A CLOSER LOOK AT FEW-SHOT CLASSIFICATION

W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang
Published at ICLR 2019

Presented by Kaustubh Sridhar on 02/15/2021 in CIS 620

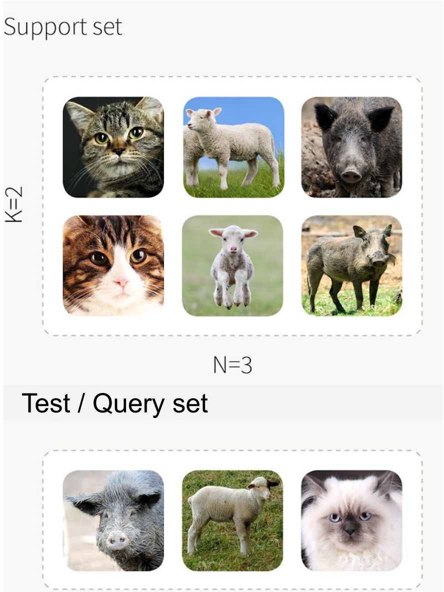
Contents of this presentation

- What is few-shot classification
    - N-way k-shot task
- Few-shot classification’s training paradigms
    - Baseline: transfer learning
    - Meta-learning
- Meta-learning methods
    - Distance metric based: Matching Net, Prototype Net, Relation Net
    - Initialisation based: MAML (Model Agnostic Meta Learning)
- Empirical comparison between baseline and meta-learning methods

Few Shot Classification

Given abundant training examples for the base classes, few-shot learning algorithms aim to learn to recognizing novel classes with a limited amount of labeled examples.

At test time : n-way k-shot task

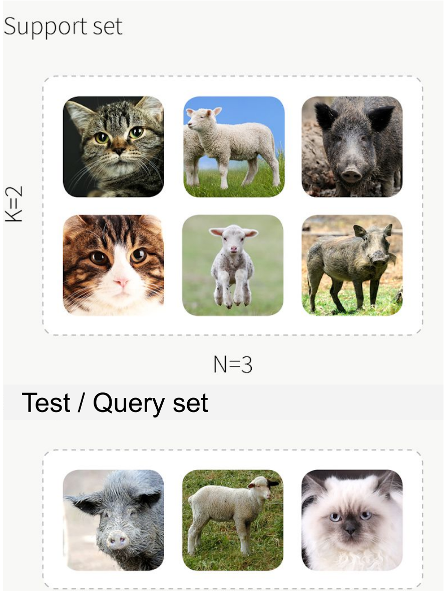


[1] Borealis AI, Tutorial on few-shot learning and meta-learning, link
[2] Zsolt Kira, "Low-Label ML Formulations", CS 4803 Course Presentation, link
[3] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Few Shot Classification

Given abundant training examples for the base classes, few-shot learning algorithms aim to learn to recognizing novel classes with a limited amount of labeled examples.

At test time : n-way k-shot task



Given: limited novel-labelled Support Set with K images from each of N novel classes

All classes are numbered (*e.g.* mini-imagenet dataset has 100 classes numbered 0-99) and a label of an image is the number of the class it belongs to.

Task: classify test (*a.k.a. query*) set images with "novel labels" (labels not present in base data but available in support set)

The test set is more often called the query set because the support set is not available during *training* and only given at the *testing* phase. Thus some authors refer to the combined support+query sets as test sets.

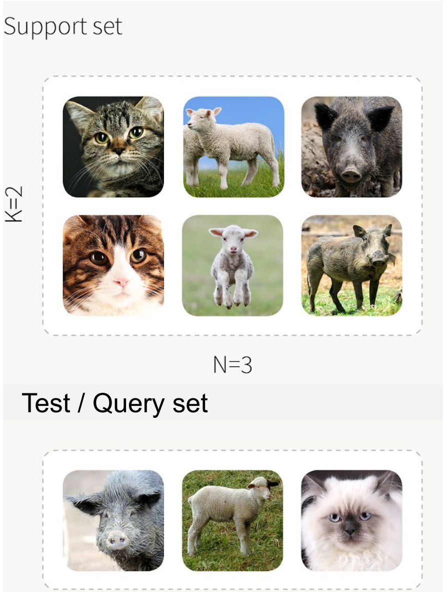
[1] Borealis AI, Tutorial on few-shot learning and meta-learning, link
[2] Zsolt Kira, "Low-Label ML Formulations", CS 4803 Course Presentation, link
[3] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Few Shot Classification

How do I use this?

Given abundant training examples for the base classes, few-shot learning algorithms aim to learn to recognizing novel classes with a limited amount of labeled examples.

At test time : n-way k-shot task



Given: limited novel-labelled Support Set with K images from each of N novel classes

All classes are numbered (*e.g.* mini-imagenet dataset has 100 classes numbered 0-99) and a label of an image is the number of the class it belongs to.

