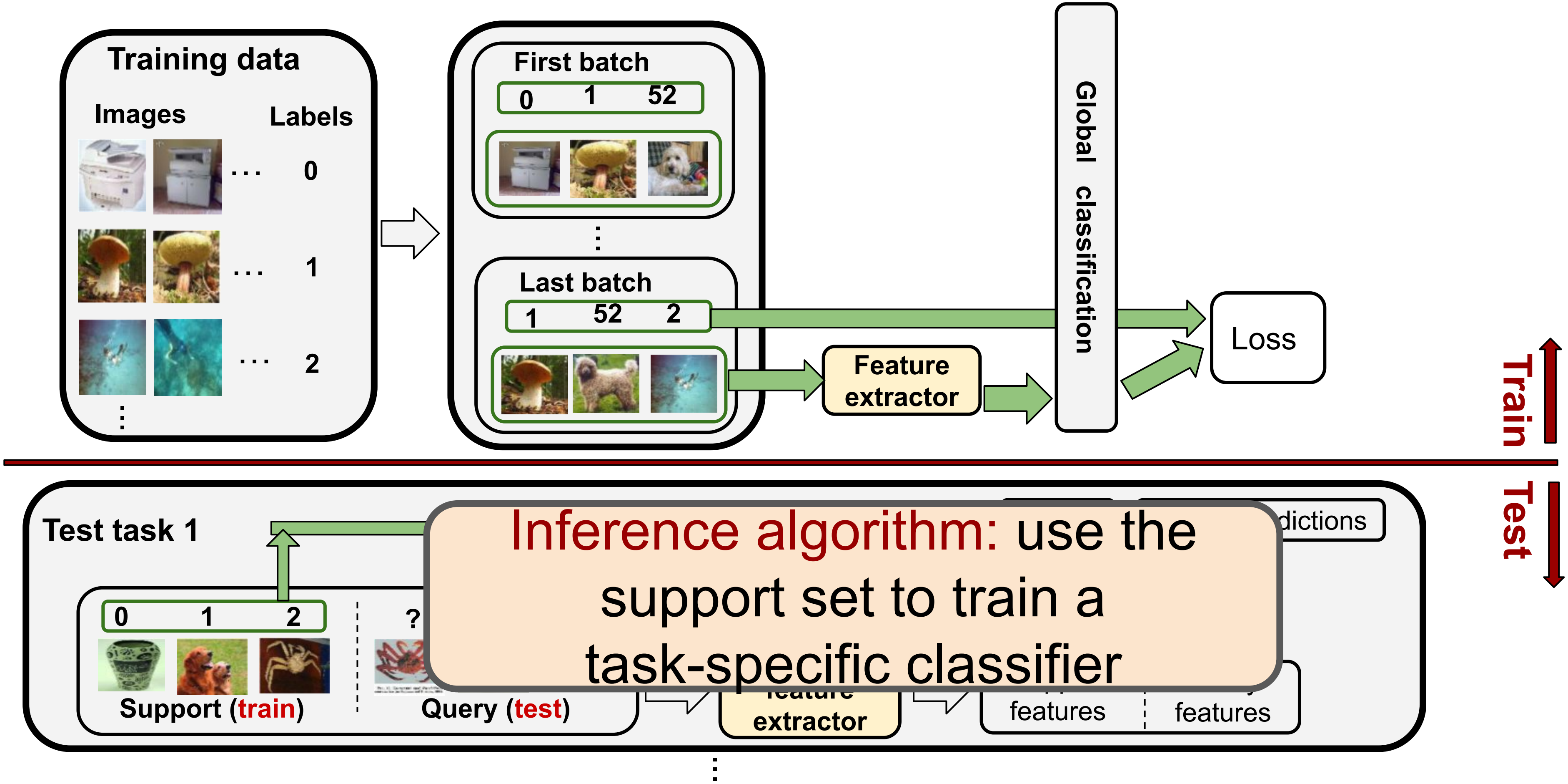
Transfer learning ‘baseline’



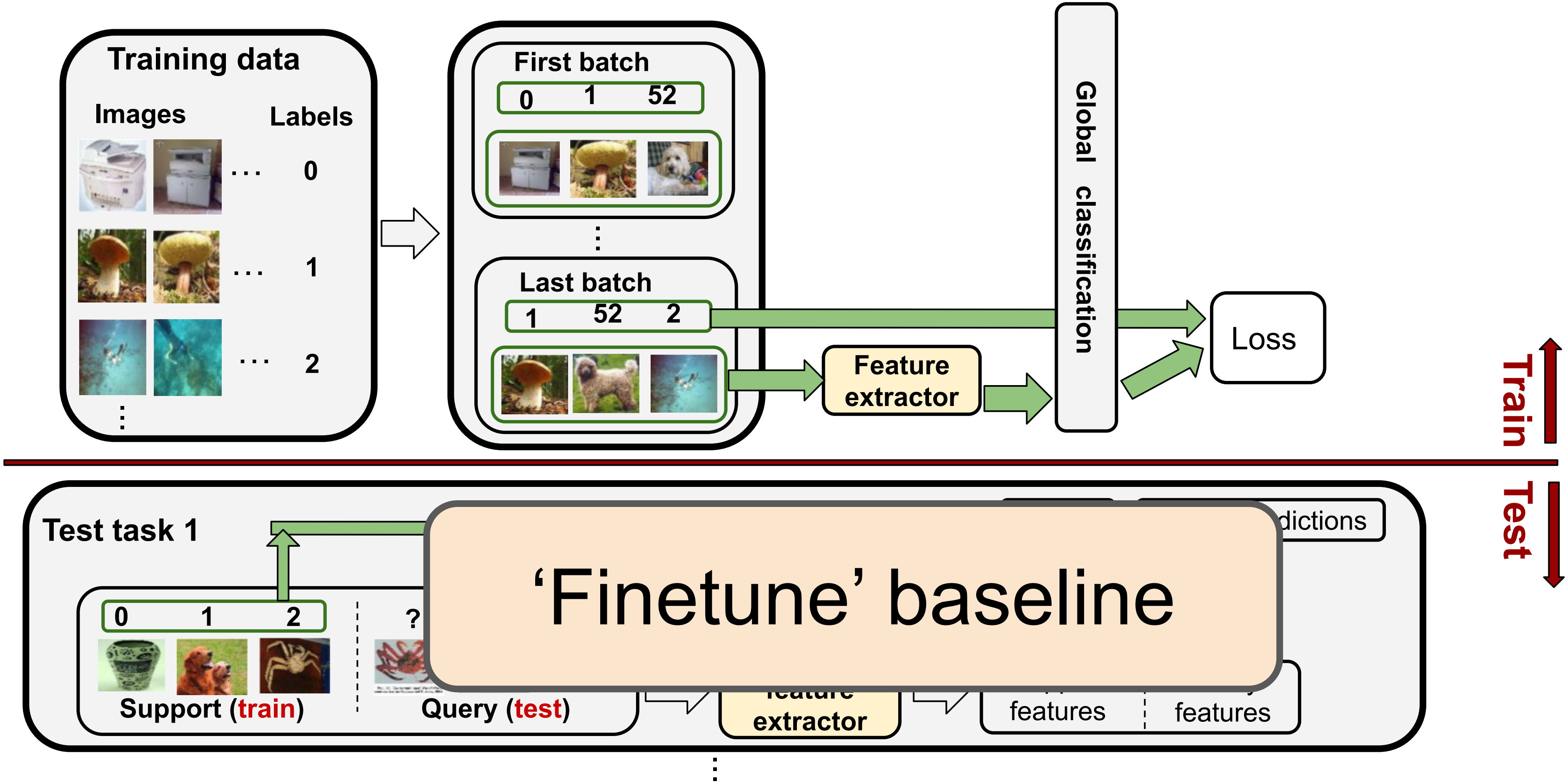
Transfer learning ‘baseline’



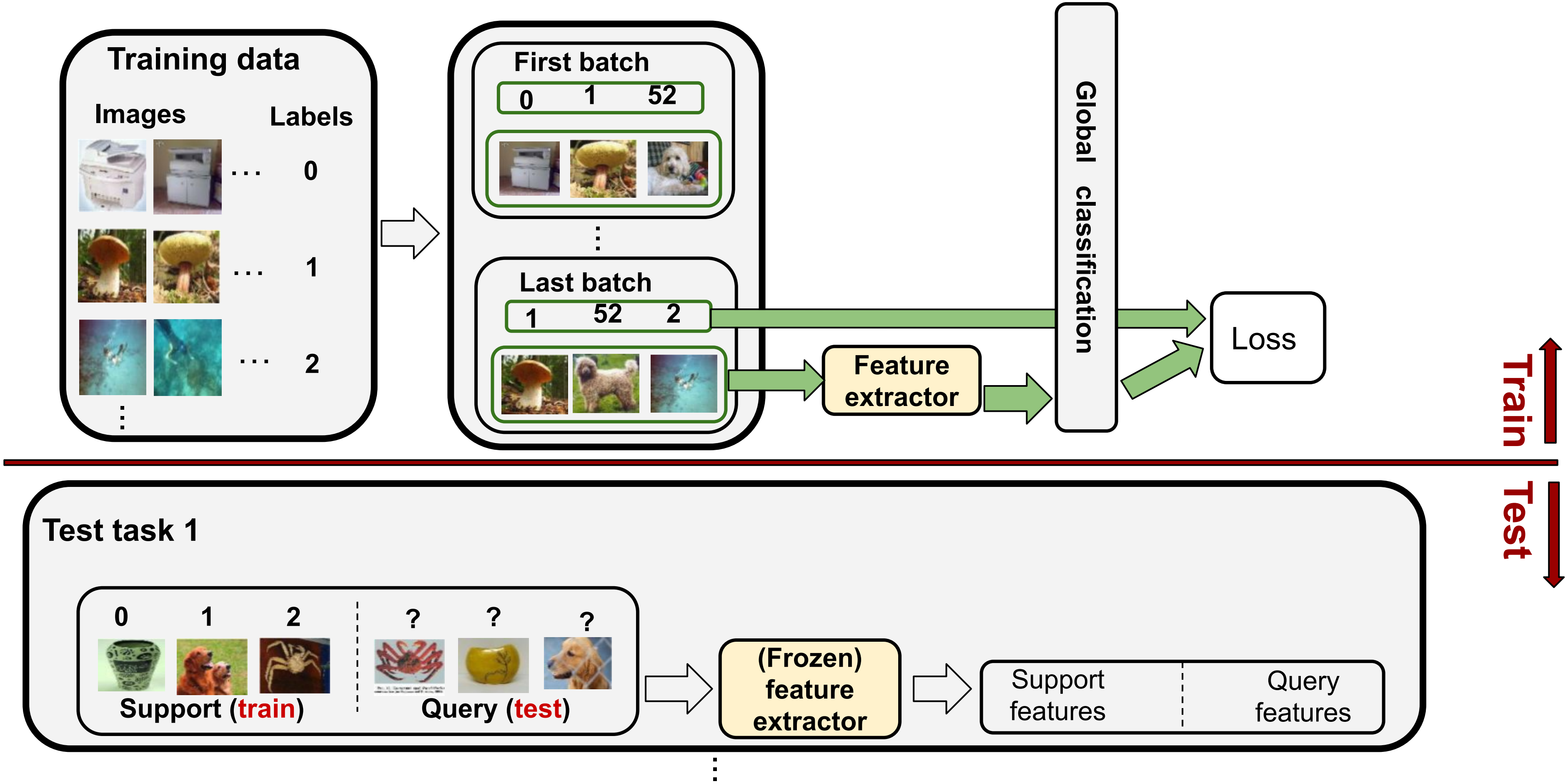
Transfer learning ‘baseline’



