A Tutorial on Few-Shot Learning

ICCV 2019 Tutorial on Learning with Limited Data

Kevin Swersky, Google Research

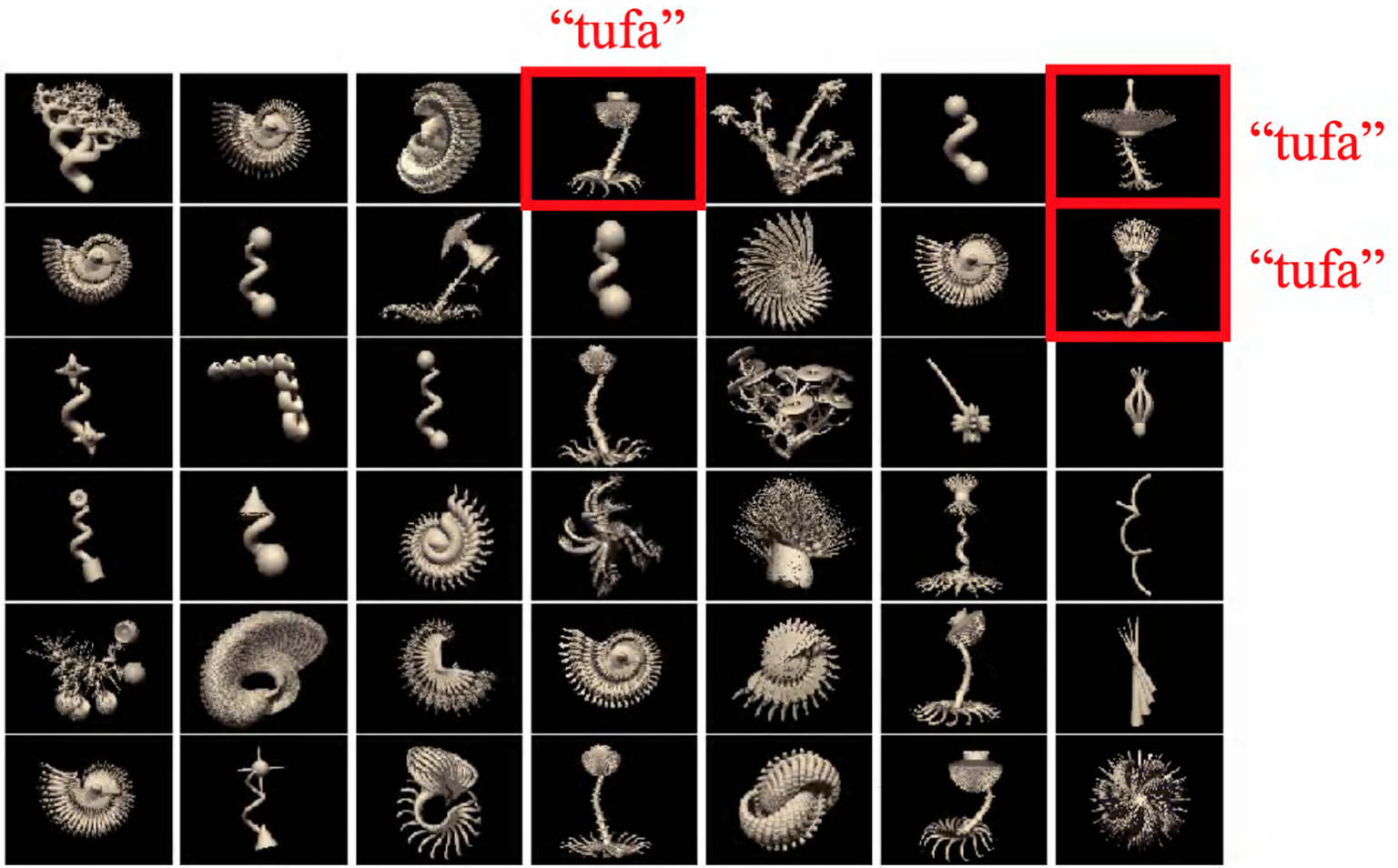
Nov 2, 2019

Special thanks to Hugo Larochelle, Pascal Lamblin, and Vincent Dumoulin for important contributions to this talk

Outline

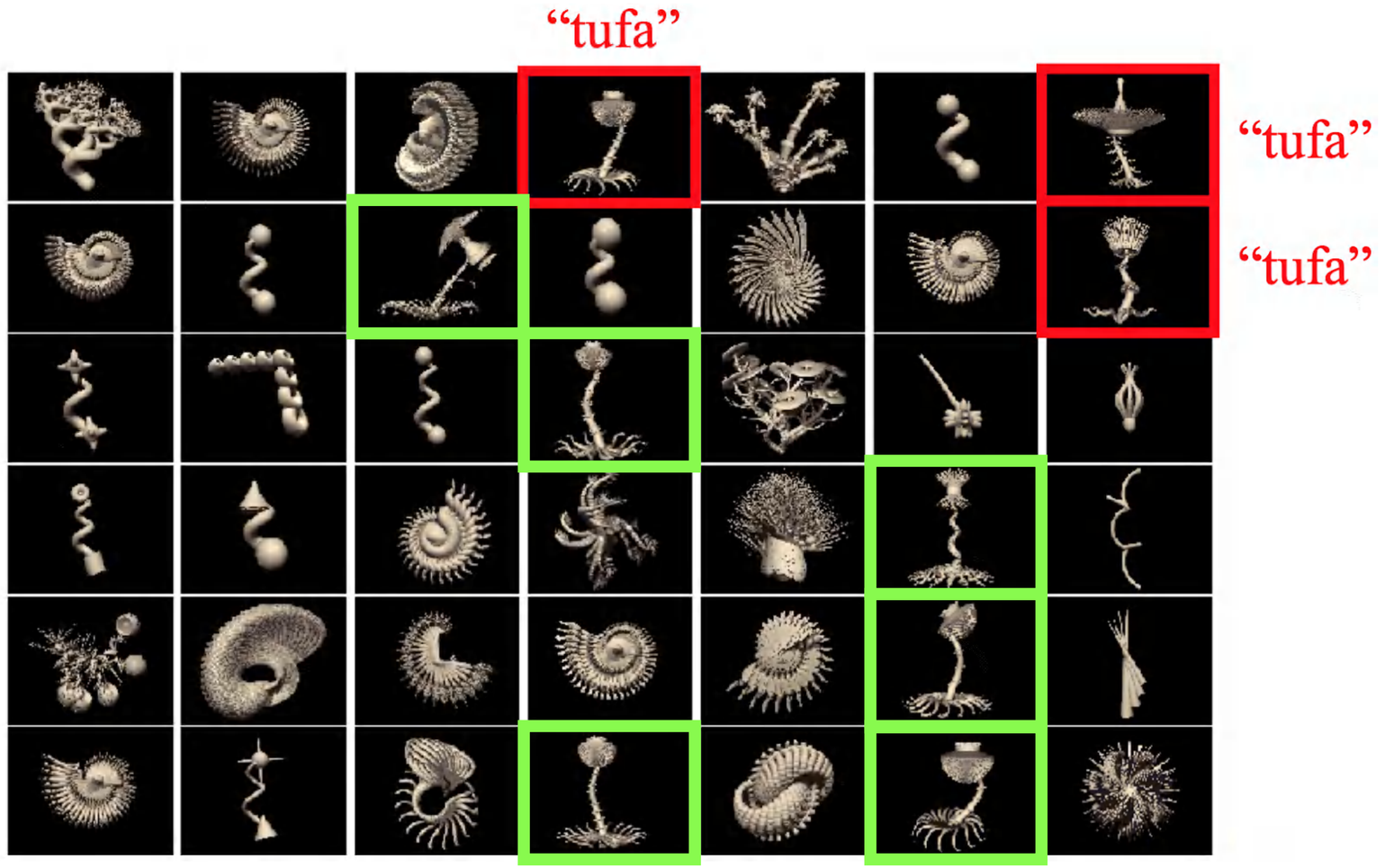
- Problem setup
- Related work
- Batch vs episodic learning
- Recent approaches
- Extensions and variations

Human Few-Shot Learning



Source: Josh Tenenbaum

Human Few-Shot Learning



Source: Josh Tenenbaum

Human Few-Shot Learning

Human-level concept learning through probabilistic program induction
Brenden M. Lake, Ruslan Salakhutdinov, Joshua B. Tenenbaum

Human Few-Shot Learning

Human-level concept learning through probabilistic program induction
Brenden M. Lake, Ruslan Salakhutdinov, Joshua B. Tenenbaum

Deep learning is data hungry

- Key issue: training on limited data leads to overfitting
- We can exploit other sources of information to build robust learners
  - Unsupervised learning
  - Transfer learning
  - Data augmentation (best results require good priors)
- Few-shot learning uses auxiliary data to build useful representations and to learn good update rules for the limited data setting

Related work

Related work: transfer learning

- Large image datasets (e.g. ImageNet) have been shown to allow training representations that transfer to other problems
  - DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition (2014)
    Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng and Trevor Darrell
  - *CNN Features off-the-shelf: an Astounding Baseline for Recognition (2014)*
    Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson
- In few-shot learning, we aim at transferring the complete training of the model on new datasets (not just transferring the features or initialization)
  - ideally there should be no human involved in producing a model for new datasets

