Meta-learning with differentiable closed-form solvers

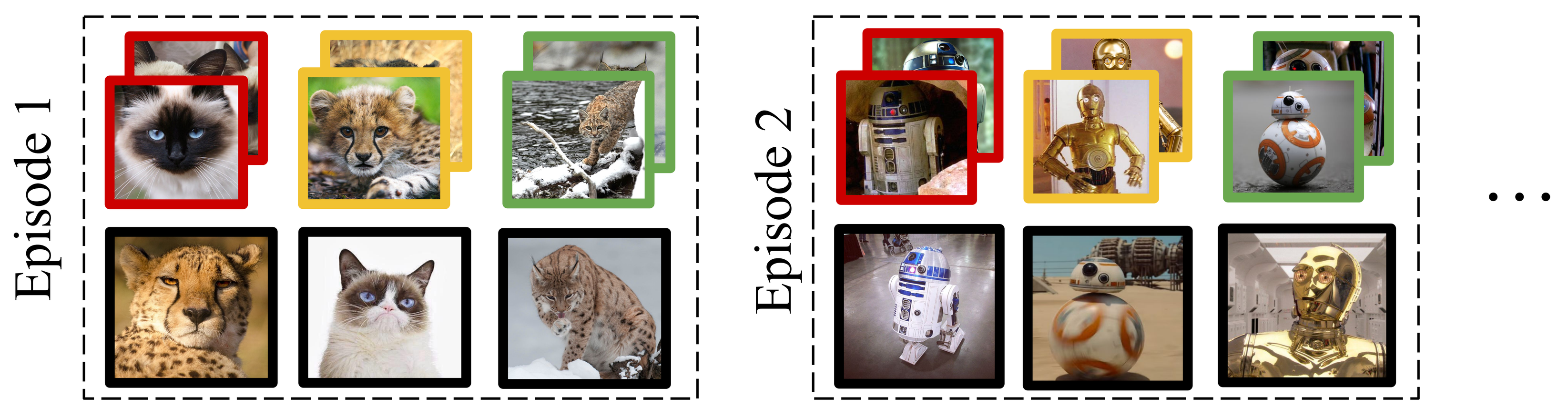
Luca Bertinetto[1,2] João F. Henriques[1] Philip H.S. Torr[1,2] Andrea Vedaldi[1]

[1]University of Oxford [2]FiveAI Ltd.

Code available at github.com/bertinetto/r2d2

One/few-shot learning

Learning to discriminate between previously unseen classes using only a handful of training examples from these classes.



- Episode: small subset sampled from {train,validation,test} split; it is in its turn divided into *base-train* and *base-test*.
- $I_\star$ and $C_\star$: set of images and classes from a data split $\star$.
  In standard classification, $I_{\text{train}} \cap I_{\text{test}} = \varnothing$ and $C_{\text{train}} = C_{\text{test}}$.
  Few-shot learning requires $I_{\text{train}} \cap I_{\text{test}} = \varnothing$ and $C_{\text{train}} \cap C_{\text{test}} = \varnothing$.
- Datasets: Omniglot, *mini*ImageNet, CIFAR-FS.
  Also: Visual Decathlon[1] and Meta-dataset[2] span multiple domains.
- Often tackled with meta-learning.

Meta-learning

- Thrun&Pratt[3] (inspired by Mitchell[4]): when an "algorithm's performance on new tasks improves with experience and with the number of tasks" (by dynamically adapting its inductive bias).
- Modern use (e.g. Ravi&Larochelle[5]): training is conducted at two (nested) levels.
  - Former operates within the scope of individual episodes (i.e. new learning tasks).
  - Latter guides the former and tries to improve it across episodes.

Related work and motivation

- Metric learning-based (e.g. matching[7]/proto[8] networks): simple and fast, but no adaptation to new episodes.
- Iterative (e.g. MAML[6]): adaptation of all parameters in new episodes, but quite slow.
- Our aim: allow fast adaptation to new episodes.
  Intuition: backpropagate through the solution of an efficient learning problem like ridge regression.