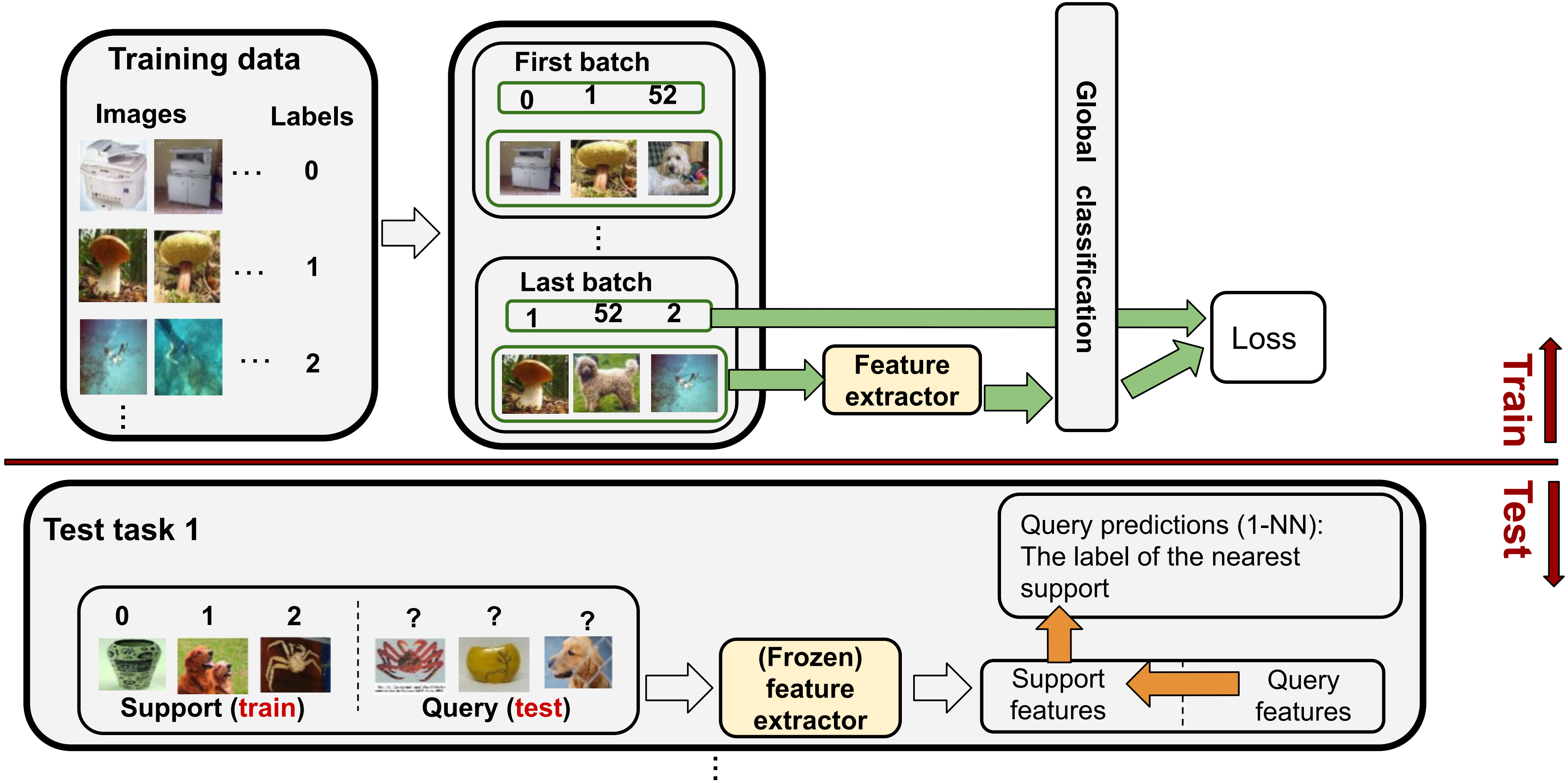
Transfer learning ‘baseline’



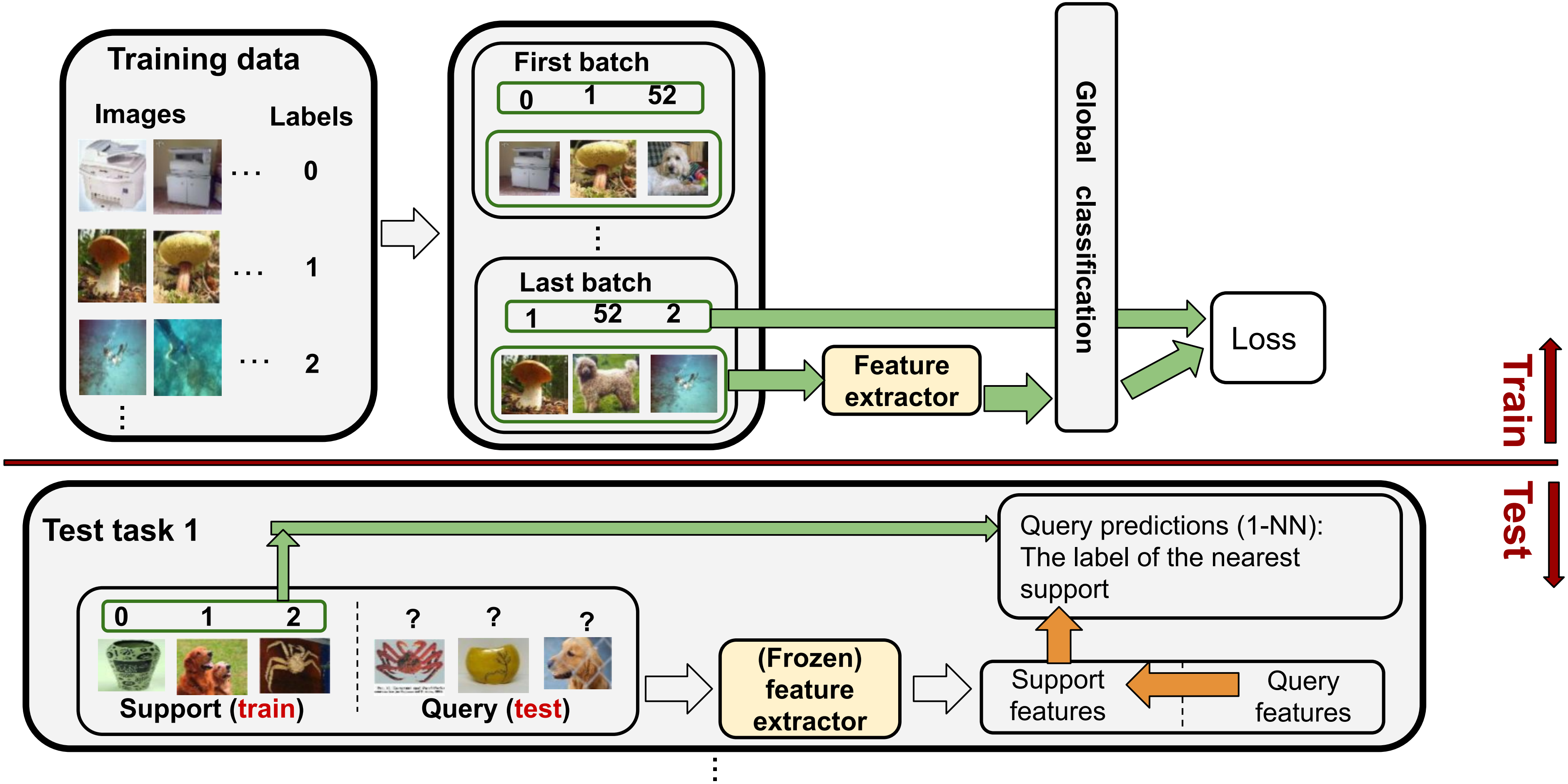
Transfer learning ‘baseline’



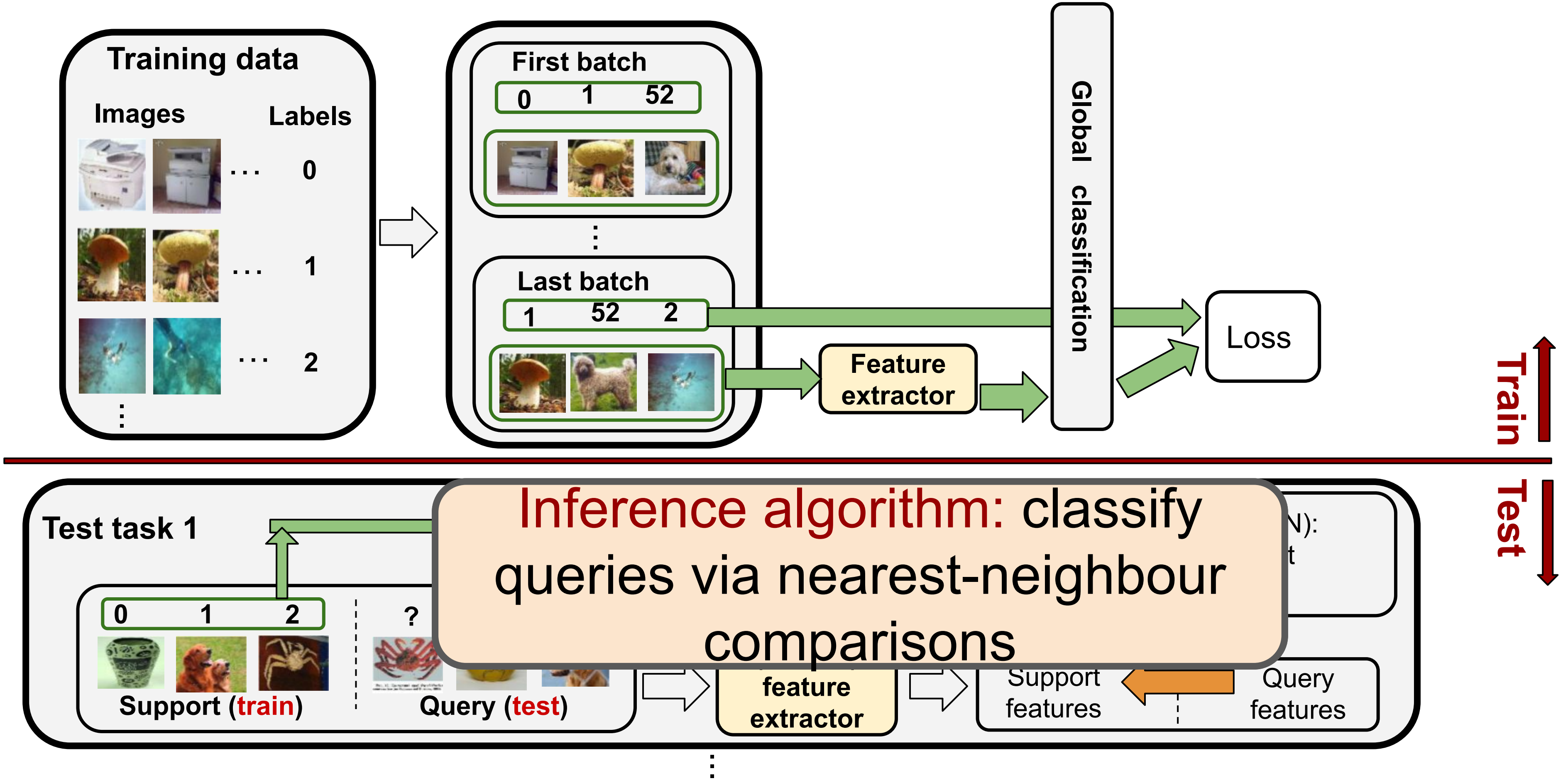
Transfer learning ‘baseline’



Transfer learning ‘baseline’