Related work: one-shot learning

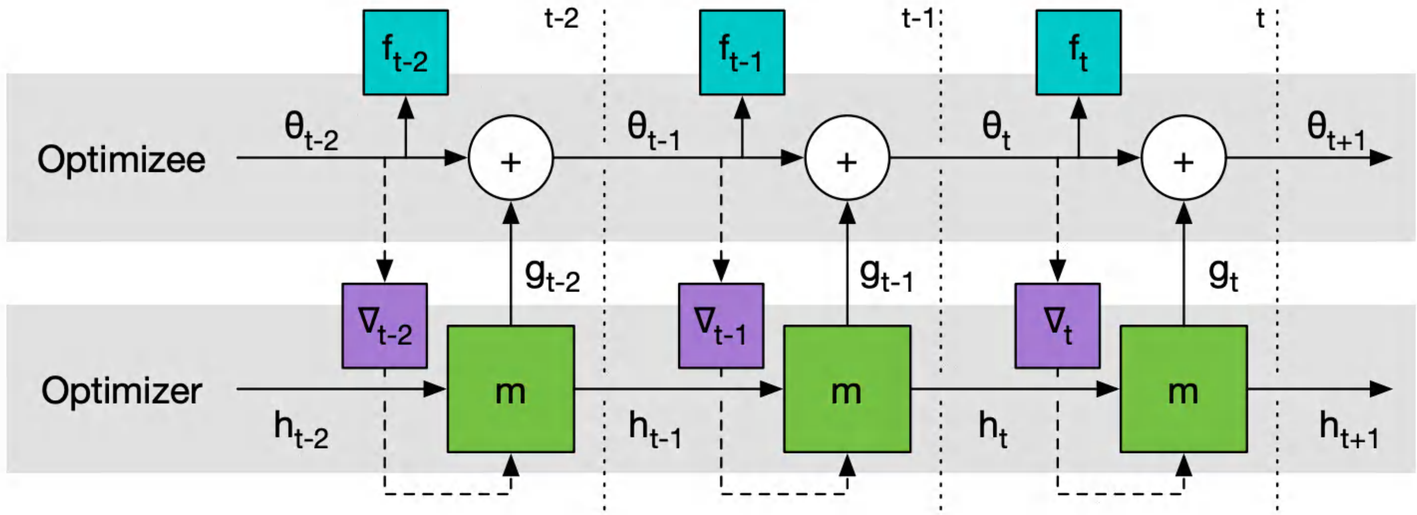
- One-shot learning has been studied before
  - One-Shot learning of object categories (2006)
    Fei-Fei Li, Rob Fergus and Pietro Perona
  - Knowledge transfer in learning to recognize visual objects classes (2004)
    Fei-Fei Li
  - Object classification from a single example utilizing class relevance pseudo-metrics (2004)
    Michael Fink
  - Cross-generalization: learning novel classes from a single example by feature replacement (2005)
    Evgeniy Bart and Shimon Ullman
- These largely relied on hand-engineered features
  - with recent progress in end-to-end deep learning, we hope to learn a representation better suited for few-shot learning

Related work: meta-learning

- Early work on learning an update rule
  - Learning a synaptic learning rule (1990)
    Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier
  - The Evolution of Learning: An Experiment in Genetic Connectionism (1990)
    David Chalmers
  - On the search for new learning rules for ANNs (1995)
    Samy Bengio, Yoshua Bengio, and Jocelyn Cloutier
- Early work on recurrent networks modifying their weights
  - Learning to control fast-weight memories: An alternative to dynamic recurrent networks (1992)
    Jürgen Schmidhuber
  - A neural network that embeds its own meta-levels (1993)
    Jürgen Schmidhuber

Related work: meta-learning

- Training a recurrent neural network to optimize
  - Outputs update, so can decide to do something else than gradient descent

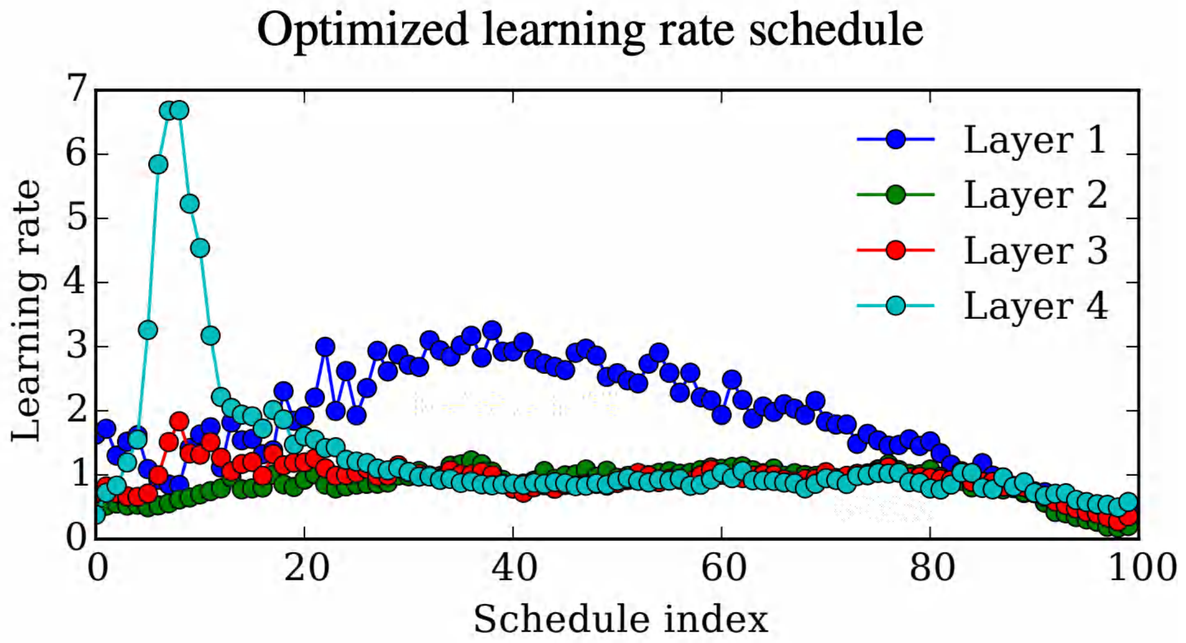


- Learning to learn by gradient descent by gradient descent (2016)
  Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W. Hoffman, David Pfau, Tom Schaul, and Nando de Freitas

- Learning to learn using gradient descent (2001)
  Sepp Hochreiter, A. Steven Younger, and Peter R. Conwell

Related work: meta-learning

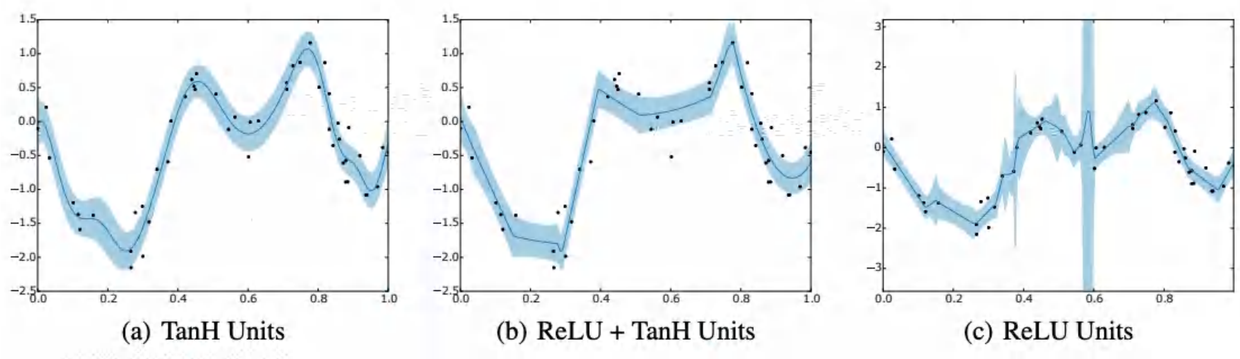
- Hyper-parameter optimization
  - Idea of learning the learning rates and initialization conditions

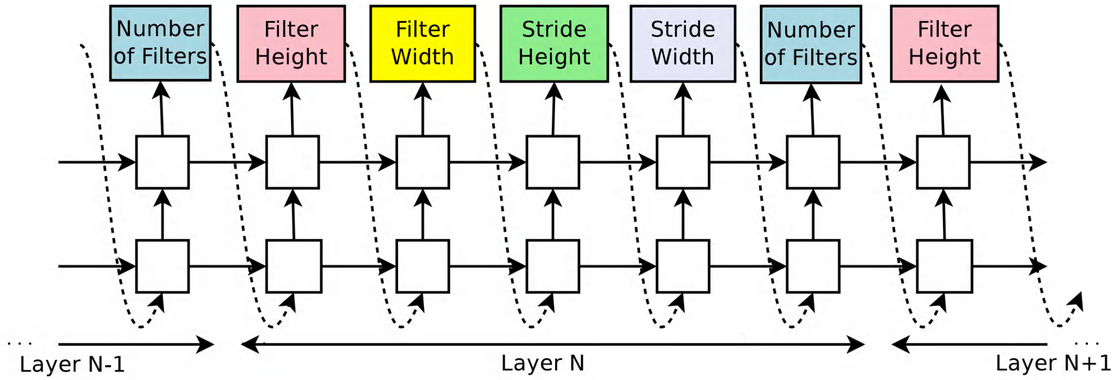


- Gradient-based hyperparameter optimization through reversible learning (2015)
  Dougal Maclaurin, David Duvenaud, Ryan P. Adams

Related work: meta-learning

- AutoML





- Neural architecture search with reinforcement learning (2017)
  Barret Zoph and Quoc Le

- Scalable Bayesian optimization using deep neural networks (2015)
  Jasper Snoek, Oren Rippel, Kevin Swersky, Jamie Kiros, Nadathur Satish, Narayanan Sundaram, Md. Mostofa Ali Patwary, Prabhat, Ryan P. Adams

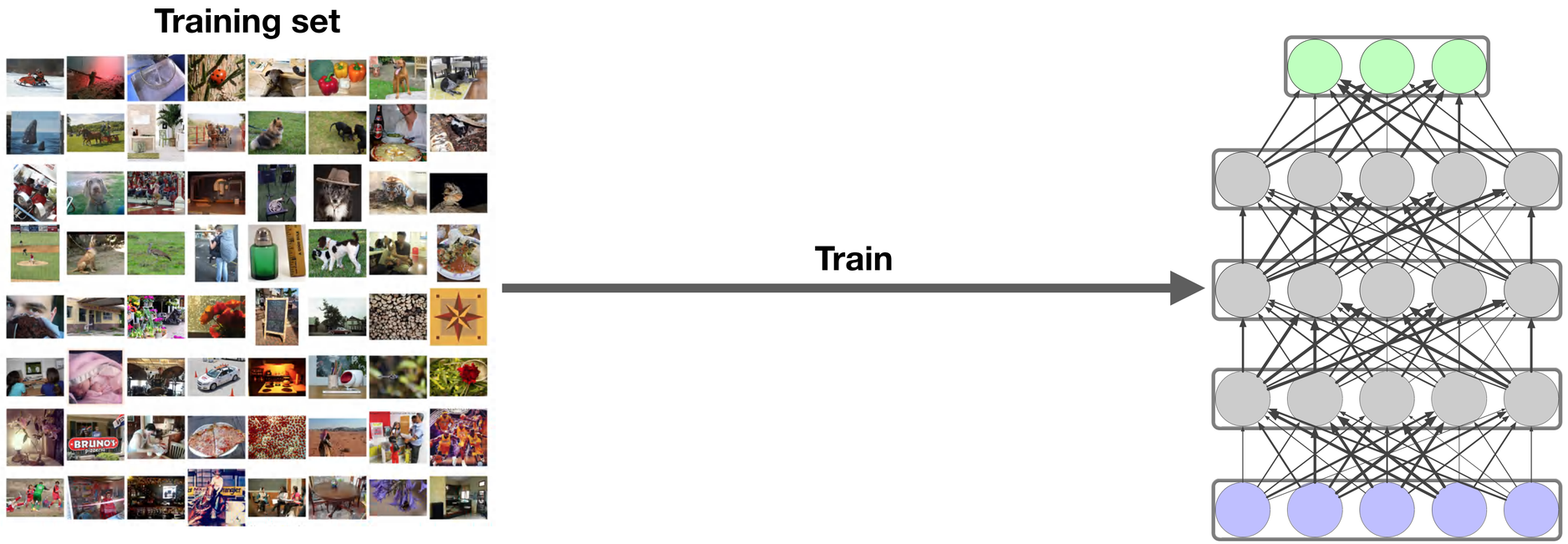
Few-shot classification setup

- Given several examples of several classes, create a classifier to classify unseen instances.