Task: classify test (*a.k.a. query*) set images with "novel labels" (labels not present in base data but available in support set)

The test set is more often called the query set because the support set is not available during *training* and only given at the *testing* phase. Thus some authors refer to the combined support+query sets as test sets.

[1] Borealis AI, Tutorial on few-shot learning and meta-learning, link
[2] Zsolt Kira, "Low-Label ML Formulations", CS 4803 Course Presentation, link
[3] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Few Shot Classification: Training (with base class data) paradigms

Transfer learning baselines

Meta-learning

1. Pre-training: supervised learning of model on base data
2. Fine-tuning: supervised learning of (parts of or whole) model on labelled support data
3. Testing on test / query dataset

"Learning to learn": a paradigm specifically for the k-shot n-way task that uses base data to "learn to learn", *i.e.* learn a *meta-learner*, and applies the meta-learner on the testing phase (support + query) data.

Expected to not perform well on test / query dataset without large support set (in fine-tuning stage)

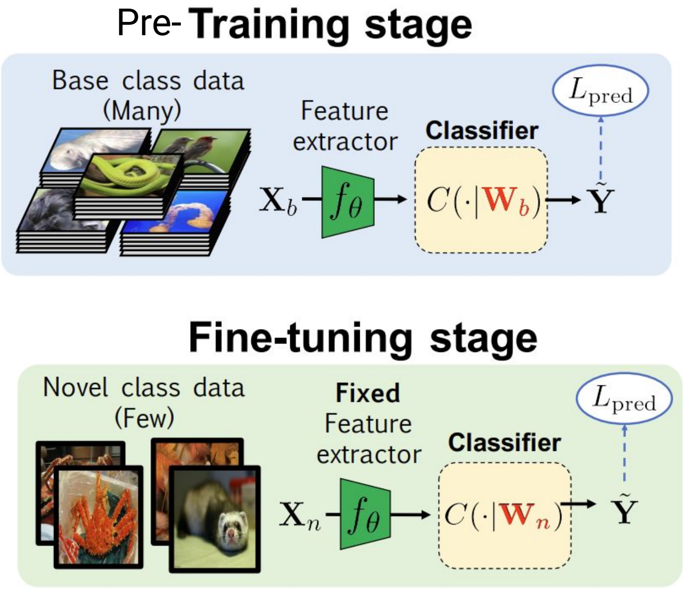
[1] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Few Shot Classification: Training (with base class data) paradigms

Transfer learning baselines

1. Pre-training: supervised learning of model on base data
2. Fine-tuning: supervised learning of (parts of or whole) model on labelled support data
3. Testing on test / query dataset

Pre-



Retrain only classifier

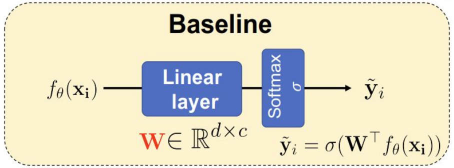
[1] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Few Shot Classification: Training (with base class data) paradigms

Transfer learning baselines

1. Pre-training: supervised learning of model on base data
2. Fine-tuning: supervised learning of (parts of or whole) model on labelled support data
3. Testing on test / query dataset

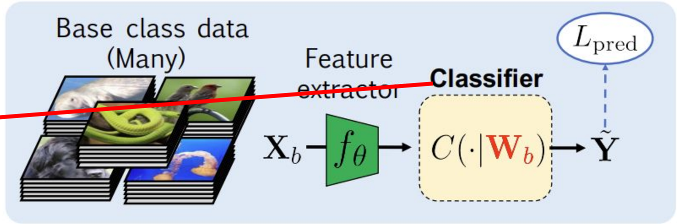
Choice of Classifier



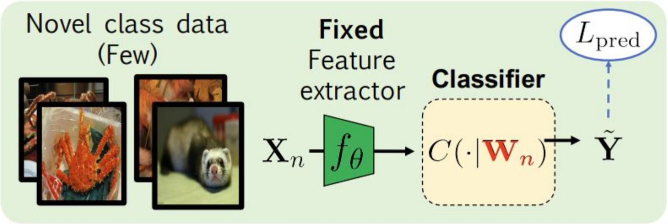
(Standard procedure)

Commonly seen last layer (*a.k.a.* logits) in a deep neural network classifying image into one of classes by min loss = f(predicted label probability vector, true one-hot encoded label).