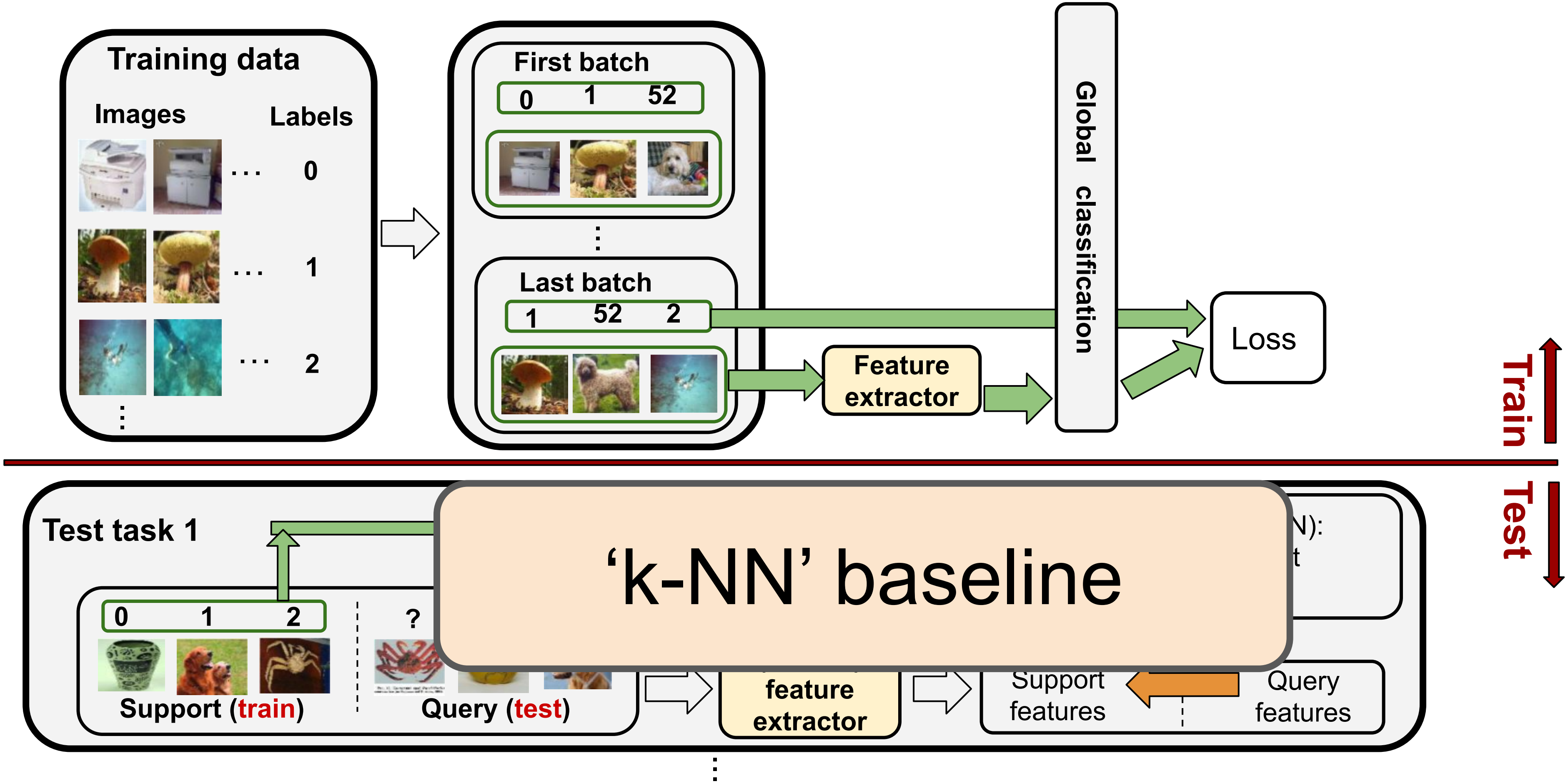


Transfer learning 'baseline'

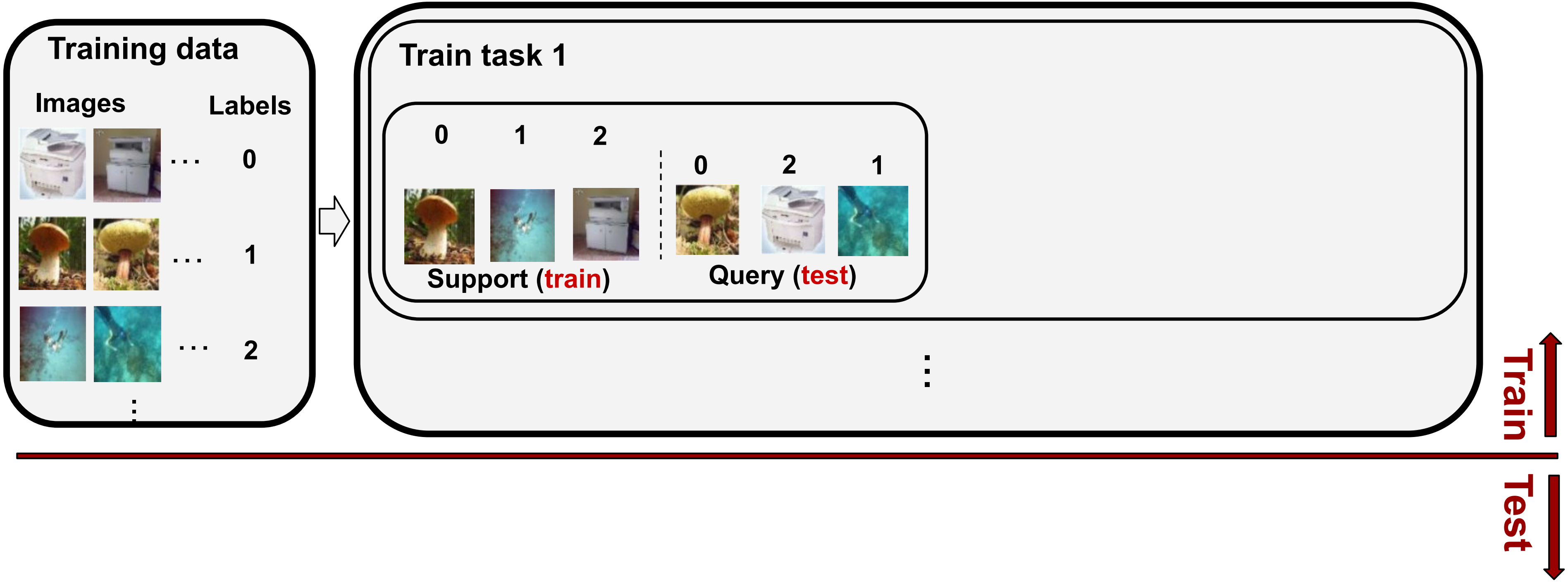


Inference algorithm: classify queries via nearest-neighbour comparisons

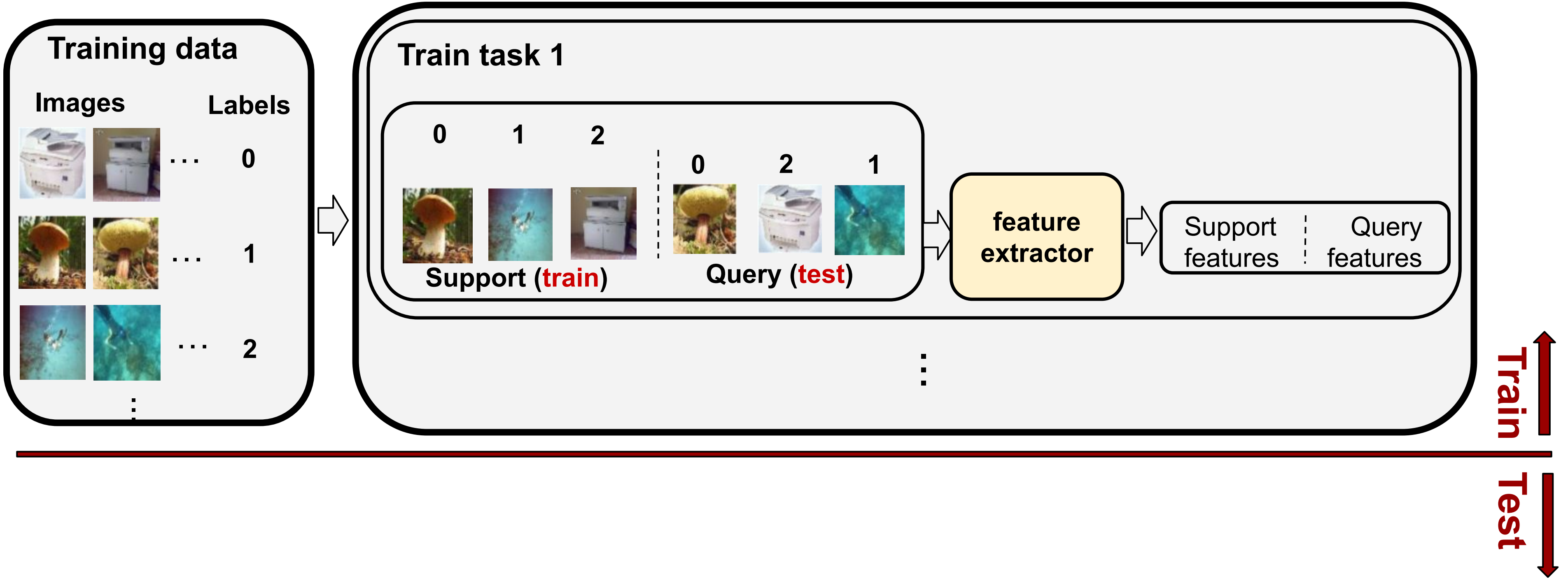
Transfer learning ‘baseline’