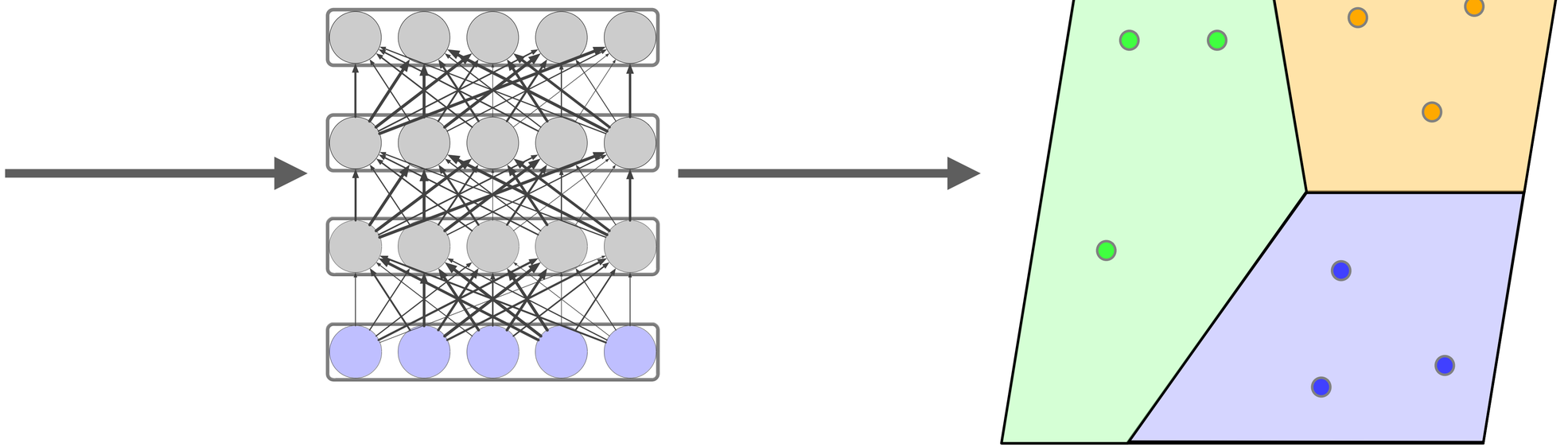
Simple strategies (batch training)

Learn good features and use simple classifiers to avoid overfitting

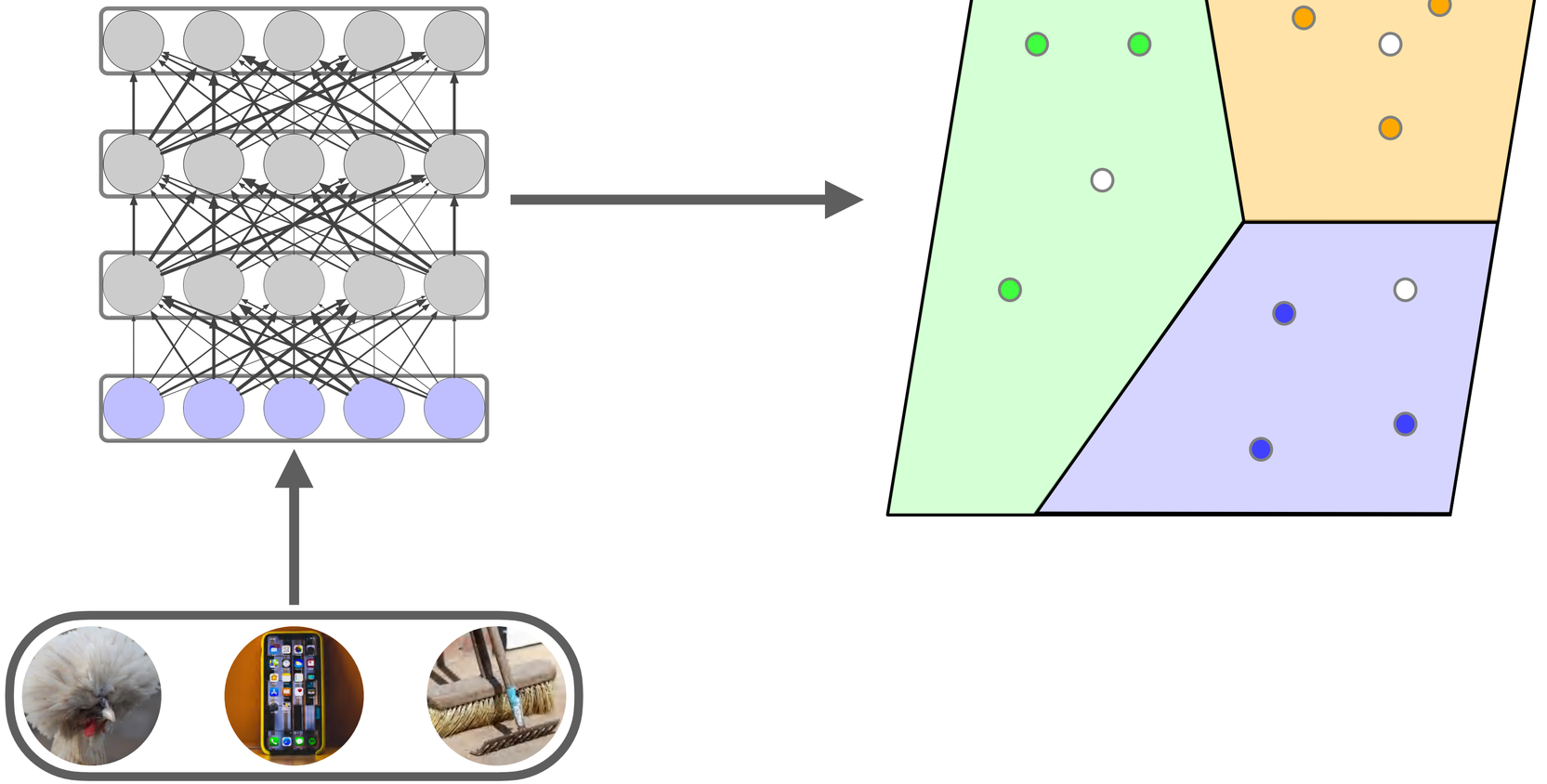
Simple strategies



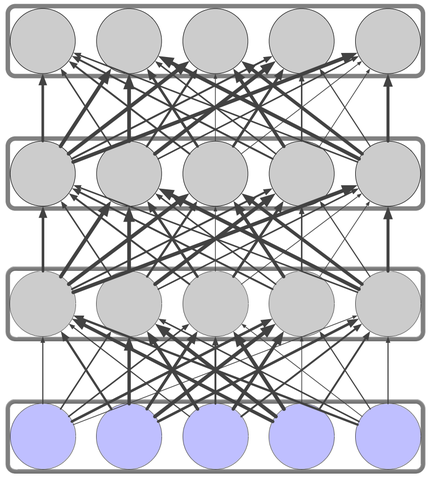
Simple strategy 1 (knn)

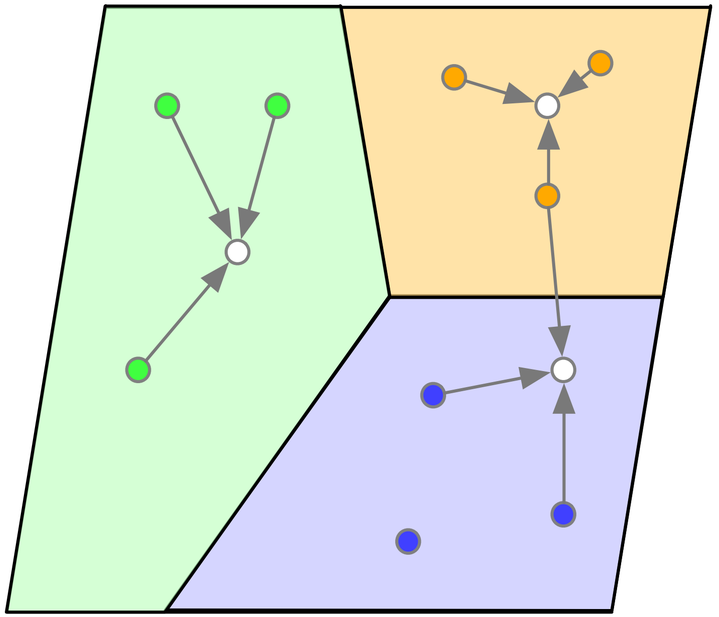


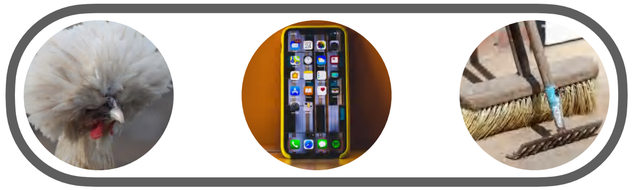
Simple strategy 1 (knn)