Pre- **Training stage**



Fine-tuning stage



Retrain only classifier

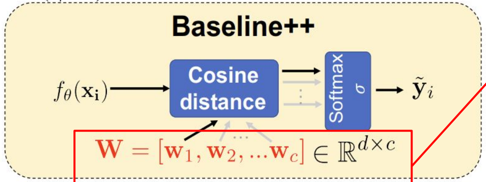
[1] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Few Shot Classification: Training (with base class data) paradigms

Transfer learning baselines

1. Pre-training: supervised learning of model on base data
2. Fine-tuning: supervised learning of (parts of or whole) model on labelled support data
3. Testing on test / query dataset

Reimagining the Classifier



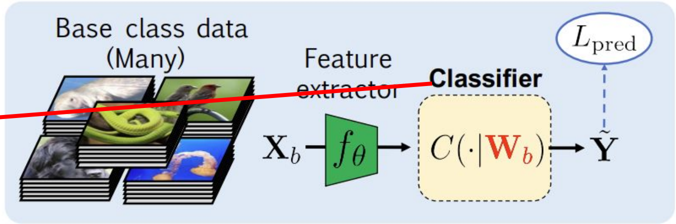
Classifier weight thought of as d-dimensional weight vectors for each of c classes

Similarity scores for each class $[s_{i,1}, s_{i,2}, \cdots, s_{i,c}]$ obtained with cosine distance between logits (feature $f_\theta(x_i)$) and weights

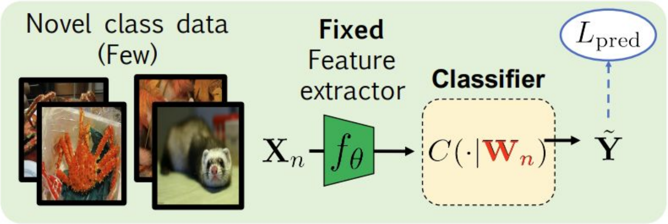
$$s_{i,j} = f_\theta(\mathbf{x}_i)^\top \mathbf{w}_j / \|f_\theta(\mathbf{x}_i)\| \|\mathbf{w}_j\|$$

Labels are the same as in baseline (one-hot encoded vectors) but a value of 1 can be thought of as a similarity score of 1

Pre- **Training stage**



Fine-tuning stage



Retrain only classifier

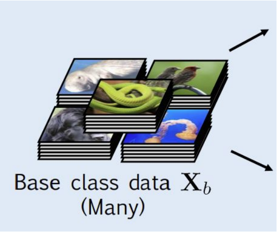
[1] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, “A Closer Look At Few-Shot Classification”, Proceedings of the International Conference on Learning Representations (ICLR) 2019

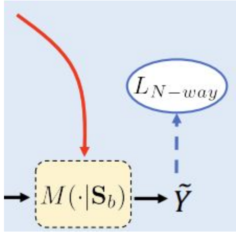
Few Shot Classification: Training (with base class data) paradigms

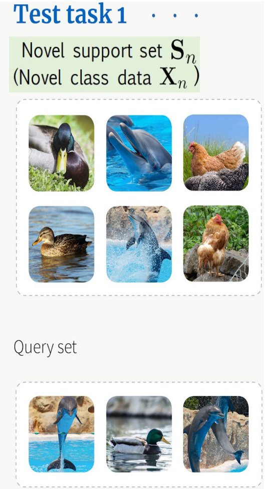
Meta-learning

"Learning to learn": a paradigm specifically for the k-shot n-way task that uses base data to "learn to learn", *i.e.* learn a *meta-learner*, and applies the meta-learner on the testing phase (support + query) data.

Randomly sample N classes and rearrange base class data into meta-training tasks that simulate test (usually same k, N).







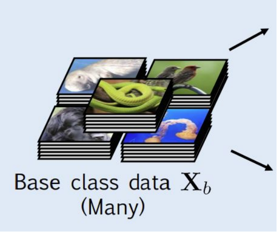
[1] Borealis AI, Tutorial on few-shot learning and meta-learning, link
[2] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Few Shot Classification: Training (with base class data) paradigms

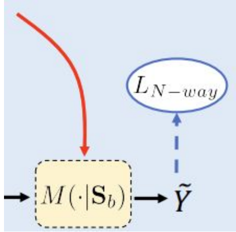
Meta-learning

"Learning to learn": a paradigm specifically for the k-shot n-way task that uses base data to "learn to learn", *i.e.* learn a *meta-learner*, and applies the meta-learner on the testing phase (support + query) data.

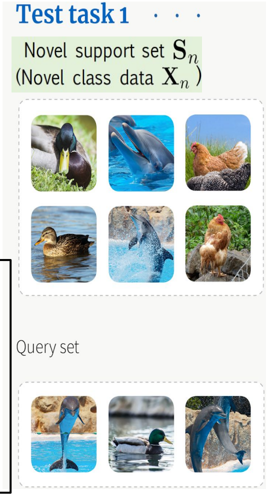
Randomly sample N classes and rearrange base class data into meta-training tasks that simulate test (usually same k, N).