General framework

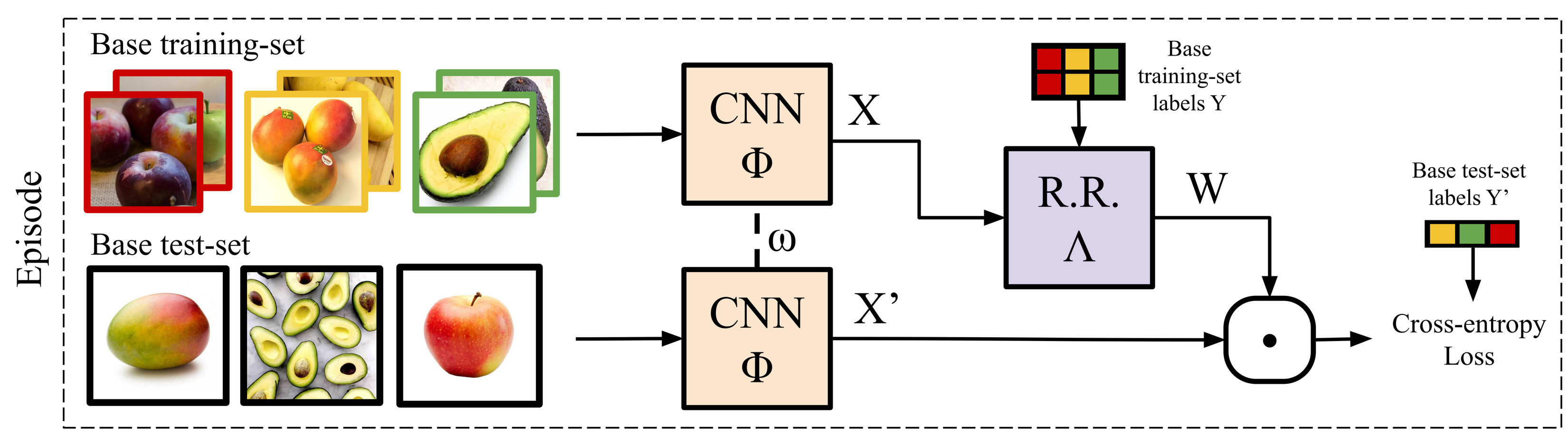
To train+evaluate the predictor on one episode, we use training samples $Z_\mathcal{E} = \{(x_i, y_i)\} \sim \mathcal{E}$ and test samples $Z'_\mathcal{E} = \{(x'_i, y'_i)\} \sim \mathcal{E}$.

$$\min_{\omega,\rho} \frac{1}{|\mathbb{E}| \cdot |Z'_\mathcal{E}|} \sum_{\mathcal{E}\in\mathbb{E}} \sum_{(x',y')\in Z'_\mathcal{E}} L\left(f\left(\phi\left(x'\,;\,\omega\right)\,;\,W\right),\,y'\right),$$

$$\text{with } W = \Lambda(\phi(Z_\mathcal{E}\,;\,\omega)\,;\,\rho)$$

Base learner Λ can be implemented in many ways; we experiment with ridge regression and logistic regression.

R2-D2: ridge regression differentiable discriminator

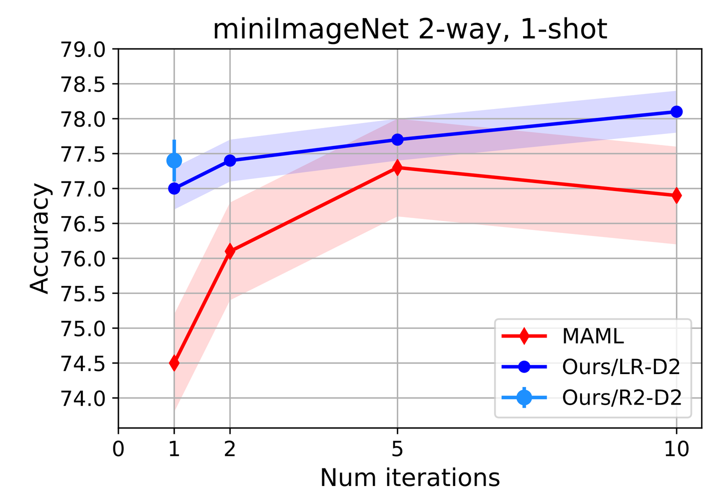


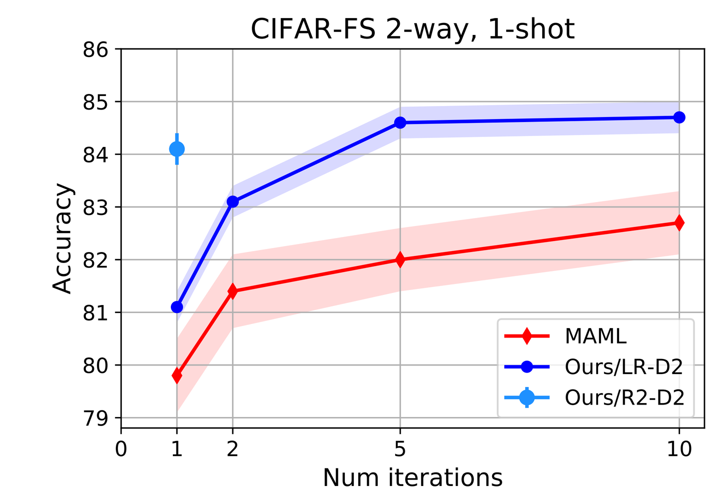
$$\begin{aligned}\Lambda(Z) &= \arg\min_W \|XW - Y\|^2 + \lambda \|W\|^2 \\ &= (X^TX + \lambda I_{e,e})^{-1} X^TY \\ &= X^T(XX^T + \lambda I_{n,n})^{-1} Y \quad \text{(Woodbury identity)}\end{aligned}$$

The Woodbury identity makes the matrix to invert quadratic in n (num examples, typically 1 or 5) rather than in e (embedding size, typically 100-1000): big computational gain in few-shot learning scenario.