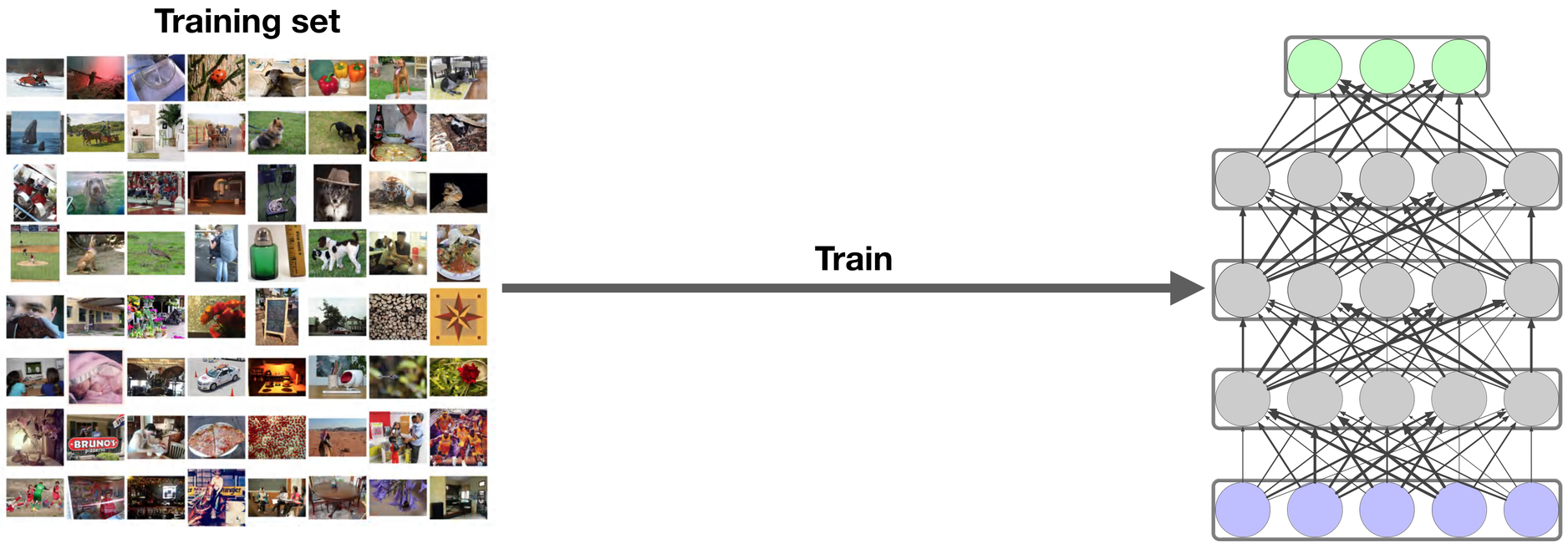
Simple strategy 1 (knn)





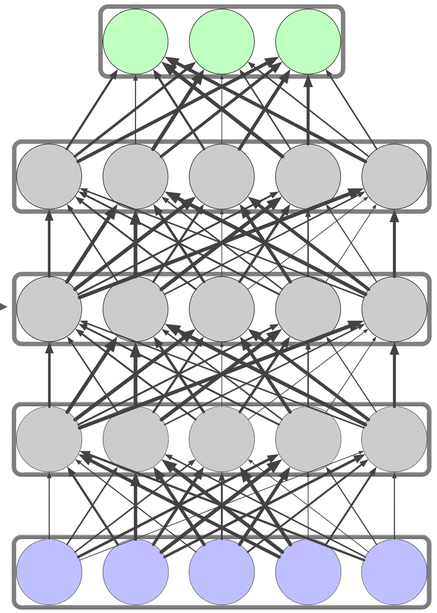


Simple strategies



Simple strategy 2 (fine-tune)

Fine-tune with a few steps of gradient descent



- A Closer Look at Few-shot Classification
  Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, Jia-Bin Huang

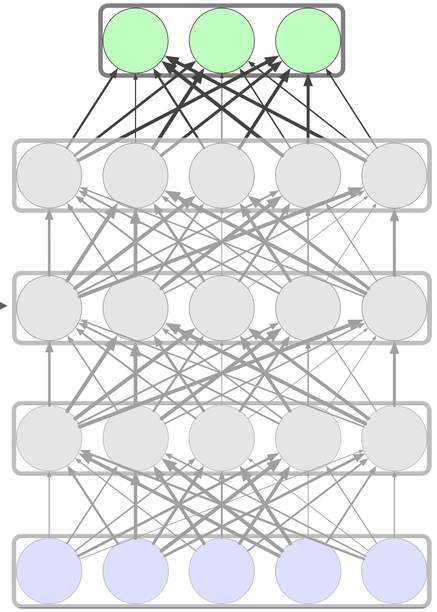
Simple strategy 2 (fine-tune)

Hen

Broom

Cell phone

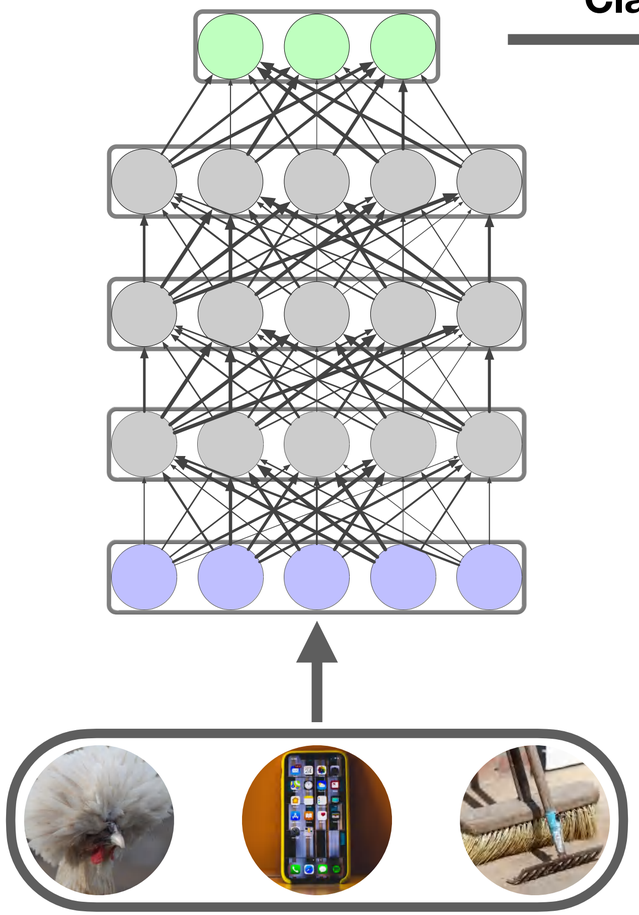
Fine-tune with a few steps of gradient descent



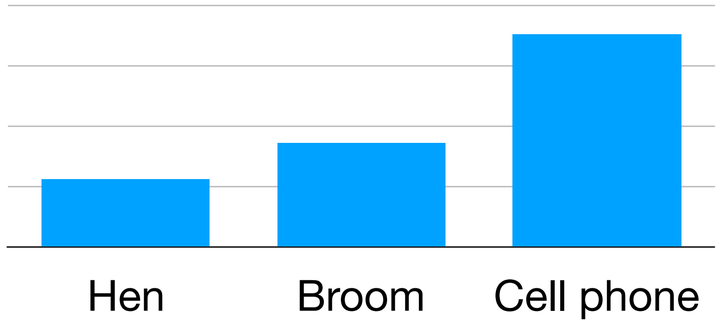
Optionally freeze lower layer weights

- A Closer Look at Few-shot Classification
  Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, Jia-Bin Huang

Simple strategy 2 (fine-tune)



Classify



- A Closer Look at Few-shot Classification
  Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, Jia-Bin Huang

Episodic training

Match the training environment to the testing environment and learn to avoid overfitting

Notation

- An **episode** is a self-contained batch of data representing a new classification problem. It can be thought of as a **task**.
- Most few-shot learners have two components
  - **Embedding model** maps inputs into feature space
  - **Base learner** creates a classifier from features
  - We'll discuss several examples of base learners
- **Meta-training**: train the embedding model and base learner across many few-shot tasks in order to generalize to held-out few-shot tasks
- **Meta-testing**: testing the model and learner on held-out few-shot tasks

Notation

- N classes with k examples per class (typically $1 \leq k \leq 5$)
  - N is referred to as the “way” e.g., 20-way
  - k is referred to as the “shot”
- Total number of images per episode is $K = kN$
- Input vectors $\mathbf{x} \in \mathbb{R}^D$, labels $y \in \{1, \ldots, N\}$
- Support set $\mathcal{S} = \left\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_K, y_K)\right\}$
  - Used to “train” the base-learner in an episode
- Query set $\mathcal{Q} = \left\{(\mathbf{x}_1^*, y_1^*), (\mathbf{x}_2^*, y_2^*), \ldots (\mathbf{x}_T^*, y_T^*)\right\}$
  - Used to “test” the base-learner in an episode
- We’ll assume access to a set of training classes $\mathscr{C}_{\text{train}}$ for meta-training
- We’ll test on a held-out set of classes $\mathscr{C}_{\text{test}}$ for meta-testing