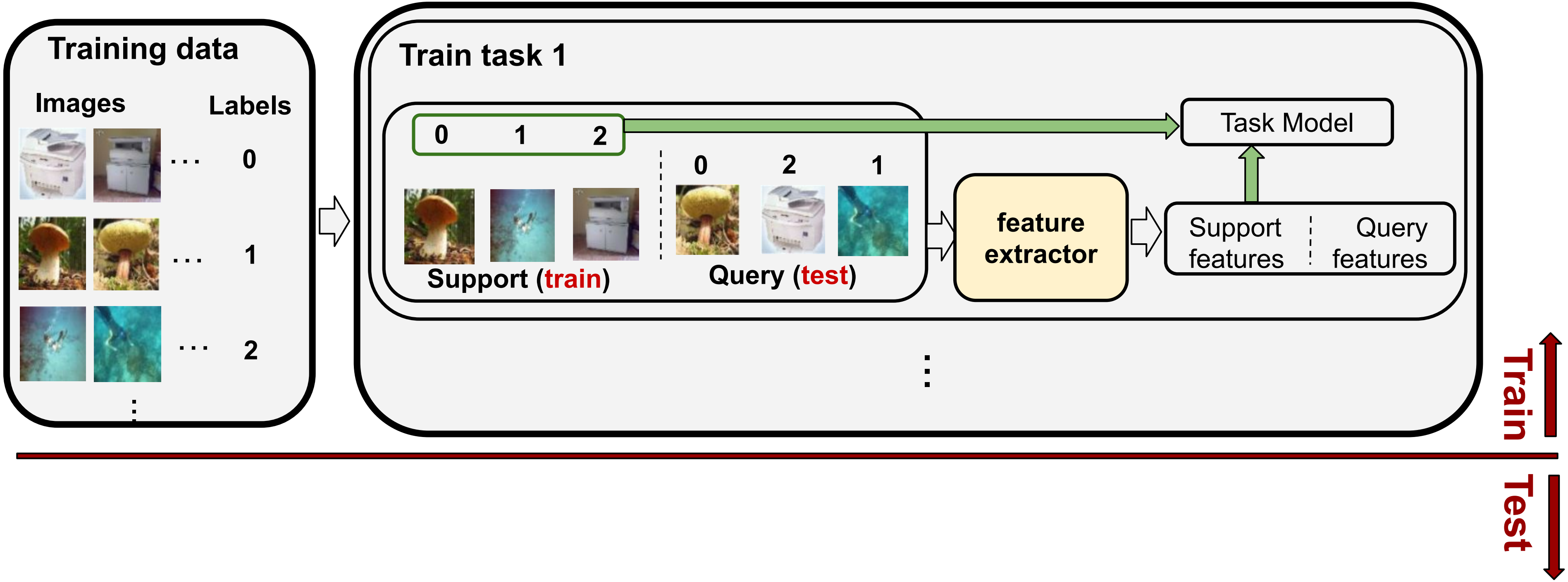
The meta-learning approach



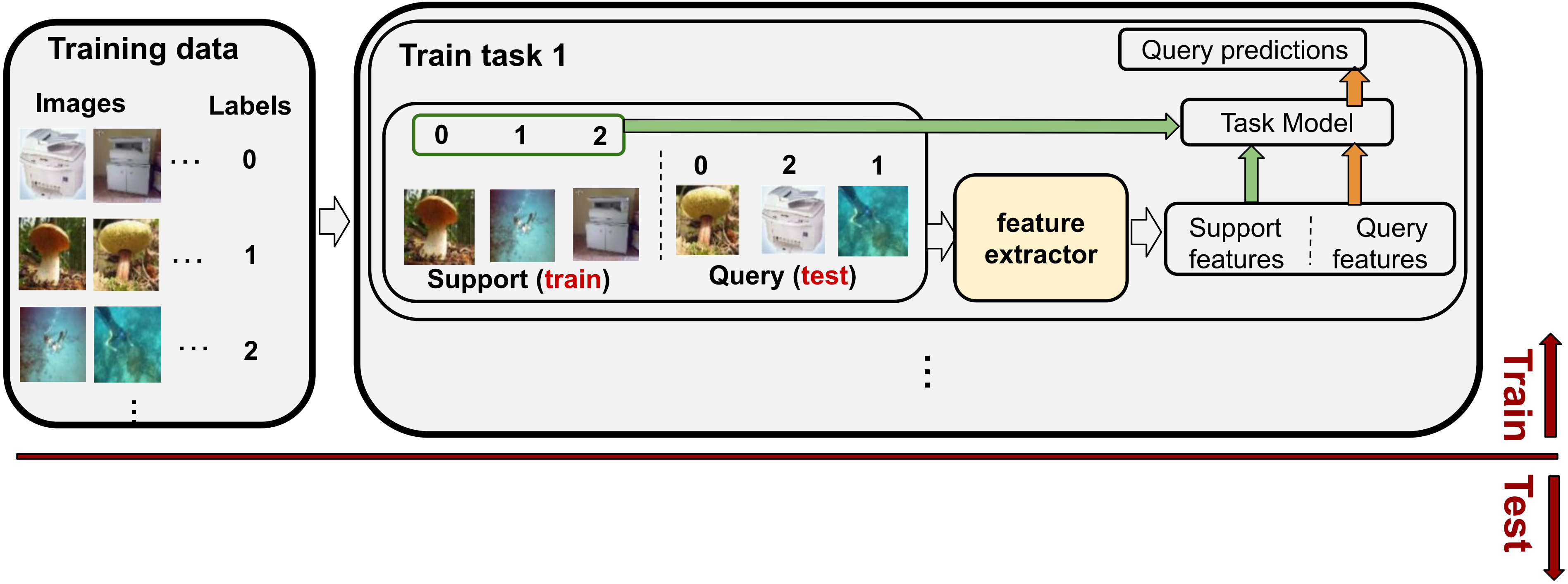
The meta-learning approach



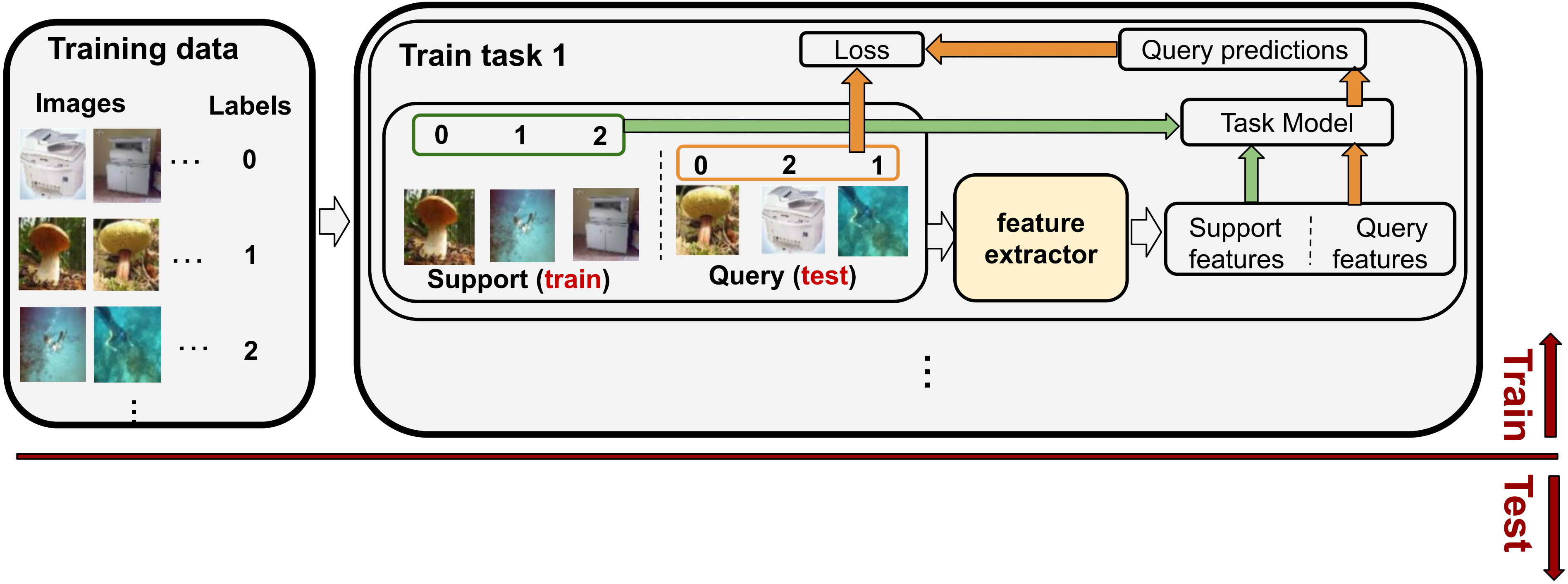
The meta-learning approach



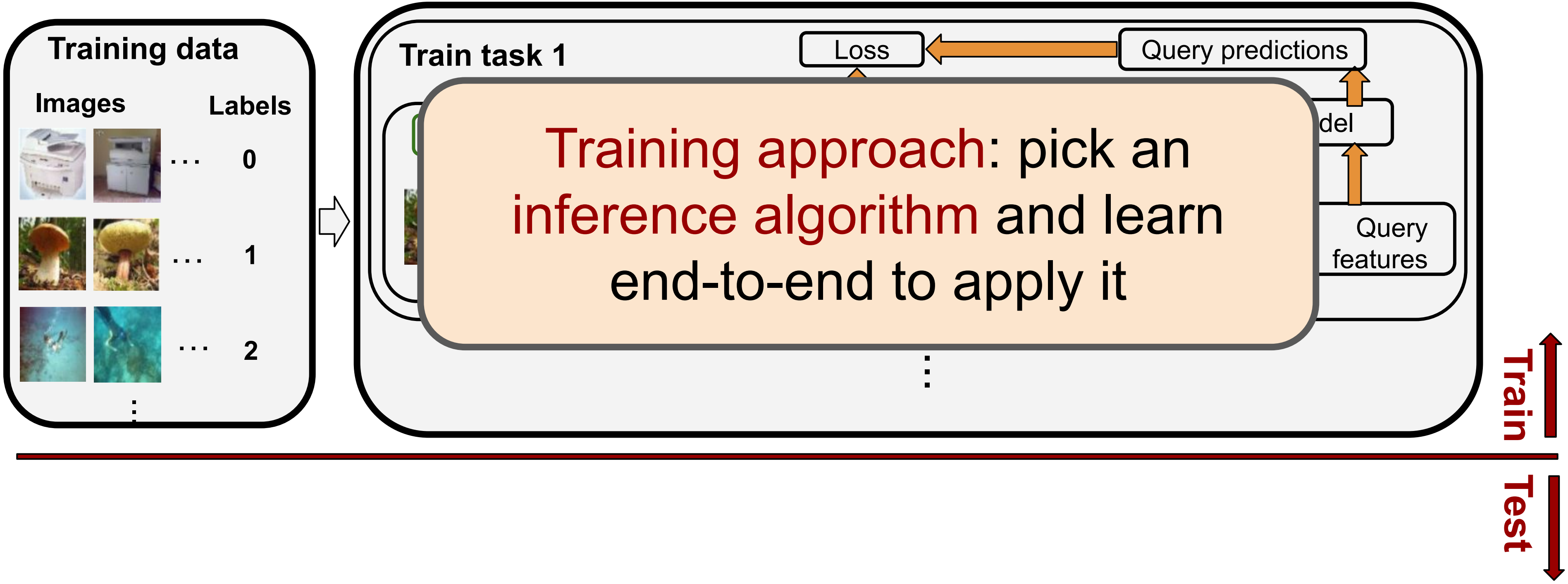
The meta-learning approach



The meta-learning approach

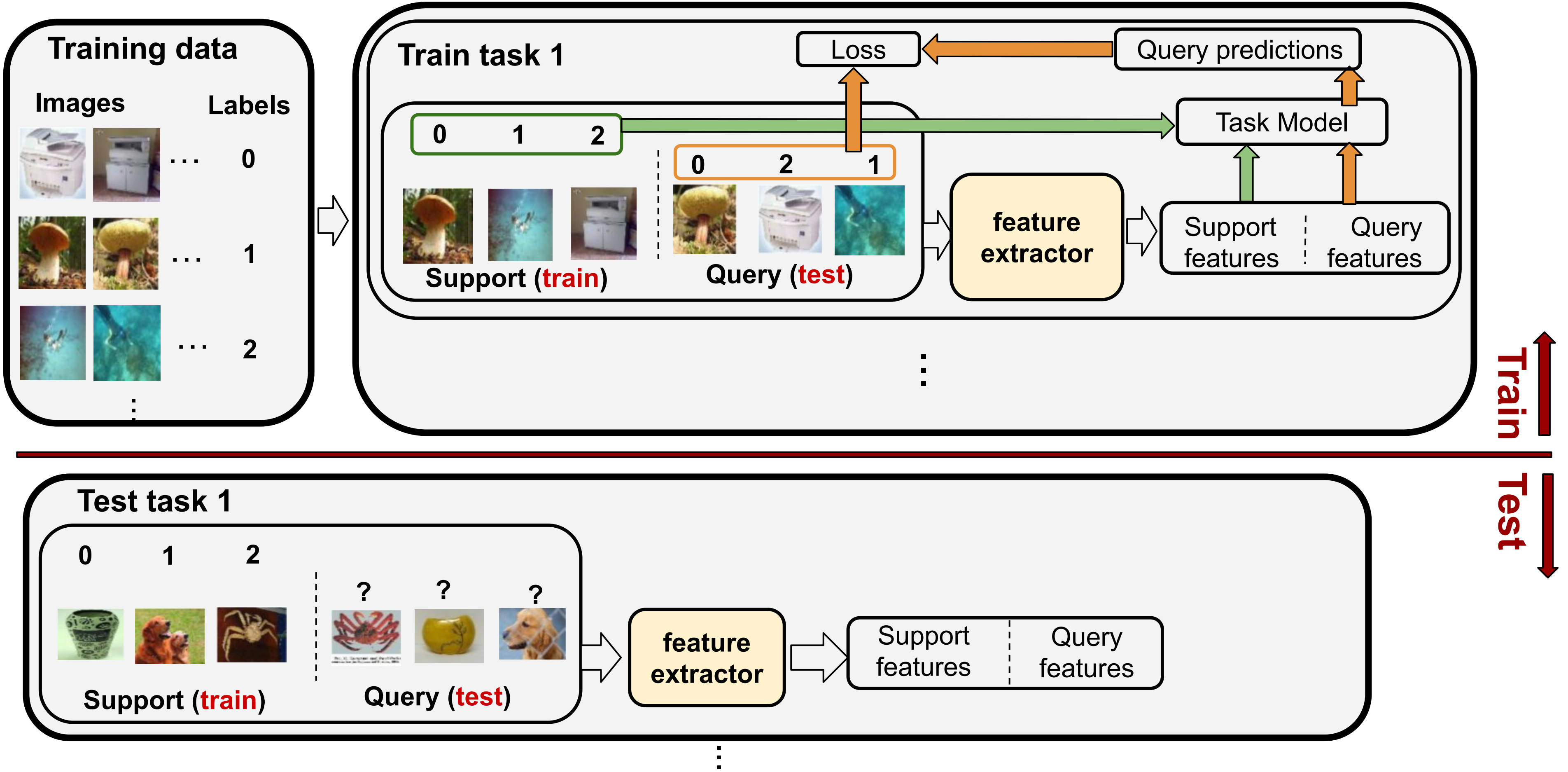


The meta-learning approach

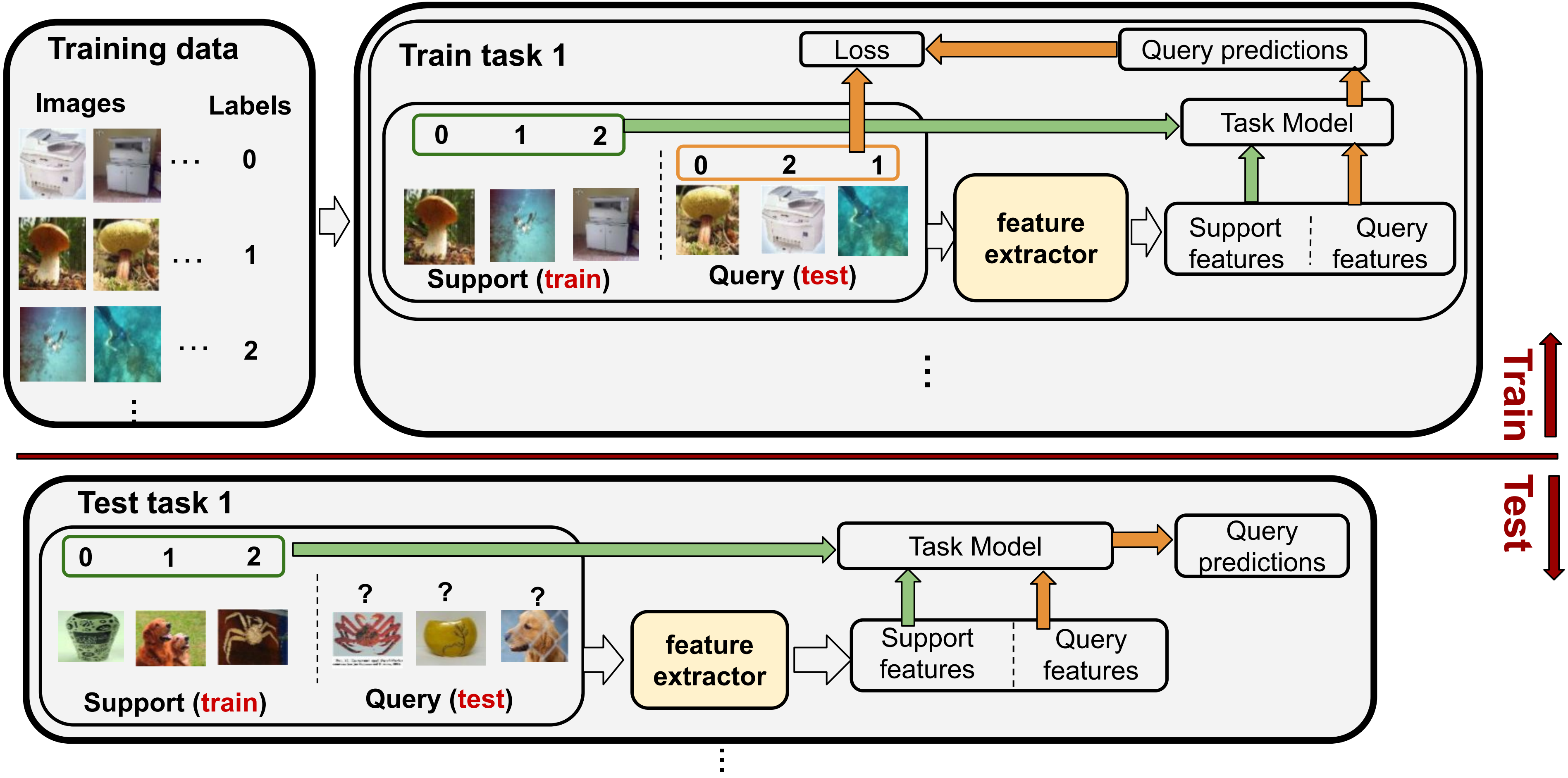


Training approach: pick an inference algorithm and learn end-to-end to apply it

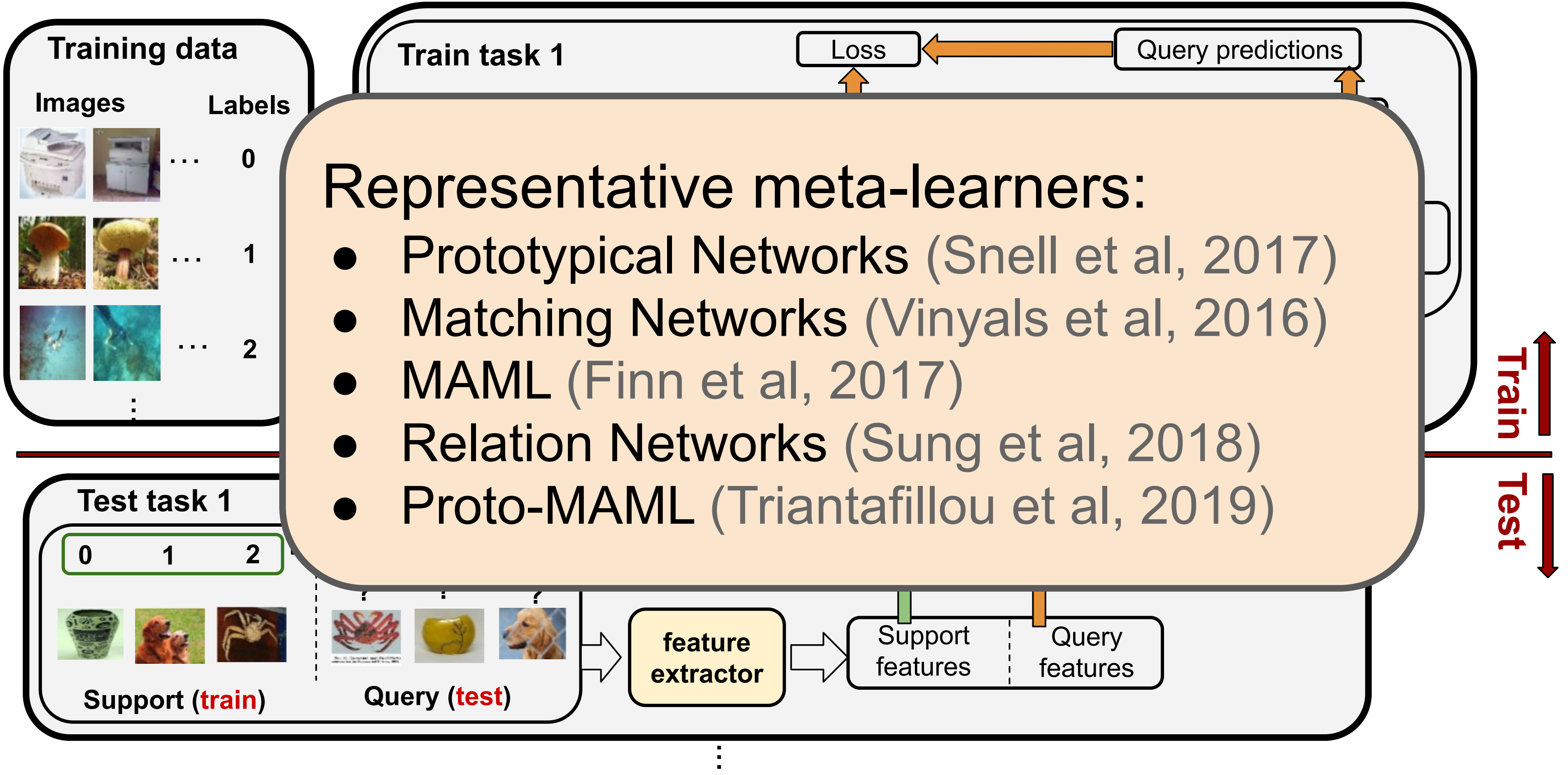
The meta-learning approach



The meta-learning approach



The meta-learning approach



Representative meta-learners:

- Prototypical Networks (Snell et al, 2017)
- Matching Networks (Vinyals et al, 2016)
- MAML (Finn et al, 2017)
- Relation Networks (Sung et al, 2018)
- Proto-MAML (Triantafillou et al, 2019)

Matching Networks (Vinyals et al., 2016)

Defines an inference algorithm for computing $p(y^*|\mathcal{S}, x^*)$ that labels query points x^* *conditioned on a support set* $\mathcal{S}$.

$$p(y^*|\mathcal{S}, x^*) = \sum_{i=1}^{|\mathcal{S}|} \alpha(x^*, x_i) y_i$$

where $\alpha(x^*, x_i) = softmax(\frac{f(x^*) \cdot f(x_i)}{||f(x^*)|| \cdot ||f(x_i)||})$ with f denoting the embedding function

Episodic Training Objective

$$\theta = \arg\max_{\theta} \mathop{\mathbb{E}}_{L} \mathop{\mathbb{E}}_{\mathcal{S},\mathcal{Q}\sim L} \sum_{(x^*,y^*)\in\mathcal{Q}} \log p_\theta(y^*|x^*, \mathcal{S})$$

where L denotes a label set e.g. $\{cat, dog\}$ sampled from **training classes** and $\mathcal{Q}$ is the query set.

Prototypical Networks (Snell et al., 2017)

- Follows episodic training
- But defines a different inference algorithm for computing $p(y^*|\mathcal{S}, x^*)$

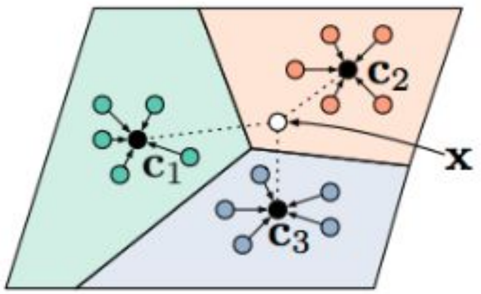


Figure: from (Snell et al., 2017)

Prototypical Network Classifier:

$$p_\theta(y^* = k|\mathcal{S}) = \frac{\exp(-d(f(x^*), c_k))}{\sum_{k' \in \{1,\dots,N\}} \exp(-d(f(x^*), c_{k'}))}$$