Build a support set conditioned model by min N-way loss = f(label prediction of query images, true query labels)



A meta-learner that has learnt how to learn **from** support images **to classify** query images



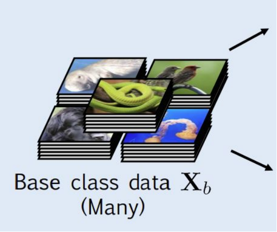
[1] Borealis AI, Tutorial on few-shot learning and meta-learning, link
[2] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Few Shot Classification: Training (with base class data) paradigms

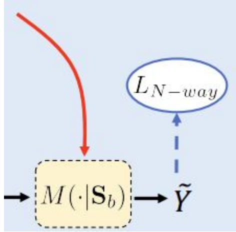
Meta-learning

"Learning to learn": a paradigm specifically for the k-shot n-way task that uses base data to "learn to learn", *i.e.* learn a *meta-learner*, and applies the meta-learner on the testing phase (support + query) data.

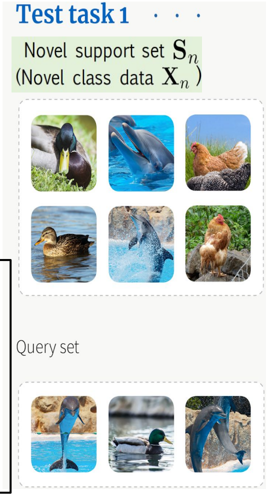
Randomly sample N classes and rearrange base class data into meta-training tasks that simulate test (usually same k, N).



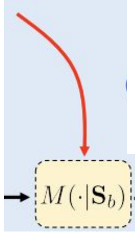
Build a support set conditioned model by min N-way loss = f(label prediction of query images, true query labels)



A meta-learner that has learnt how to learn **from** support images **to classify** query images



Evaluate model on novel-label test tasks



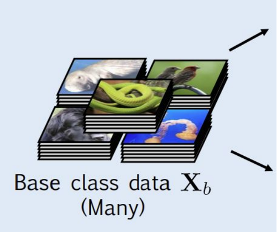
[1] Borealis AI, Tutorial on few-shot learning and meta-learning, link
[2] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Few Shot Classification: Training (with base class data) paradigms

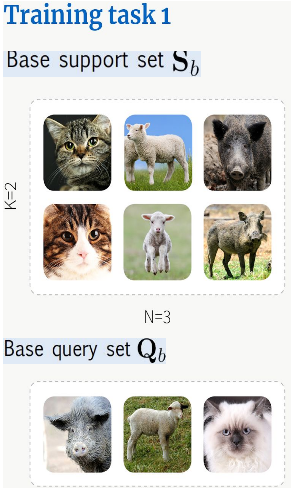
Meta-learning

"Learning to learn": a paradigm specifically for the k-shot n-way task that uses base data to "learn to learn", *i.e.* learn a *meta-learner*, and applies the meta-learner on the testing phase (support + query) data.

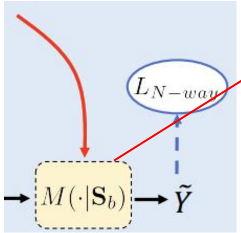
Randomly sample N classes and rearrange base class data into meta-training tasks that simulate test (usually same k, N).



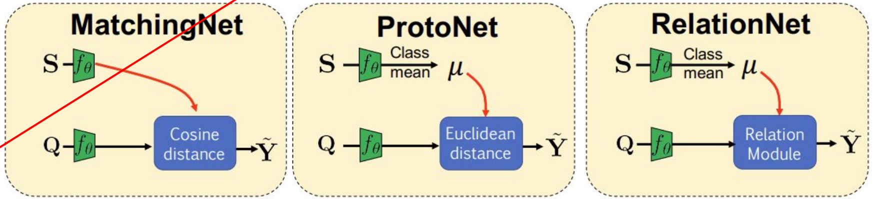
Meta-training stage

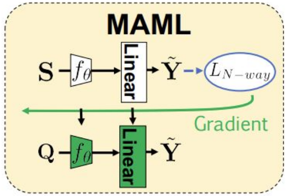


Build a support set conditioned model by min N-way loss



Choice of Model





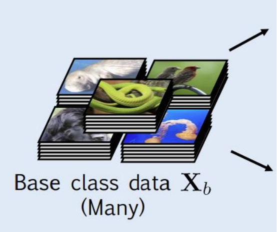
[1] Borealis AI, Tutorial on few-shot learning and meta-learning, link
[2] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Few Shot Classification: Training (with base class data) paradigms

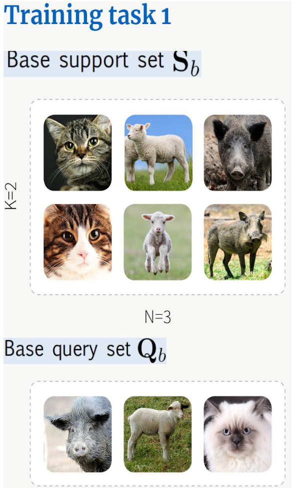
Meta-learning

"Learning to learn": a paradigm specifically for the k-shot n-way task that uses base data to "learn to learn", *i.e.* learn a *meta-learner*, and applies the meta-learner on the testing phase (support + query) data.