Episodic Training

Hen

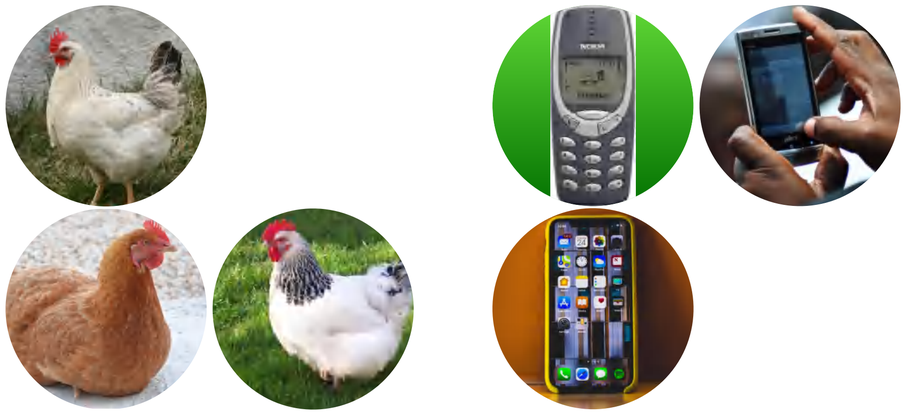
Broom

Cell phone

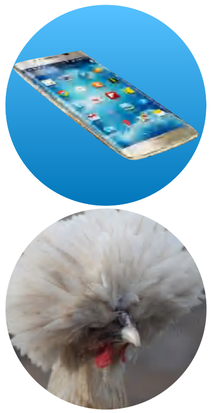
Guacamole

Dog

Support set



Query set



Episodic Training

Hen

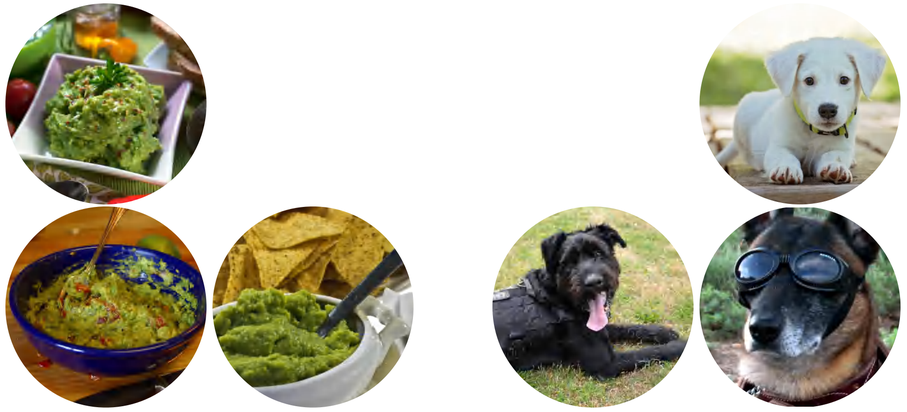
Broom

Cell phone

Guacamole

Dog

Support set



Query set

Episodic Training

Hen

Broom

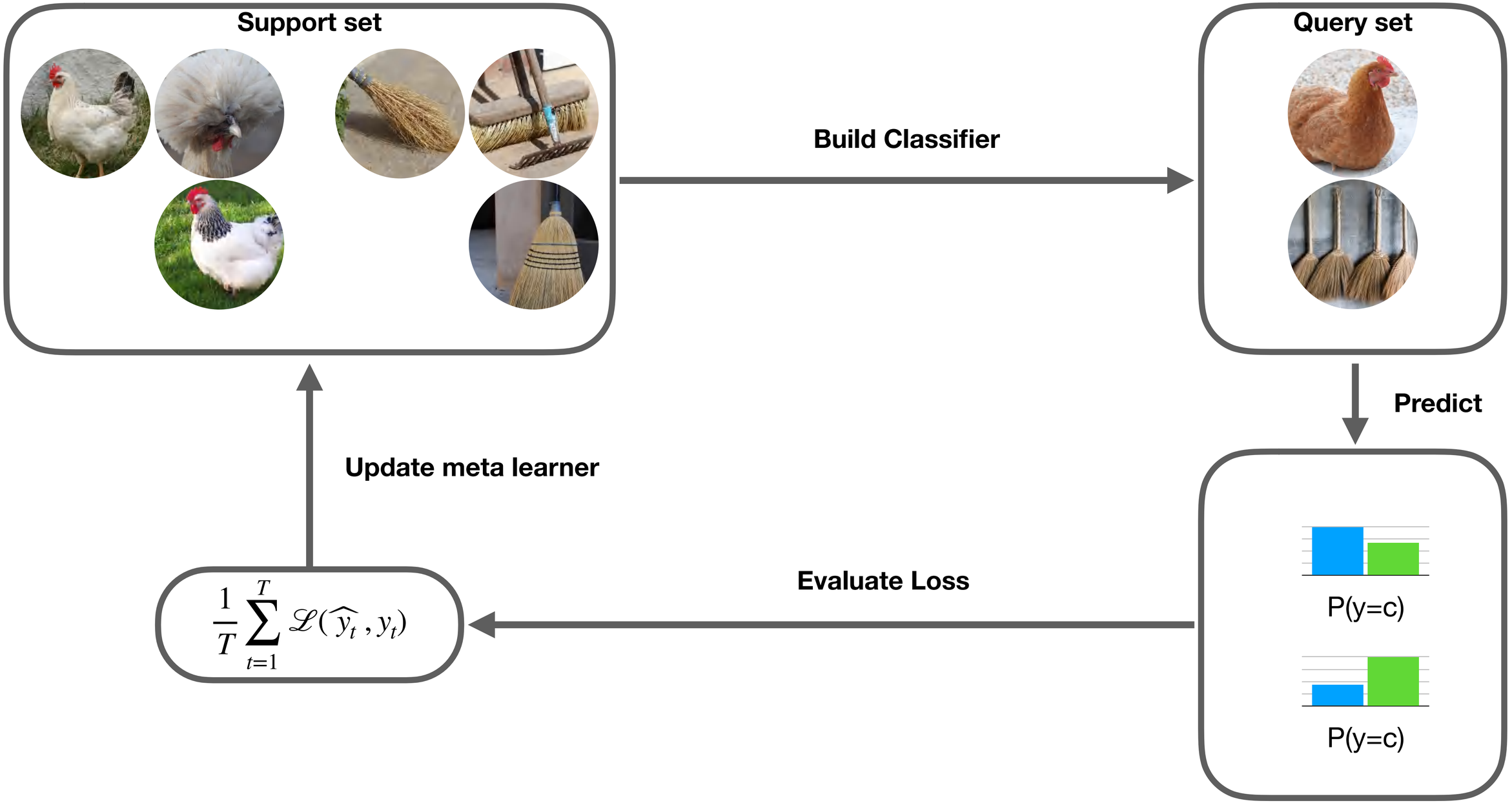
Cell phone

Guacamole

Dog

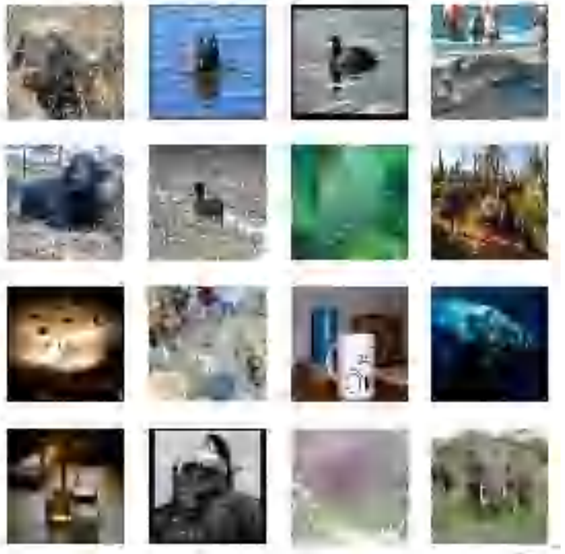
Support set

Query set


Support set
Query set
Build Classifier
Predict
Update meta learner
Evaluate Loss
$\frac{1}{T}\sum_{t=1}^{T}\mathscr{L}(\widehat{y_t}, y_t)$
P(y=c)
P(y=c)

Benchmark datasets

Mini-Imagenet (Vinyals et al., 2016)



- 100 classes from Imagenet
- 64 training classes
- 16 validation classes
- 20 test classes
- 600 examples per class

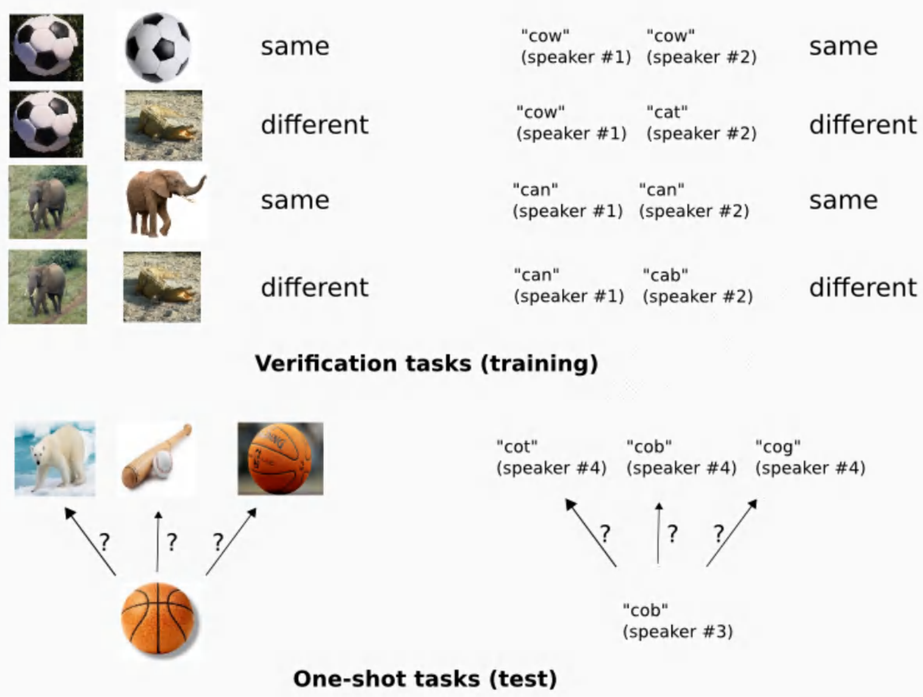
Omniglot (Lake et al., 2015)

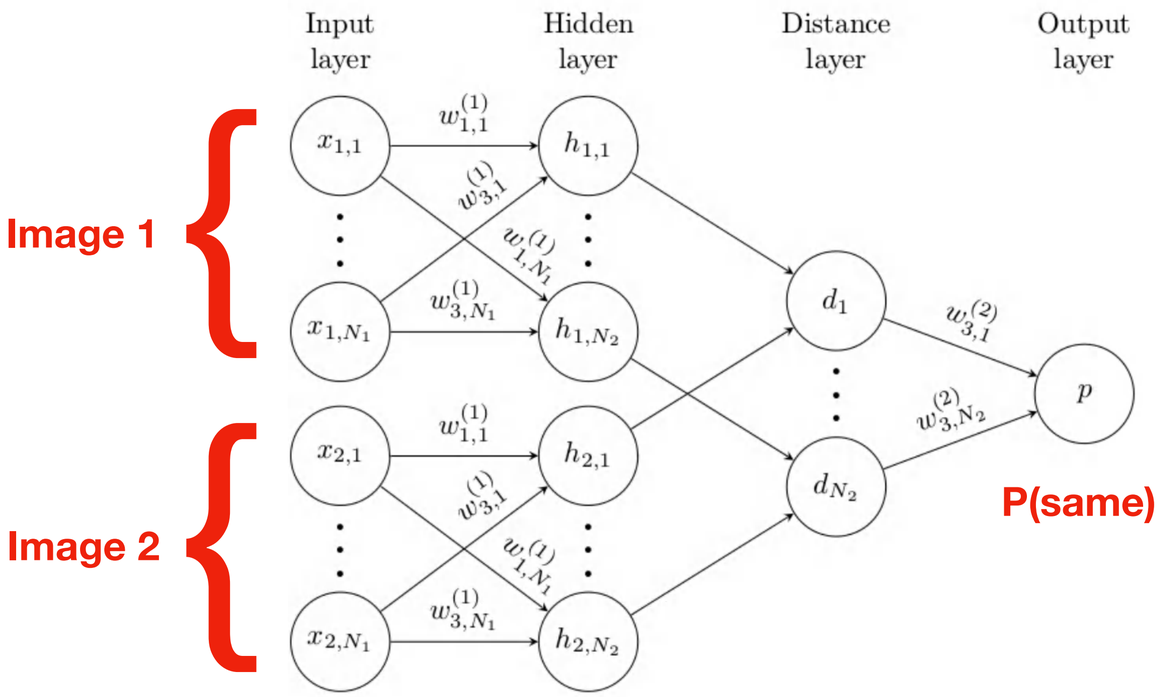
- 1623 handwritten characters
- 50 alphabets
- 20 examples per character