where c_k is the prototype for class k computed as: $c_k = \frac{1}{|\mathcal{S}_k|} \sum_{x_i \in \mathcal{S}_k} f(x_i)$ where $\mathcal{S}_k$ is the support points for class k and f is the embedding function.

Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017)

- MAML offers another way of meta-learning the learner parameters θ
- It learns a **common initialization** across tasks that is easily adaptable for each new task

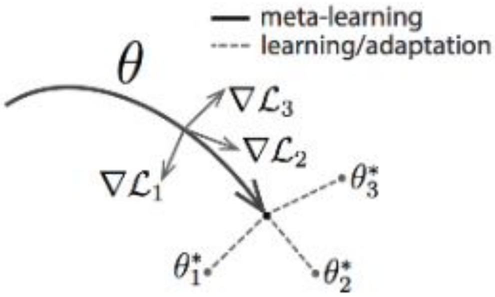


Figure: (Finn et al., 2017)

Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017)

Let $\mathcal{L}_\theta(S)$ denote the loss on set S using parameters θ.

MAML's training objective

In each training episode with support / query sets $\mathcal{S}, \mathcal{Q}$:

$$\min \mathop{\mathbb{E}}_{L} \mathop{\mathbb{E}}_{\mathcal{S},\mathcal{Q} \sim L} \mathcal{L}_{\theta - \alpha \nabla \mathcal{L}_\theta(\mathcal{S})}(\mathcal{Q})$$

As before L is a label set of training classes, e.g. $\{cat, dog\}$

This suggests a simple algorithm...

In each training episode with support / query sets $\mathcal{S}, \mathcal{Q}$:

1. **Learning:** Fine-tune θ for the episode's task via 1 step of SGD for the support loss: $\theta' = \theta - \alpha \nabla \mathcal{L}_\theta(\mathcal{S})$
2. **Meta-Learning:** Update θ (by SGD) to minimize the loss on the query set $\mathcal{L}_{\theta'}(\mathcal{Q})$, computed using the updated θ'

Proto-MAML (Triantafillou et al., 2019)

Aims to capture the best of Prototypical Nets and MAML: inductive bias of the former and flexibility of the latter.

Re-interpreting Prototypical Networks as a linear classifier

Let c_k be the prototype of class k, f the embedding function and x^* a query. The 'logit' for x^* being classified as k is:

$$\begin{aligned} -||f(x^*) - c_k||^2 &= -f(x^*)^T f(x^*) + 2c_k^T f(x^*) - c_k^T c_k = \\ &= 2c_k^T f(x^*) - ||c_k||^2 + constant \end{aligned}$$

which defines a linear layer: $w_k = 2c_k$ and $b_k = -||c_k||^2$, as also shown in (Snell et al., 2017).

Proto-MAML: For each task, initialize MAML's linear layer from the above prototypical classifier, then do task adaptaton as usual.