LR-D2: logistic regression differentiable discriminator

A similar derivation is also possible for iterative solvers with differentiable operations. In particular, we experiment with Newton's method applied to logistic regression (aka Iteratively Reweighted Least Squares).





Results on *mini*ImageNet and CIFAR-FS

	*mini*ImageNet, 5-way		CIFAR-FS, 5-way	
Method	1-shot	5-shot	1-shot	5-shot
Matching net	44.2%	57%	—	—
MAML	48.7±1.8%	63.1±0.9%	58.9±1.9%	71.5±1.0%
MAML $*$	40.9±1.5%	58.9±0.9%	53.8±1.8%	67.6±1.0%
Meta-LSTM	43.4±0.8%	60.6±0.7%	—	—
Proto net	47.4±0.6%	65.4±0.5%	55.5±0.7%	72.0±0.6%
Proto net $*$	42.9±0.6%	65.9±0.6%	57.9±0.8%	76.7±0.6%
Relation net	50.4±0.8%	65.3±0.7%	55.0±1.0%	69.3±0.8%
SNAIL (with *ResNet*)	*55.7±1.0%*	*68.9±0.9%*	—	—
SNAIL (with 32C)	45.1%	55.2%	—	—
GNN	50.3%	66.4%	61.9%	75.3%
GNN$*$	50.3%	**68.2**%	56.0%	72.5%
Ours/R2-D2 (with 64C)	49.5±0.2%	65.4±0.2%	**62.3**±0.2%	**77.4**±0.2%
Ours/R2-D2	**51.8**±0.2%	**68.4**±0.2%	**65.4**±0.2%	**79.4**±0.2%
Ours/LR-D2 (1 iter.)	51.0±0.2%	65.6±0.2%	**64.5**±0.2%	75.8±0.2%
Ours/LR-D2 (5 iter.)	**51.9**±0.2%	**68.7**±0.2%	**65.3**±0.2%	**78.3**±0.2%

Results on Omniglot

	Omniglot, 5-way		Omniglot, 20-way	
Method	1-shot	5-shot	1-shot	5-shot
Siamese net	96.7%	98.4%	88%	96.5%
Matching net	98.1%	98.9%	93.8%	98.5%
MAML	**98.7**±0.4%	**99.9**±0.1%	95.8±0.3%	98.9±0.2%
Proto net	98.5±0.2%	99.5±0.1%	95.3±0.2%	98.7±0.1%
SNAIL	**99.07**±0.16%	**99.77**±0.09%	**97.64**±0.30%	**99.36**±0.18%
GNN	**99.2%**	**99.7%**	**97.4%**	99.0%
Ours/R2-D2 (with 64C)	98.55±0.05%	99.66±0.02%	94.70±0.05%	98.91±0.02%
Ours/R2-D2	**98.91**±0.05%	**99.74**±0.02%	96.24±0.05%	**99.20**±0.02%

Vanilla transfer learning

Loss in (absolute) accuracy for not considering base learner Λ during training.

	R2-D2
*mini*ImageNet (1-shot)	-13.8%
*mini*ImageNet (5-shot)	-11.6%
CIFAR-FS (1-shot)	-11.5%
CIFAR-FS (5-shot)	-5.9%

Speed

Time required to solve 10,000 episodes.

	5-way/1-shot
Ours/R2-D2	1'23"
Ours/R2-D2 (64C)	1'4"
MAML (32C)	6'35"
Ours/LR-D2 (32C)	5'48"
Ours/R2-D2 (32C)	57"
Proto nets (32C)	24"

References

[1] S.Rebuffi *et al.* Learning multiple visual domains with residual adapters. In NeurIPS'17.
[2] E.Triantafillou *et al.* Meta-dataset: A dataset of datasets for learning to learn from few examples. In *arXiv:1903.03096*, 2019.
[3] S.Thrun and L.Pratt. *Learning to learn*, 1998.
[4] T. Mitchell. *Machine Learning*, 1997.
[5] S.Ravi and H.Larochelle. Optimization as a model for few-shot learning. In ICLR'17.
[6] C.Finn *et al.* Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In ICML'17.
[7] O.Vinyals *et al.* Matching networks. In NeurIPS'16.
[8] J.Snell *et al.* Prototypical Networks for Few-shot Learning. In NeurIPS'17.

Personal page: www.robots.ox.ac.uk/~luca

Get in touch! luca@robots.ox.ac.uk