A sample of recent few-shot learning approaches

Siamese Networks

Train a network to recognize instances of the same class

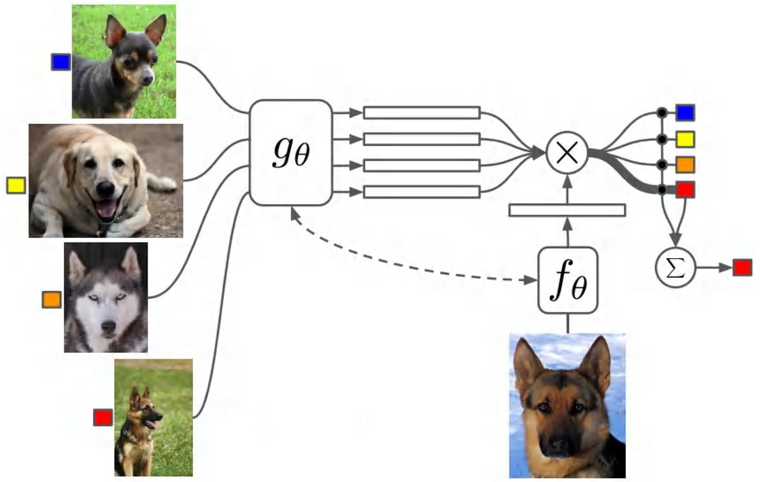


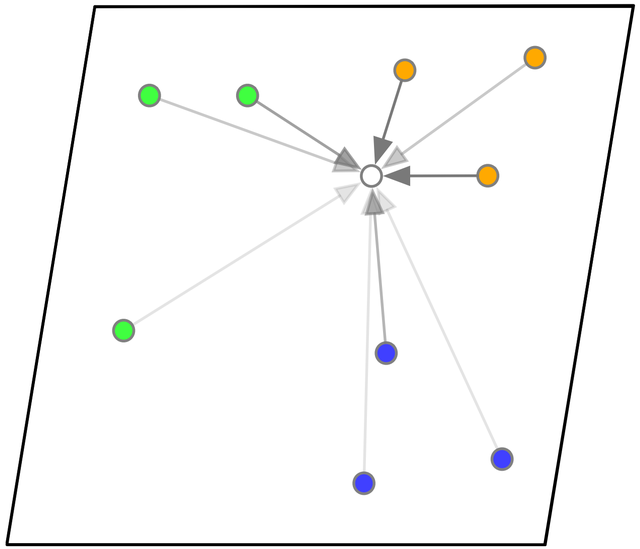


- Siamese neural networks for one-shot image recognition (2015)
  Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov

Matching Networks

Learn an embedding where distance-weighted nearest neighbour classification works well





$$P(\hat{y}_k = 1 \mid \hat{x}, \mathcal{S}) = \sum_{(x,y)\in\mathcal{C}_k} a(\hat{x}, x)y$$

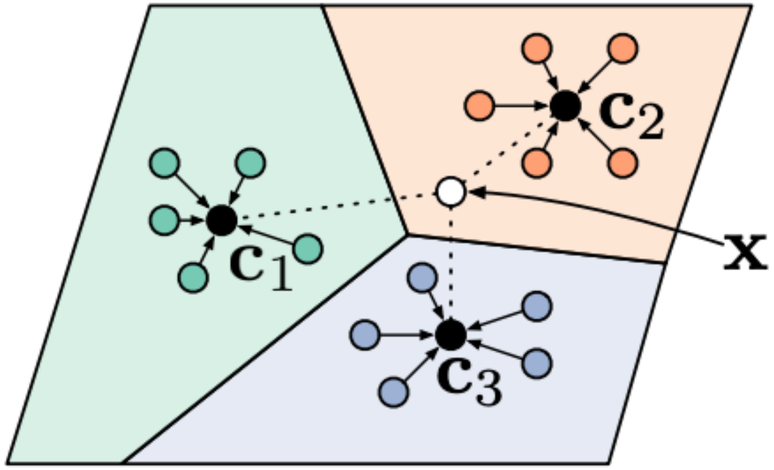
$$a(\hat{x}, x) = \frac{\exp\left(c(f(\hat{x}), g(x))\right)}{\sum_{i=1}^{K} \exp\left(c(f(\hat{x}), g(x_i))\right)}$$

$$c(u, v) = u^\top v$$

- Matching networks for one-shot learning (2015)
  Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra

Prototypical Networks

Map class embeddings to prototypes using their average and do nearest-prototype classification



$$P(\hat{y}_k = 1 \mid \hat{x}, \mathcal{S}) = \frac{\exp\left(c(f(\hat{x}), p_k)\right)}{\sum_{i=1}^{N} \exp\left(c(\hat{x}, p_i)\right)}$$

$$p_k = \frac{1}{|\mathcal{S}_k|} \sum_{x \in \mathcal{S}_k} f(x)$$

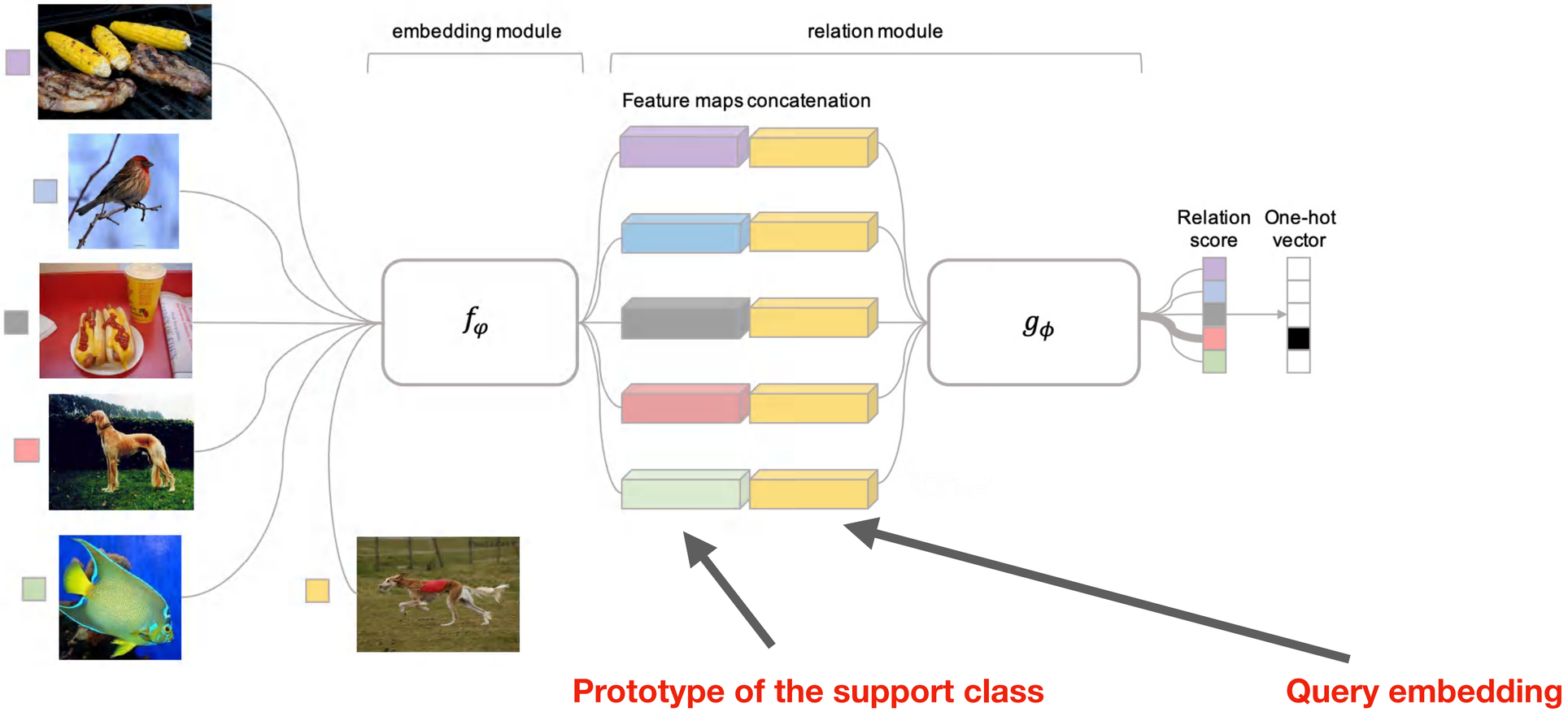
$$c(u, v) = -\|u - v\|^2$$

- Essentially a Gaussian classifier applied to the few-shot setting
- Prototypical networks for few-shot learning (2017)
  Jake Snell, Kevin Swersky, Richard Zemel
- Metric Learning for Large Scale Image Classification:
  Generalizing to New Classes at Near-Zero Cost (2012)
  Thomas Mensink, Jakob Verbeek, Florent Perronnin, Gabriela Csurka

Same probability distribution as matching networks in the one-shot case, when using dot product

Relation Networks

Combine prototypical networks and siamese networks

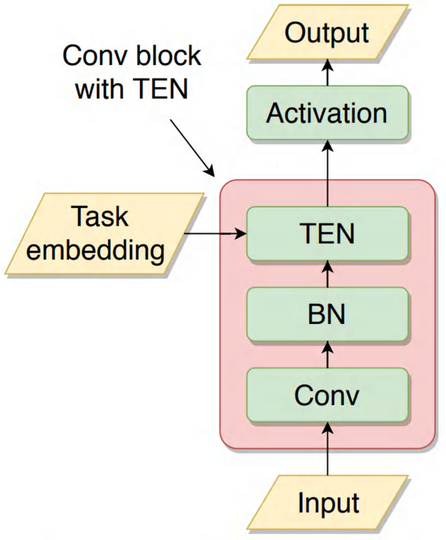


- Learning to compare: relation network for few-shot learning (2018)
  Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, Timothy M. Hospedales

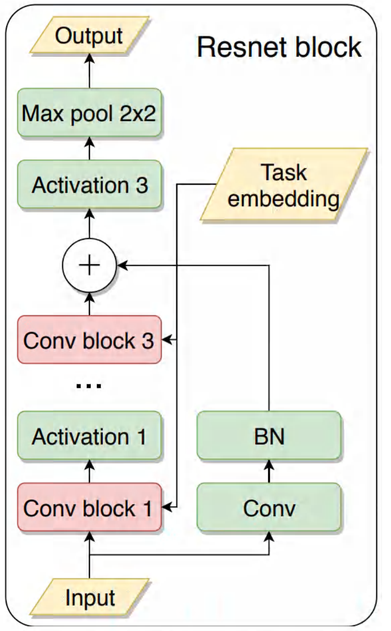
TADAM

Scale the distance metric and learn a task embedding

- Several contributions
  1. Scale the distance metric used in protonets
  2. Average the prototypes to get a task embedding
     - Use this as input to a "task embedding network" (TEN)
     - Use the TEN to rescale conv blocks
     - This is similar to conditional batchorm[1]



(a) Convolutional block with TEN.