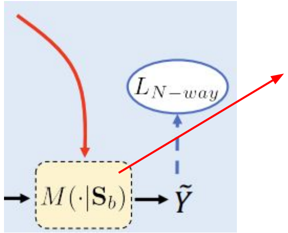
Randomly sample N classes and rearrange base class data into meta-training tasks that simulate test (usually same k, N).



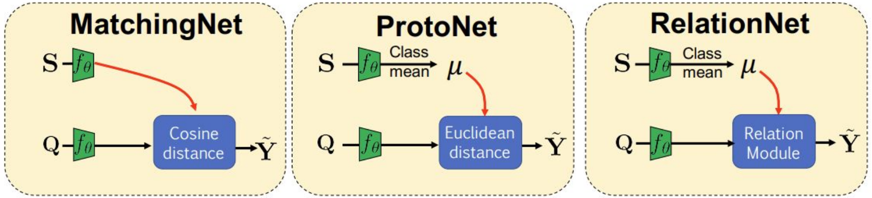
Meta-training stage

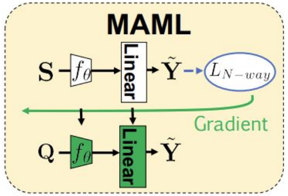


Build a support set conditioned model by min N-way loss



Distance metric based models





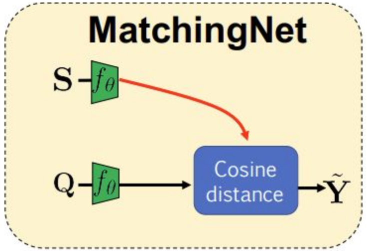
Initialisation based model

Learns a *good model initialization* to accurately classify novel class images with few labelled examples

[1] Borealis AI, Tutorial on few-shot learning and meta-learning, link
[2] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Learning a meta-learner with distance metric: cosine

Meta-learning

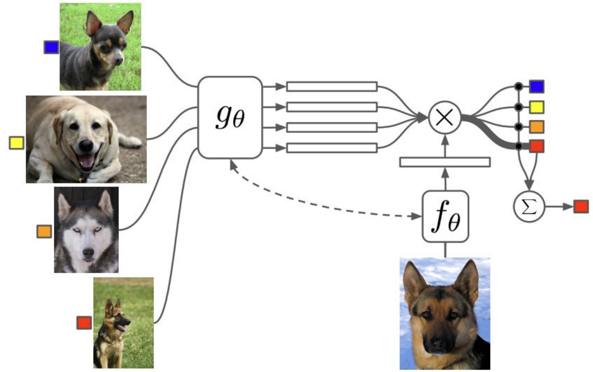


➔ Cosine distance between support features and query features computed

$$\hat{y} = \sum_{i=1}^{k} a(\hat{x}, x_i) y_i$$

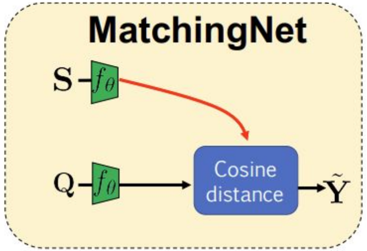
➔ Attention mechanism a(.,.) is chosen as softmax (not shown in image but present in loss) of cosine distance between labelled support samples' features (x_i, y_i) and query features $\hat{x}$

Toy 1-shot 4-way (4 classes of dog breeds) with 1 query example



Learning a meta-learner with distance metric: cosine

Meta-learning



➔ Cosine distance between support features and query features computed

$$\hat{y} = \sum_{i=1}^{k} a(\hat{x}, x_i) y_i$$

➔ Attention mechanism a(.,.) is chosen as softmax (not shown in image but present in loss) of cosine distance between labelled support samples' features (x_i, y_i) and query features $\hat{x}$

➔ Loss:

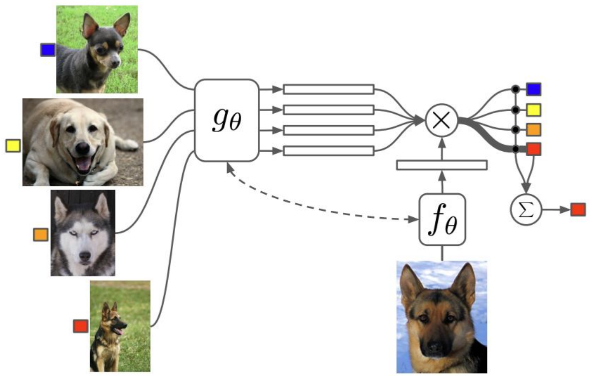
Sample Support and Query from Task

$$L_{N\text{-way}} = \mathbb{E}_{\text{Task}_N \sim \text{ Base}} \left[\mathbb{E}_{S_b \sim \text{Task}_N, Q_b \sim \text{Task}_N} \left[- \sum_{(\hat{x}, \hat{y}) \in Q_b} \log P(\hat{y} | \hat{x}, S_b) \right] \right]$$