Relation Networks (Sung et al., 2018)

The inference algorithm is:

- Concatenate the query with each class prototype, and feed each such pair through a *relation module*
- The relation module predicts how likely the given query belongs to the given prototype (binary classification)
- Uses mean square error as the loss

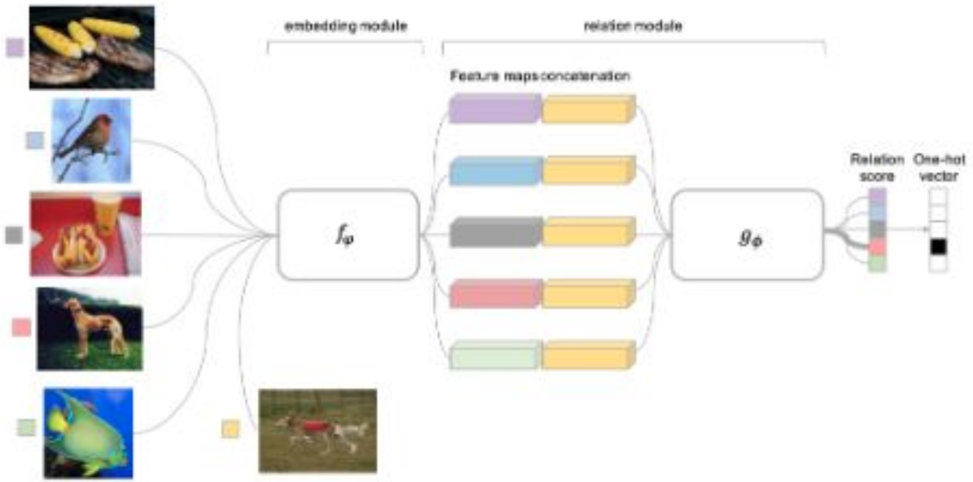


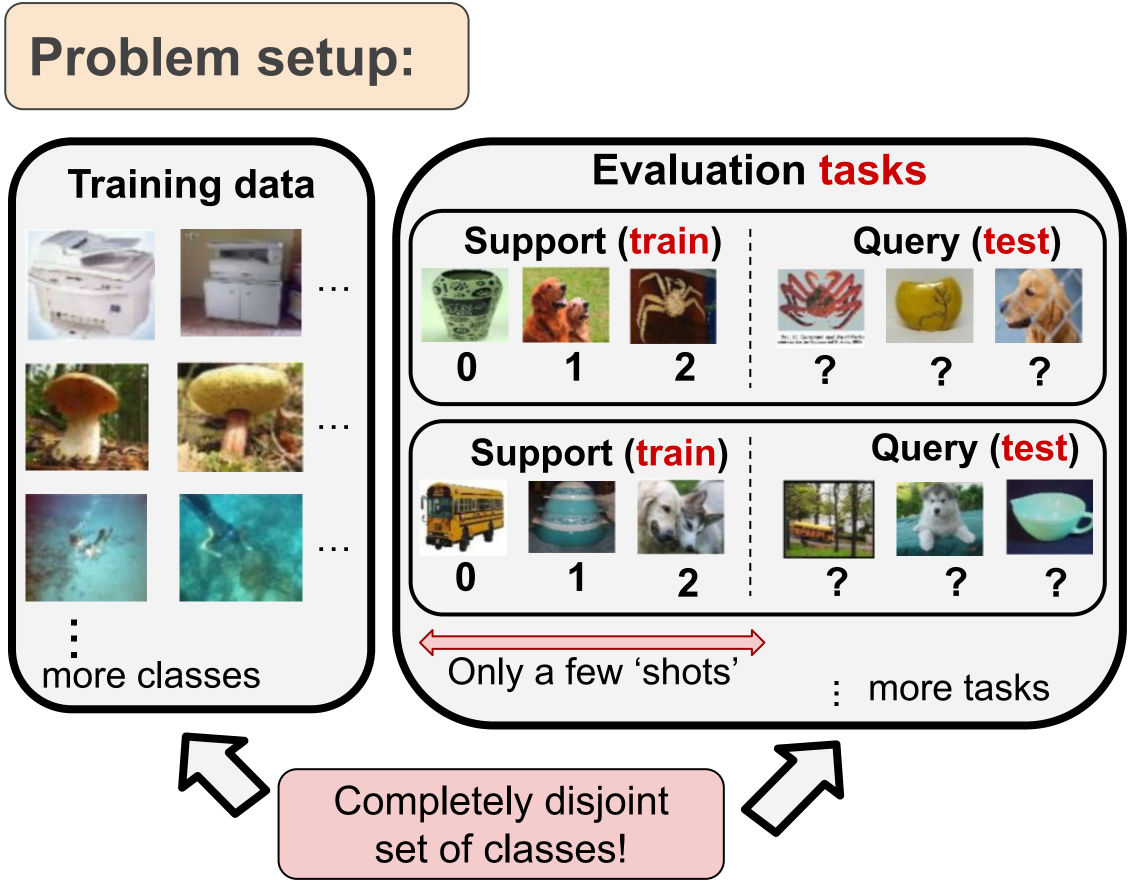
Figure: Figure from (Sung et al., 2018)

Revisiting the setup for few-shot classification

- The de facto evaluation involves benchmarks that:
  - Are comprised of a single dataset, so training and testing classes aren’t too different visually
  - Are class balanced
  - Have homogeneous tasks

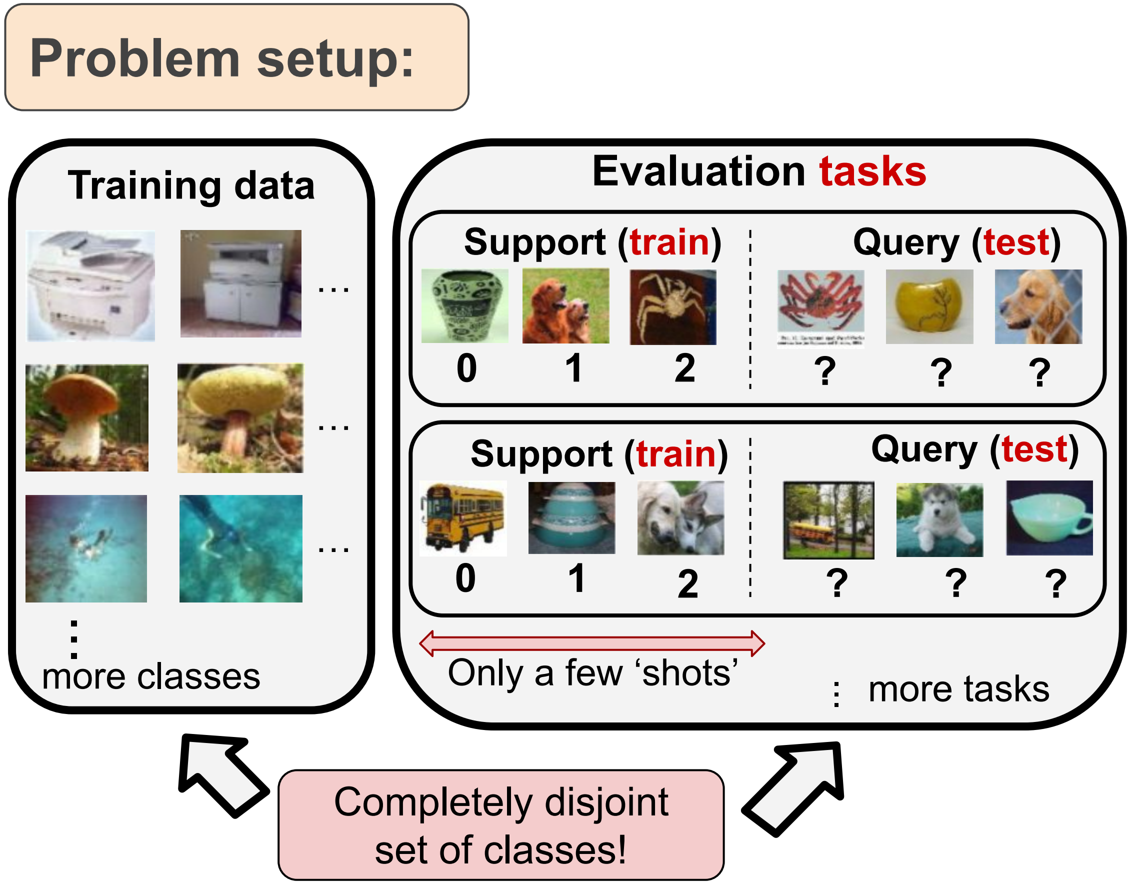
Few-shot Classification

- Learn new classes from few examples
- Practically important and scientifically interesting