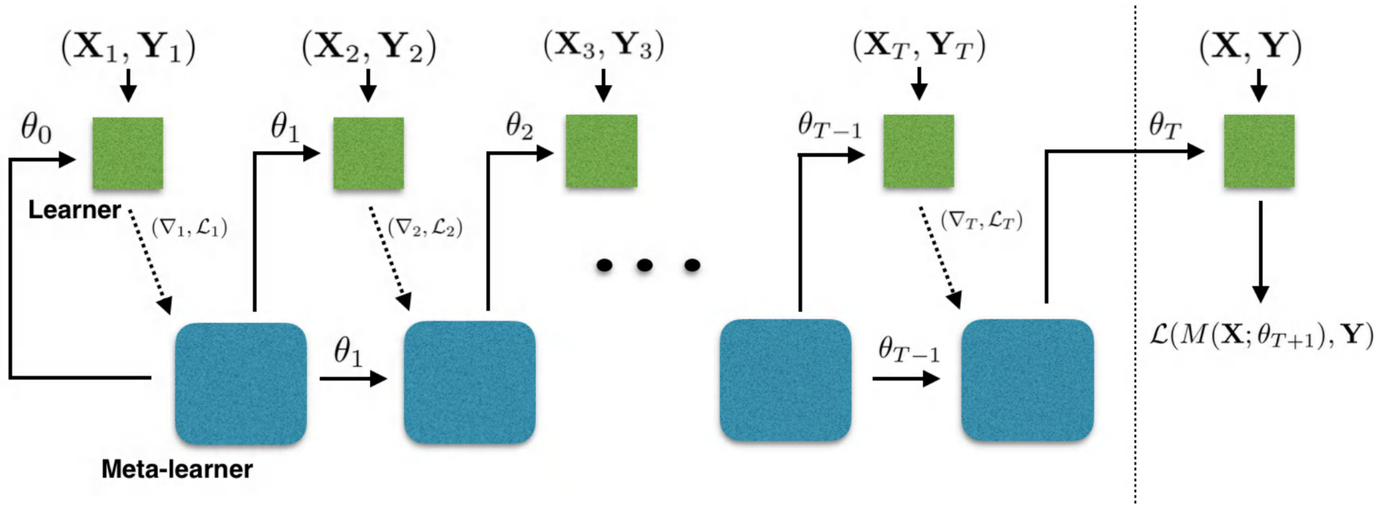
(b) Resnet block with TEN.

[1] FiLM: visual reasoning with a general conditioning layer (2017)
Ethan Perez, Harm de Vries, Florian Strub, Vincent Dumoulin, Aaron Courville

- A Simple Neural Attentive Meta-Learner (2018)
  Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, Pieter Abbeel

Optimization as a model

Use an LSTM to learn a gradient descent policy



Gradient descent: $\theta_t = \theta_{t-1} - \alpha_t \nabla \mathcal{L}(\theta_{t-1})$

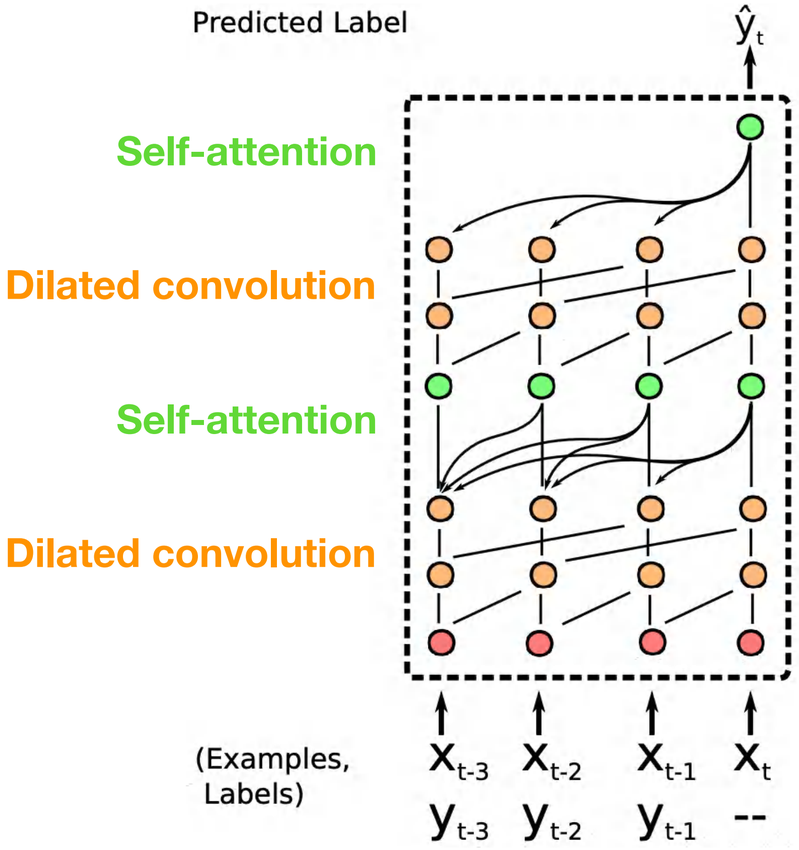
LSTM cell update: $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$

- Key idea: we can parameterize gradient descent on an episode using an LSTM.
- Learns an adaptive step size policy

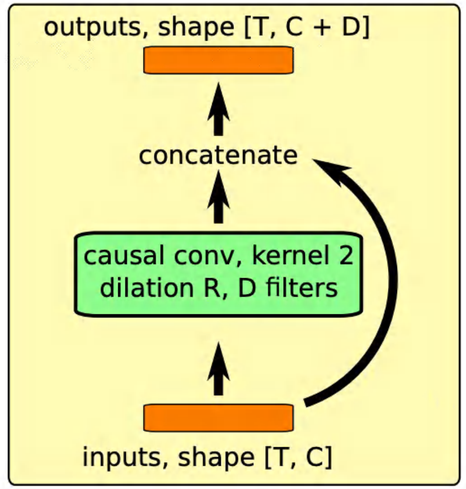
- Optimization as a model for few-shot learning (2017)
  Sachin Ravi and Hugo Larochelle

SNAIL: simple neural attentIve learner

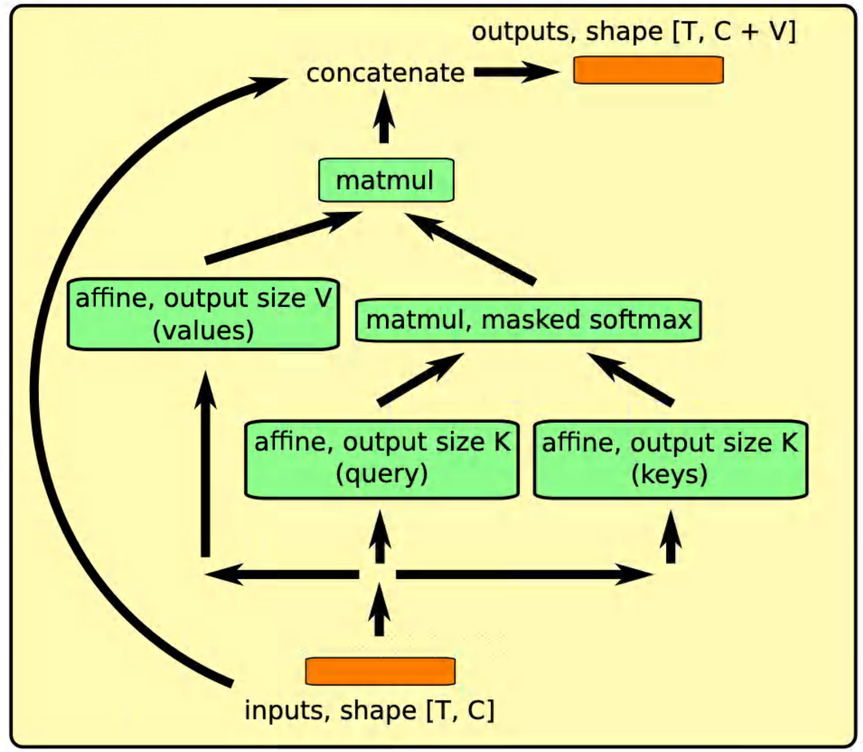
Treat the data as a time-series and apply next-step prediction using dilated convolutions and self-attention



(a) Dense Block (dilation rate R, D filters)



(b) Attention Block (key size K, value size V)

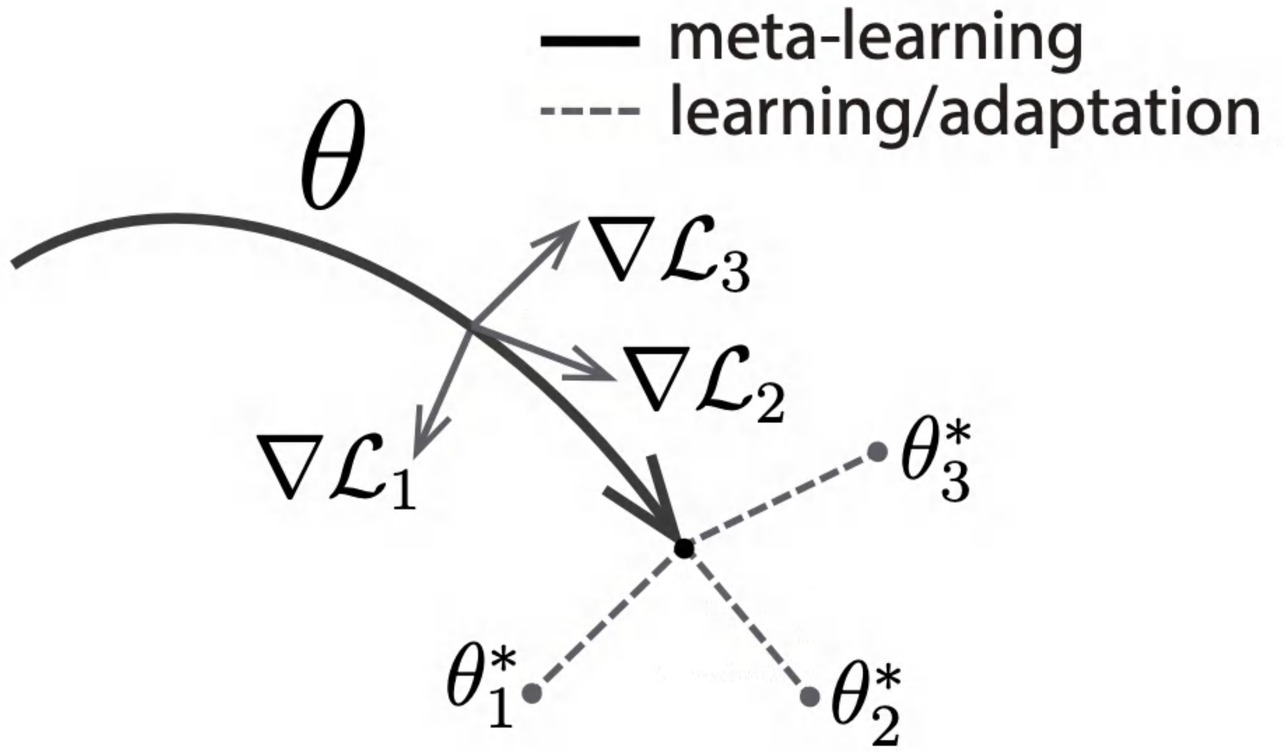


- A Simple Neural Attentive Meta-Learner (2018)
  Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, Pieter Abbeel

MAML: model agnostic meta learning

Find a robust initial set of parameters that makes gradient descent work across episodes

- MAML learns a set of weights θ_0 such that doing a few steps of gradient descent on the support set will yield good results.
- This is like fine-tuning, but all the weights are adjusted.
- The optimization itself is backpropagated through to find θ_0 with good gradient descent dynamics.
- Second-order gradients are sometimes ignored, creating first-order MAML. This doesn't seem to hurt results in few-shot learning.