Sample a N-way task from Base data

Sum log loss over Query set

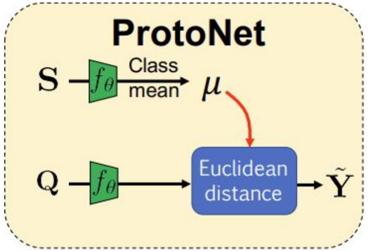
Toy 1-shot 4-way (4 classes of dog breeds) with 1 query example



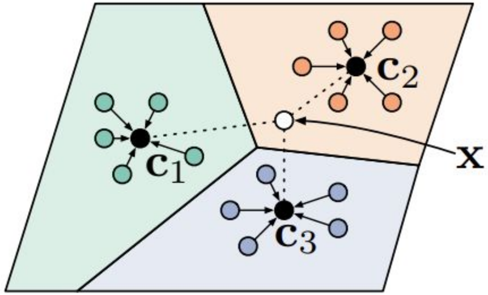
[1] Zsolt Kira, "Low-Label ML Formulations", CS 4803 Course Presentation, link
[2] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019
[3] O. Vinyals et. al., "Matching Networks for One Shot Learning", arxiv:1606.04080v2

Learning a meta-learner with distance metric: euclidean

Meta-learning



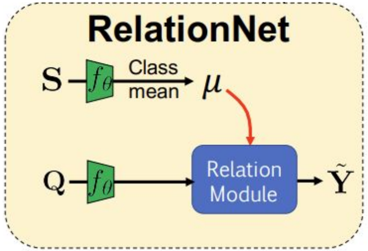
- ➔ An alternative to MatchingNet which uses, as attention mechanism, Softmax of Euclidean distance between query features and class mean of support features.
- ➔ Equivalent to learning an embedding network for a Gaussian classifier to work well
- ➔ In few-shot and zero-shot learning, prototypes are points in the feature space used to represent a single class, and distance to the prototype determines how an observation is classified.



[1] Zsolt Kira, “Low-Label ML Formulations”, CS 4803 Course Presentation, link
[2] J. Snell et. al., “Prototypical Networks for Few-shot Learning”, arxiv:1703.05175v2
[3] Li, Oscar, et al. "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018

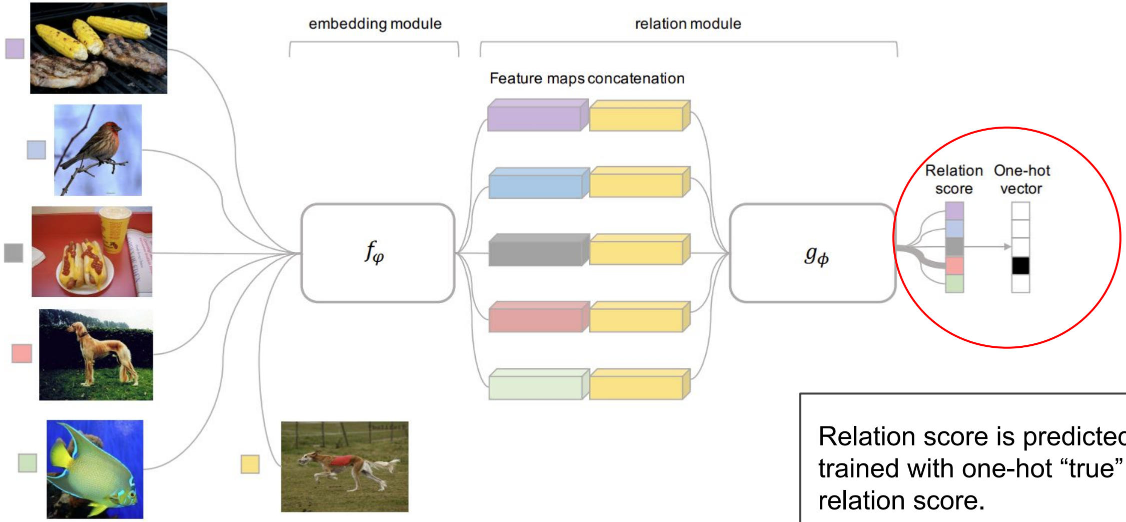
Learning a meta-learner with distance metric *that is learnt*

Meta-learning



→ Another alternative with a learnable attention mechanism.

Toy 1-shot 5-way task with 1 query example

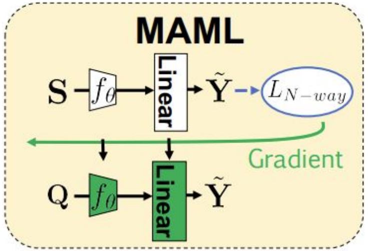


Relation score is predicted and trained with one-hot "true" relation score.