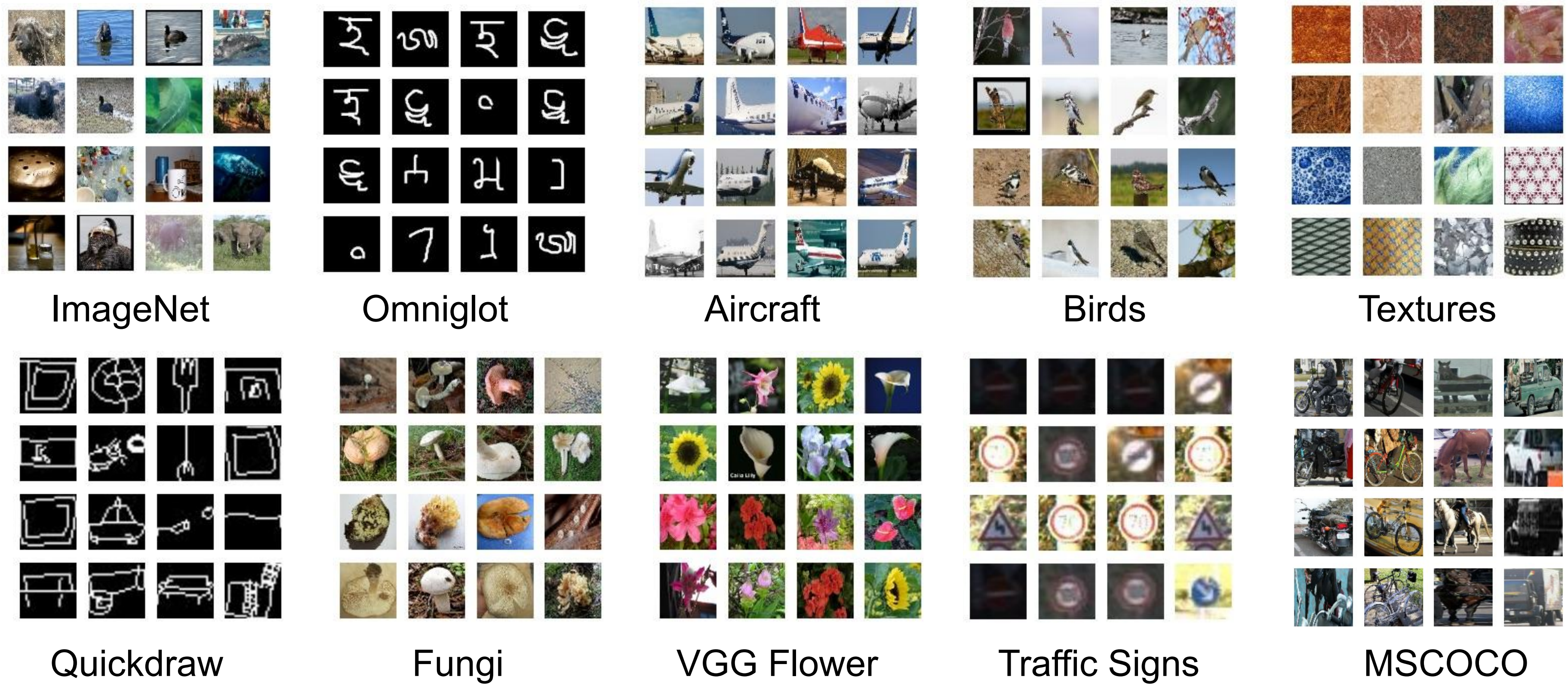


Few-shot Classification

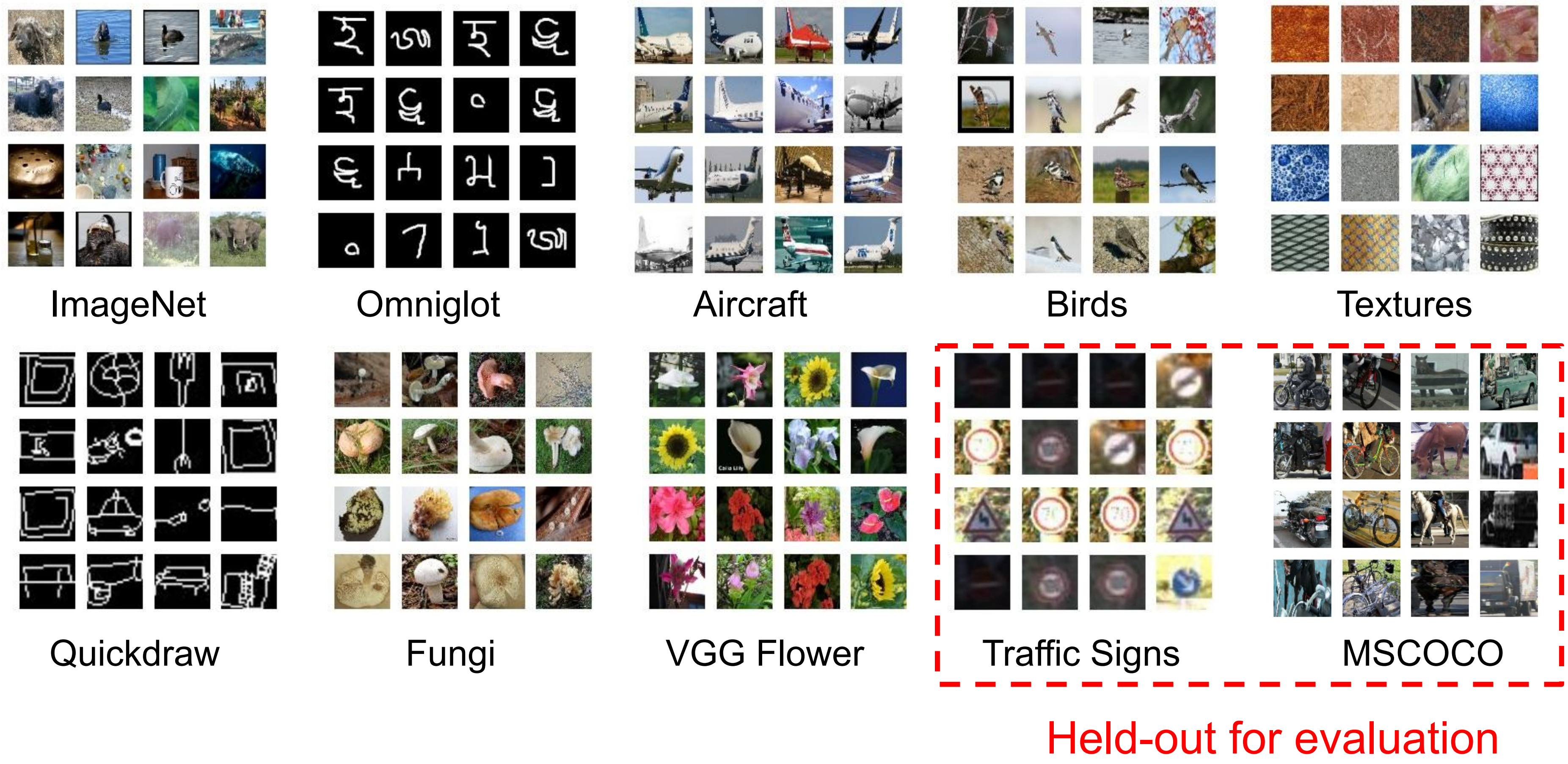
- Learn new classes from few examples
- Practically important and scientifically interesting
- Previous datasets (e.g. mini-ImageNet) don't evaluate on substantially different classes
- We propose a benchmark to address this



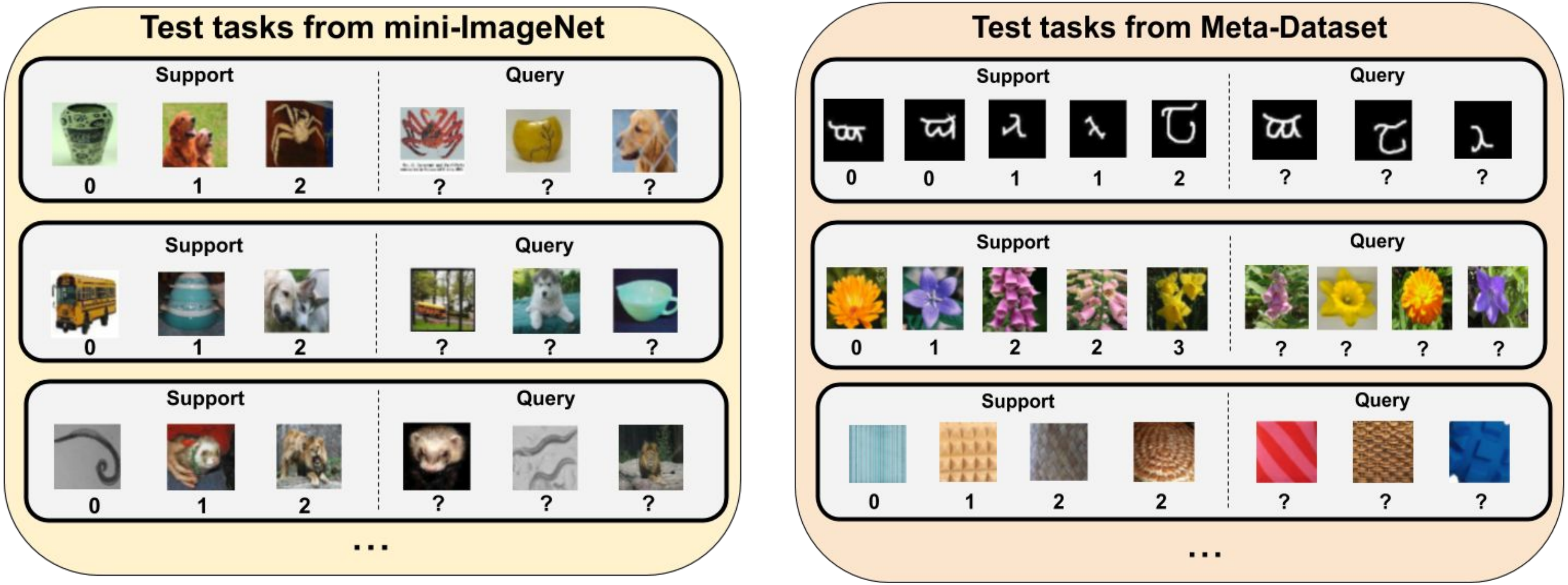
Introducing Meta-Dataset



Introducing Meta-Dataset



What's new in Meta-Dataset?



Meta-Dataset is much larger scale (larger image resolution and > 5K classes in total) than previous benchmarks (e.g. mini-ImageNet has 100 classes), and its tasks:

1. Are diverse: originate from 10 datasets.
2. Have variable ways and shots, imbalance and degrees of fine-grainedness.

Research Investigation

We investigate several important questions:

- Do we generalize better when training on more datasets?
- Can we benefit from increasing numbers of ‘shots’ at test time?
- Non-meta-learning ‘baselines’ are gaining popularity. Are these sufficient for performing well on Meta-Dataset?

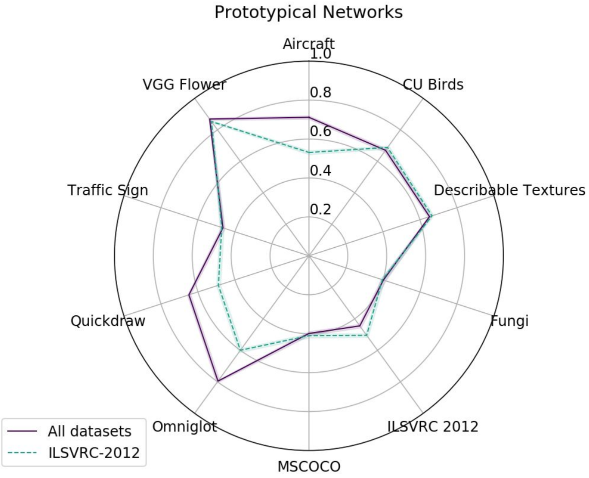
Research Investigation

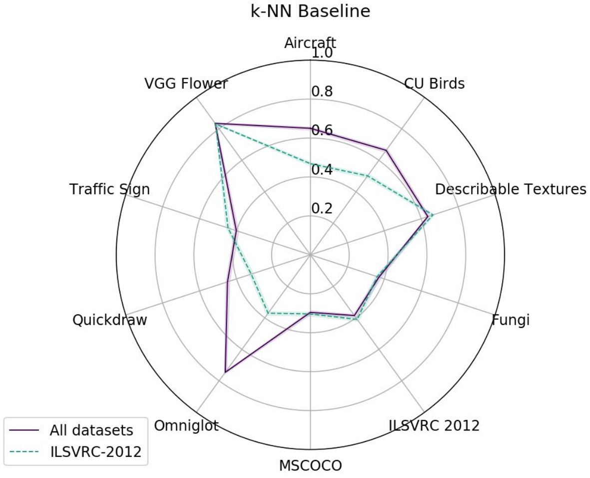
We investigate several important questions:

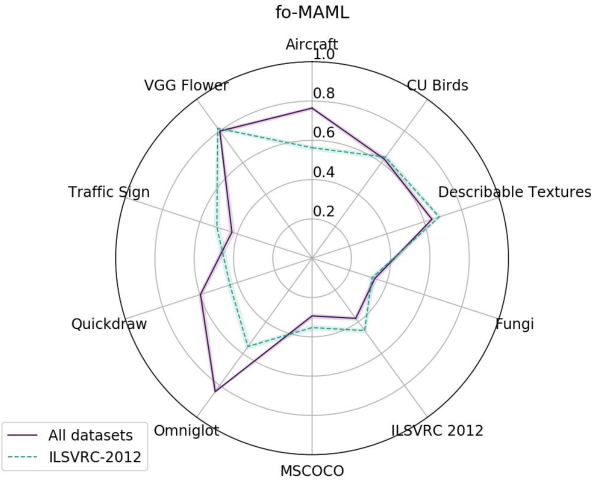
- Do we generalize better when training on more datasets?
- Can we benefit from increasing numbers of ‘shots’ at test time?
- Non-meta-learning ‘baselines’ are gaining popularity. Are these sufficient for performing well on Meta-Dataset?

Effect of training on all datasets versus ImageNet only

- Training on (the train classes of) all datasets instead of (the train classes of) ImageNet only is not helpful for all evaluation datasets.
- Further work is needed to better absorb diverse training information.







Visualization technique inspired by [Dvornik et al]() (arXiv:2003.09338)

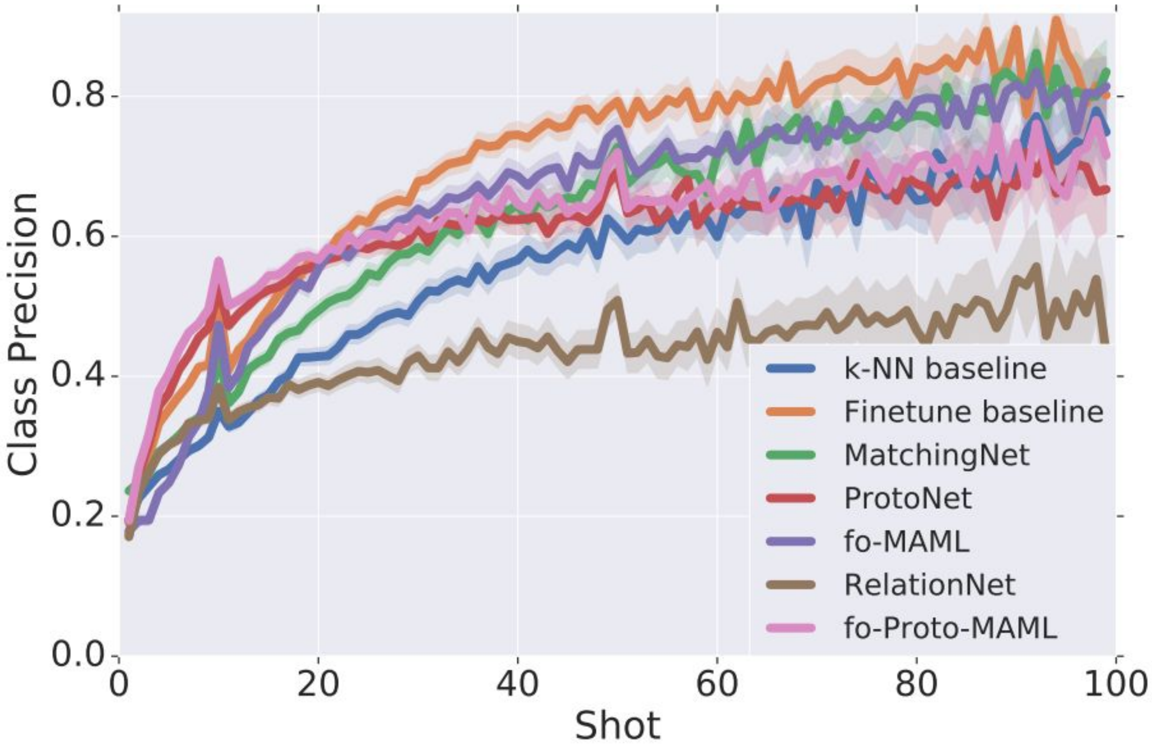
Research Investigation

We investigate several important questions:

- Do we generalize better when training on more datasets?
- Can we benefit from increasing numbers of ‘shots’ at test time?
- Non-meta-learning ‘baselines’ are gaining popularity. Are these sufficient for performing well on Meta-Dataset?