[1] Zsolt Kira, "Low-Label ML Formulations", CS 4803 Course Presentation, link
[2] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019
[3] J. Snell et. al., "Prototypical Networks for Few-shot Learning", arxiv:1703.05175v2

Learning a meta-learner with to best initialise a model

Meta-learning



Algorithm 2 MAML for Few-Shot Supervised Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
randomly initialize θ
while not done **do**
  Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
  for all $\mathcal{T}_i$ **do**
    Sample K datapoints $\mathcal{D} = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\}$ from $\mathcal{T}_i$
    Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
    Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
    Sample datapoints $\mathcal{D}_i' = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\}$ from $\mathcal{T}_i$ for the meta-update
  end for
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$ using each $\mathcal{D}_i'$
end while

Sampling multiple tasks each with images from N randomly chosen classes

Sampling base support sets

Each support set is used to adapt the initial model parameters using few gradient updates.

As different support sets have different gradient updates, the adapted model is conditioned on the support set.

Sampling base query sets

[1] Zsolt Kira, “Low-Label ML Formulations”, CS 4803 Course Presentation, link
[2] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, “A Closer Look At Few-Shot Classification”, Proceedings of the International Conference on Learning Representations (ICLR) 2019
[3] C. Finn, P. Abbeel, S. Levine, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”, Proceedings of the 34th International Conference on Machine Learning (ICML) 2017

The Paper's Experimental Contributions

Adding onto their survey of meta-learning for few-shot classification, the authors present some drawbacks with the aforementioned methods:

(1) Baseline++ transfer learning method demonstrated similar accuracy as meta-learning methods (unexpected in few-shot classification). Reported accuracy is calculated on query/test set on novel-labelled images not seen in meta-training.

Caltech-UCSD Birds (CUB) Dataset consists of images of 6033 images of 200 bird species

Method	CUB 1-shot	CUB 5-shot	*mini*-ImageNet 1-shot	*mini*-ImageNet 5-shot
Baseline	47.12 ± 0.74	64.16 ± 0.71	42.11 ± 0.71	62.53 ±0.69
Baseline++	60.53 ± 0.83	79.34 ± 0.61	48.24 ± 0.75	66.43 ±0.63
MatchingNet Vinyals et al. (2016)	61.16 ± 0.89	72.86 ± 0.70	48.14 ± 0.78	63.48 ±0.66
ProtoNet Snell et al. (2017)	51.31 ± 0.91	70.77 ± 0.69	44.42 ± 0.84	64.24 ±0.72
MAML Finn et al. (2017)	55.92 ± 0.95	72.09 ± 0.76	46.47 ± 0.82	62.71 ±0.71
RelationNet Sung et al. (2018)	62.45 ± 0.98	76.11 ± 0.69	49.31 ± 0.85	66.60 ±0.69

mini-Imagenet has 6000 images of 100 classes of objects (subset of Imagenet)

[1] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

The Paper’s Experimental Contributions

Adding onto their survey of meta-learning for few-shot classification, the authors present some drawbacks with the aforementioned methods:

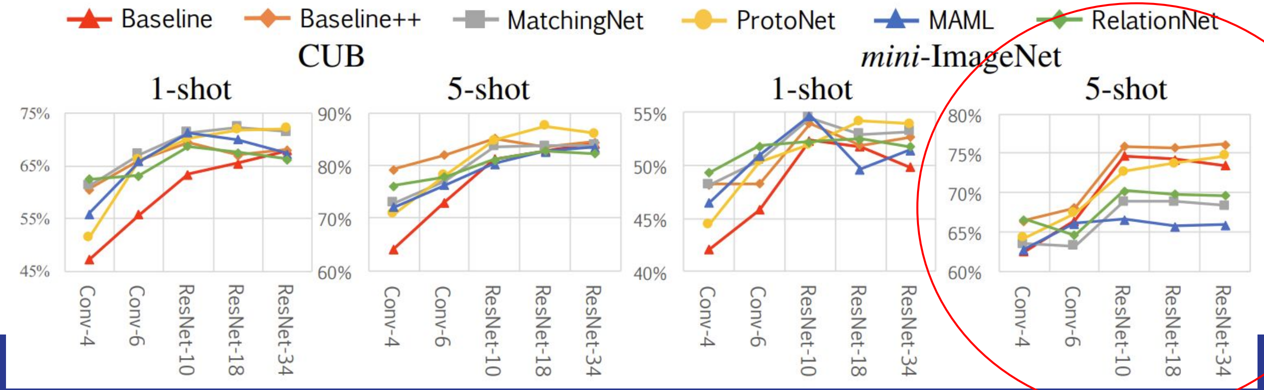
(1) Baseline++ transfer learning method demonstrated similar accuracy as meta-learning methods (unexpected in few-shot classification). Reported accuracy is calculated on query/test set on novel-labelled images not seen in meta-training.

Caltech-UCSD Birds (CUB) Dataset consists of images of 6033 images of 200 bird species

Method	CUB 1-shot	CUB 5-shot	*mini*-ImageNet 1-shot	*mini*-ImageNet 5-shot
Baseline	47.12 ± 0.74	64.16 ± 0.71	42.11 ± 0.71	62.53 ± 0.69
Baseline++	60.53 ± 0.83	79.34 ± 0.61	48.24 ± 0.75	66.43 ± 0.63
MatchingNet Vinyals et al. (2016)	61.16 ± 0.89	72.86 ± 0.70	48.14 ± 0.78	63.48 ± 0.66
ProtoNet Snell et al. (2017)	51.31 ± 0.91	70.77 ± 0.69	44.42 ± 0.84	64.24 ± 0.72
MAML Finn et al. (2017)	55.92 ± 0.95	72.09 ± 0.76	46.47 ± 0.82	62.71 ± 0.71
RelationNet Sung et al. (2018)	62.45 ± 0.98	76.11 ± 0.69	49.31 ± 0.85	66.60 ± 0.69

mini-Imagenet has 6000 images of 100 classes of objects (subset of Imagenet)

(2) Most meta-learning methods are beaten by baselines with deeper backbones in 5-shot on mini-Imagenet

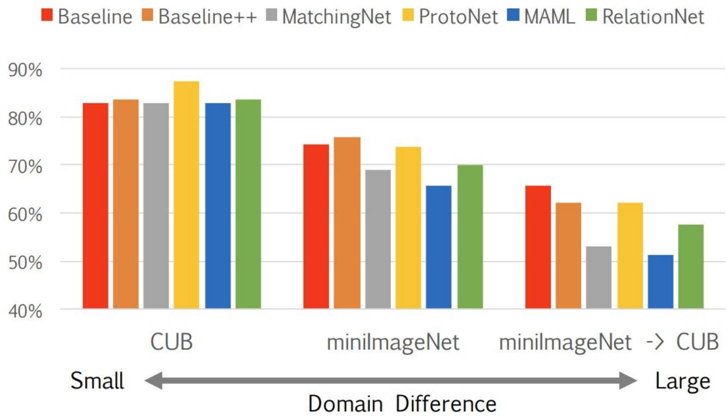


[1] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, “A Closer Look At Few-Shot Classification”, Proceedings of the International Conference on Learning Representations (ICLR) 2019

The Paper's Experimental Contributions

The authors further analyse the effects of backbone depth in a cross-domain situation (train on one dataset, draw novel classes from another)

(3) With larger domain difference (intra-class variation between training classes and novel test classes), baseline methods do better than most meta-learning methods



Caltech-UCSD Birds (CUB) Dataset has least intra-class variation with mostly images of birds in trees.

Largest cross-domain difference when meta-training on one dataset (mini-Imagenet) and meta-testing on other (CUB)

[1] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, "A Closer Look At Few-Shot Classification", Proceedings of the International Conference on Learning Representations (ICLR) 2019

Authors’ Conclusions

- ★ A codebase for comparing meta-learning methods is provided. (Helpful)
- ★ Baseline++ is comparative to SOTA in meta-learning.
- ★ Baseline++ is trained to explicitly reduce intra-class variation and in situations with large intra-class variation, performs better than meta-learning methods than are implicitly expected to perform well even under intra-class variation.
- ★ Lack of robustness to domain differences in meta-learning methods should be further studied (meta-learning should include “Learning to learn to adapt” in the meta-training stage).

[1] W. Chen, Y. Lio, Z. Kira, Y. Wang, J. Huang, “A Closer Look At Few-Shot Classification”, Proceedings of the International Conference on Learning Representations (ICLR) 2019

My Comments

The Pros:

- ❏ Bringing various meta-learning methods into one generalized paradigm where the “model” is abstracted out is helpful in building and critiquing new methods. (to the best of my knowledge, this was the first paper to do so)
- ❏ Empirical demonstration of the comparative performance and robustness of a transfer-learning baseline is indicative of the fact that some premature conclusions about the ineffectiveness of transfer learning in few-shot tasks may have been made.

The Cons:

- ❏ ProtoNet outperforms other meta-learning methods and doesn’t have any of the previous 3 drawbacks. This is overlooked in the paper. Prototypical learning’s robustness to intra-class variation can be looked at in more detail.
- ❏ The authors briefly mention “hallucination based models” (that learn how to augment training data) in the related work section but never bring it up again. They also skip other effective “initialisation-based methods”.

And more:

- ❏ Applications to NLP?
- ❏ Does meta-learning make sense?
    - ❏ Supervision issues: Does the domain distribution of the meta-learner contain regions of novel classes?
    - ❏ If not, does it make sense to apply to examples drawn from OOD (out of distribution)? [Authors indirectly arrive at this point by talking about modifying meta-learning to explicitly learn from OOD regions]
    - ❏ Other issues?