Test-time performance as a function of ‘shots’

- Trade-off: different models do better on different ‘shot’ settings at test time
- **Unmet desideratum**: optimally leverage any number of shots at test time



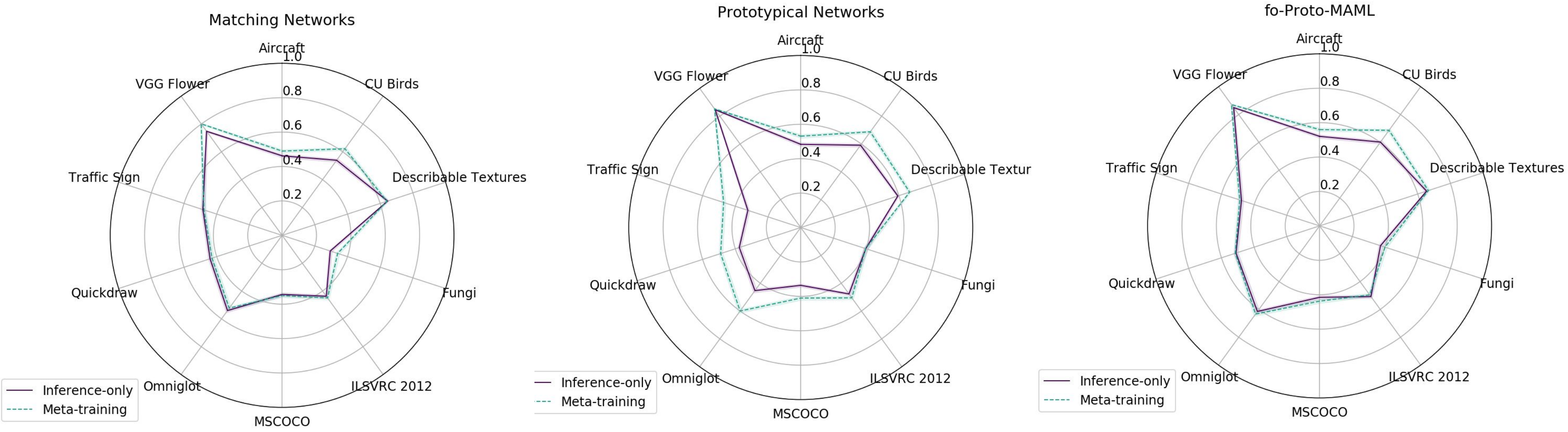
Research Investigation

We investigate several important questions:

- Do we generalize better when training on more datasets?
- Can we benefit from increasing numbers of ‘shots’ at test time?
- Non-meta-learning ‘baselines’ are gaining popularity. Are these sufficient for performing well on Meta-Dataset?

Does meta-learning improve over pre-training?

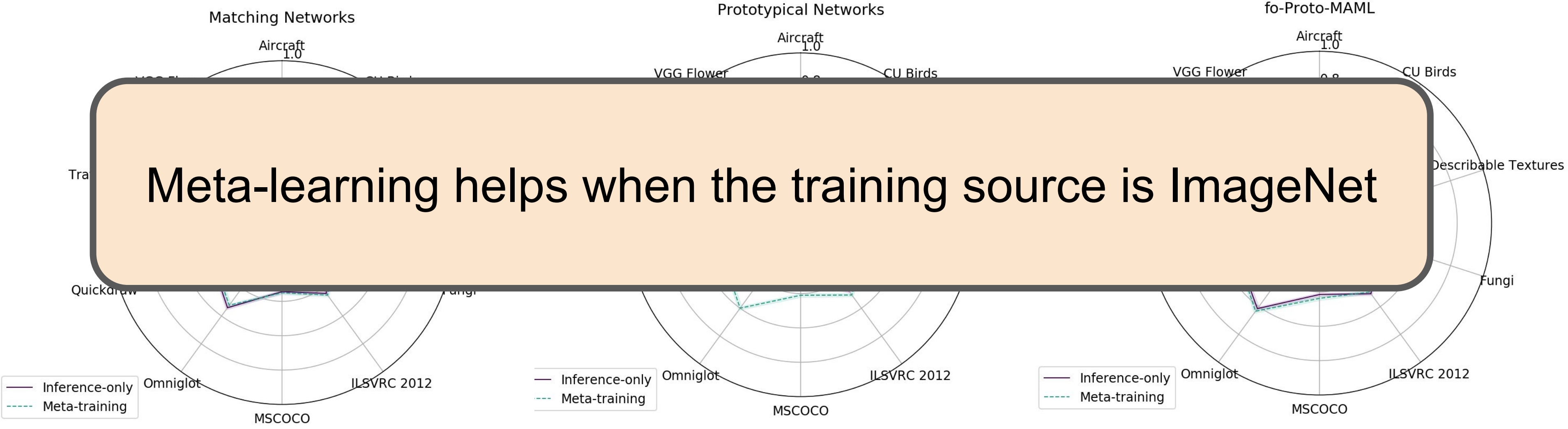
- 'Inference-only' version of each meta-learner: pre-trains a feature extractor and **at evaluation time** applies the inference algorithm of the corresponding meta-learner.



Training source: ImageNet only

Does meta-learning improve over pre-training?

- ‘Inference-only’ version of each meta-learner: pre-trains a feature extractor and **at evaluation time** applies the inference algorithm of the corresponding meta-learner.

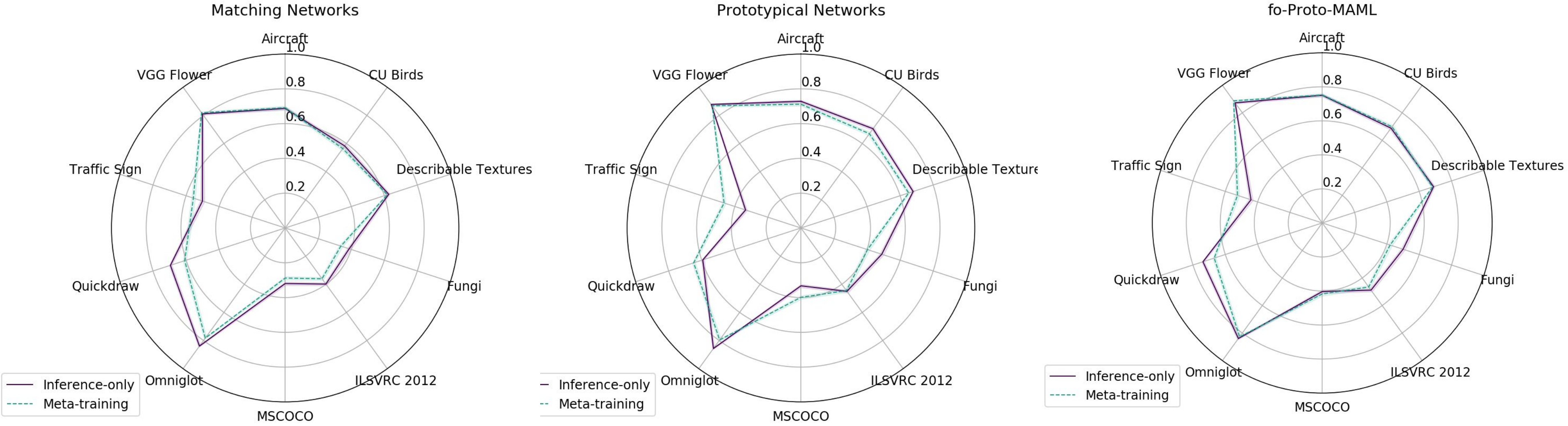


Meta-learning helps when the training source is ImageNet

Training source: ImageNet only

Does meta-learning improve over pre-training?

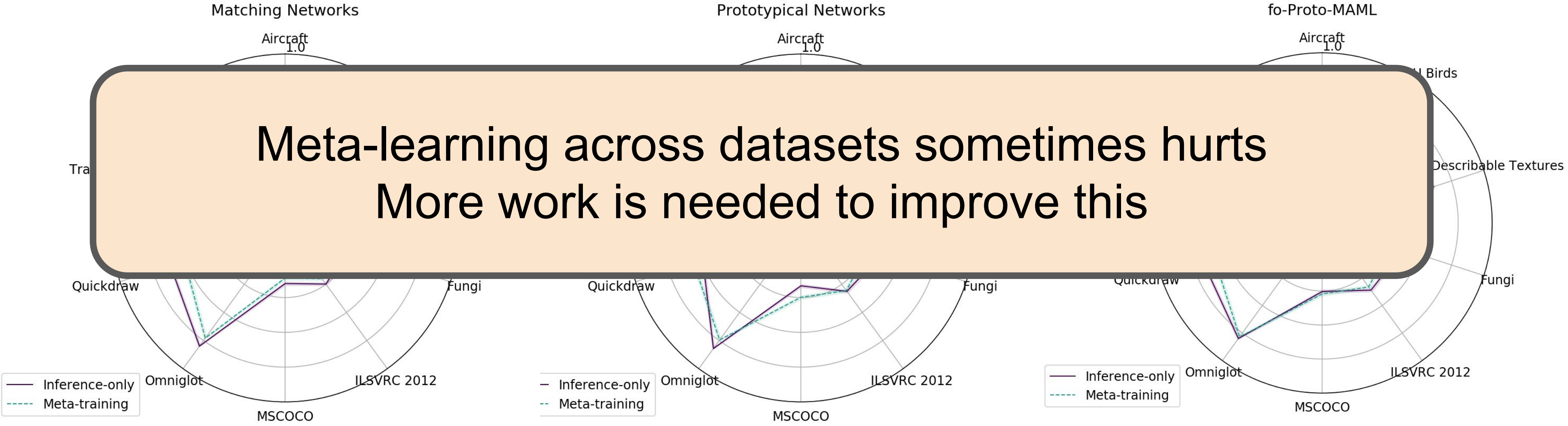
- ‘Inference-only’ version of each meta-learner: pre-trains a feature extractor and **at evaluation time** applies the inference algorithm of the corresponding meta-learner.



Training source: All datasets

Does meta-learning improve over pre-training?

- ‘Inference-only’ version of each meta-learner: pre-trains a feature extractor and **at evaluation time** applies the inference algorithm of the corresponding meta-learner.



Meta-learning across datasets sometimes hurts
More work is needed to improve this

Training source: All datasets

Progress on Meta-Dataset

Shout out to recent papers reporting results on Meta-Dataset:

- **Requeima et al, 2019.** Fast and Flexible Multi-Task Classification Using Conditional Neural Adaptive Processes.
- **Bateni et al, 2020.** Improved Few-Shot Visual Classification.
- **Saikia et al, 2020.** Optimized Generic Feature Learning for Few-shot Classification across Domains.
- **Yinbo et al, 2020.** A New Meta-Baseline for Few-Shot Learning.
- **Dvornik et al, 2020.** Selecting Relevant Features from a Universal Representation for Few-shot Classification.

Our code is public: https://github.com/google-research/meta-dataset.

More realistic forms of few-shot learning

Interesting directions that study less artificial few-shot classification tasks:

- **Semi-supervised few-shot learning**: in addition to the small labeled support set, there is a pool of possibly relevant unlabeled examples available.
    - Meta-Learning for Semi-Supervised Few-shot Classification (Ren et al, 2018)
    - Learning to Self-Train for Semi-Supervised Few-shot Classification (Li et al, 2019)
- **Incremental few-shot learning:** retain the ability to remember the train classes while learning about test classes
    - Dynamic Few-Shot Visual Learning without Forgetting (Gidaris et al, 2018)
    - Incremental Few-Shot Learning with Attention Attractor Networks (Ren et al, 2018)
- **Continual learning**: does not consider each task as an isolated learning problem.
    - Meta-Learning Representations for Continual Learning (Javed and White, 2019)
    - Learning to Continually Learn (Beaulieu et al, 2019)

Thank you for listening!