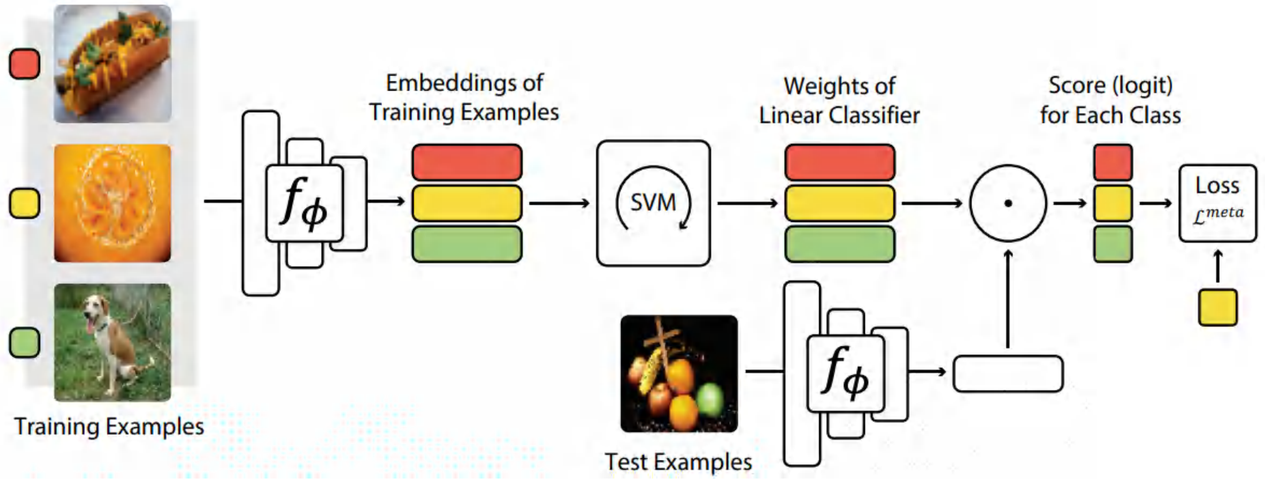


- Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks (2017)
  Chelsea Finn, Pieter Abbeel, Sergey Levine

MetaOptNet

Train a linear SVM on top of an embedding network

- OptNet showed how to backprop through differentiable quadratic programs
- MetaOptNet uses this for few-shot learning
- Idea: train a dual SVM on top of an embedding network



- Meta-Learning with Differentiable Convex Optimization (2019)
  Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, Stefano Soatto
- *OptNet: Differentiable Optimization as a Layer in Neural Networks*
  Brandon Amos, J. Zico Kolter

Proto-MAML

A better initialization for the output weights of a MAML network

- When using Euclidean distance, the energy of a prototypical network is equivalent to a linear classifier

$$
\begin{aligned}
\|f(x) - c\|^2 &= f(x)^\top f(x) - 2c^\top f(x) + c^\top c \\
&= Wx + b + const \\
W &= -2c \\
b &= c^\top c
\end{aligned}
$$

- Use the average of each class c to initialize the weights of a MAML network

- Meta-dataset: a dataset of datasets for learning to learn from few examples (2019)
  Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, Hugo Larochelle

Extra considerations

- Choose a powerful backbone
  - Often the architecture is more important than the method
- Contextual embeddings
  - Embed the support set using e.g., self-attention layers (matching nets use bidirectional LSTM) before applying a method

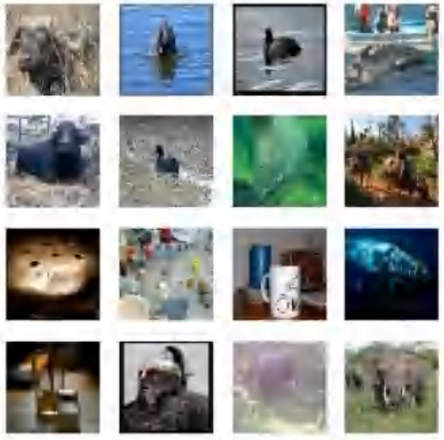
Performance

		miniImageNet 5-way	
model	**backbone**	**1-shot**	**5-shot**
Meta-Learning LSTM* [22]	64-64-64-64	43.44 ± 0.77	60.60 ± 0.71
Matching Networks* [33]	64-64-64-64	43.56 ± 0.84	55.31 ± 0.73
MAML [8]	32-32-32-32	48.70 ± 1.84	63.11 ± 0.92
Prototypical Networks*† [28]	64-64-64-64	49.42 ± 0.78	68.20 ± 0.66
Relation Networks* [29]	64-96-128-256	50.44 ± 0.82	65.32 ± 0.70
R2D2 [3]	96-192-384-512	51.2 ± 0.6	68.8 ± 0.1
Transductive Prop Nets [14]	64-64-64-64	55.51 ± 0.86	69.86 ± 0.65
SNAIL [18]	ResNet-12	55.71 ± 0.99	68.88 ± 0.92
Dynamic Few-shot [10]	64-64-128-128	56.20 ± 0.86	73.00 ± 0.64
AdaResNet [19]	ResNet-12	56.88 ± 0.62	71.94 ± 0.57
TADAM [20]	ResNet-12	58.50 ± 0.30	76.70 ± 0.30
Activation to Parameter† [21]	WRN-28-10	59.60 ± 0.41	73.74 ± 0.19
LEO† [25]	WRN-28-10	61.76 ± 0.08	77.59 ± 0.12
MetaOptNet-RR (ours)	ResNet-12	61.41 ± 0.61	77.88 ± 0.46
MetaOptNet-SVM (ours)	ResNet-12	62.64 ± 0.61	78.63 ± 0.46
MetaOptNet-SVM-trainval (ours)†	ResNet-12	**64.09 ± 0.62**	**80.00 ± 0.45**

- Meta-Learning with Differentiable Convex Optimization (2019)
  Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, Stefano Soatto

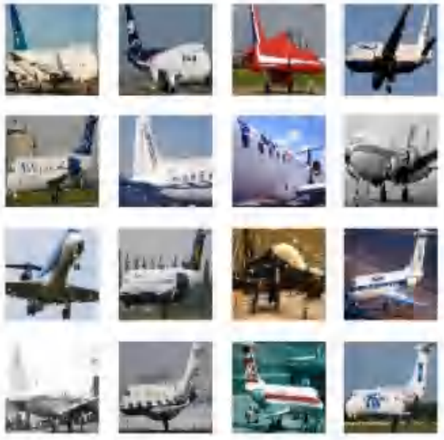
Large scale experiments using meta-dataset

- Meta-dataset: a dataset of datasets for learning to learn from few examples (2019)
  Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, Hugo Larochelle

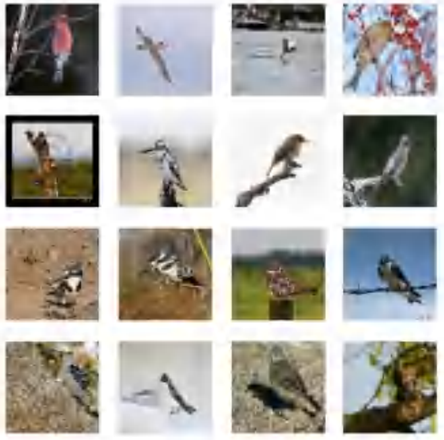


(a) ImageNet

(b) Omniglot

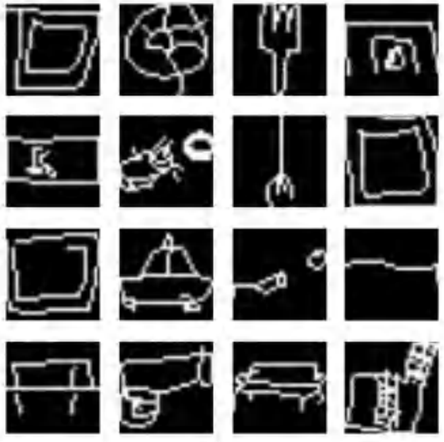


(c) Aircraft



(d) Birds

(e) DTD

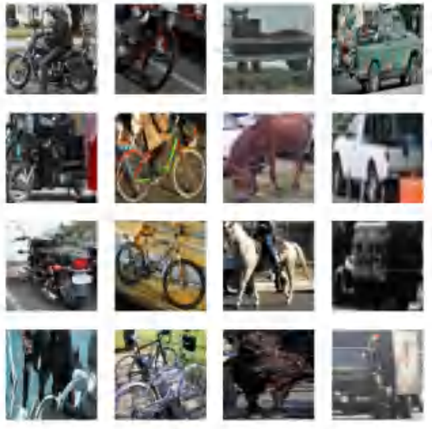


(f) Quick Draw

(g) Fungi

(h) VGG Flower

(i) Traffic Signs



(j) MSCOCO

Key features

- Many diverse image datasets
- Hierarchically aware (ImageNet ILSVRC-2012, Omniglot)
- Heterogeneous episodes
  - Different number of classes, different number of support examples per class
- Tests generalization across domains
  - Traffic Signs and MSCOCO never seen during training
- Optimize over a wide range of hyperparameter settings (including backbone architecture)
  - Allow methods to initialize from pre-trained features
- Two settings: train on ImageNet only, or train on all datasets (except Traffic Signs and MSCOCO)

Tables report accuracy

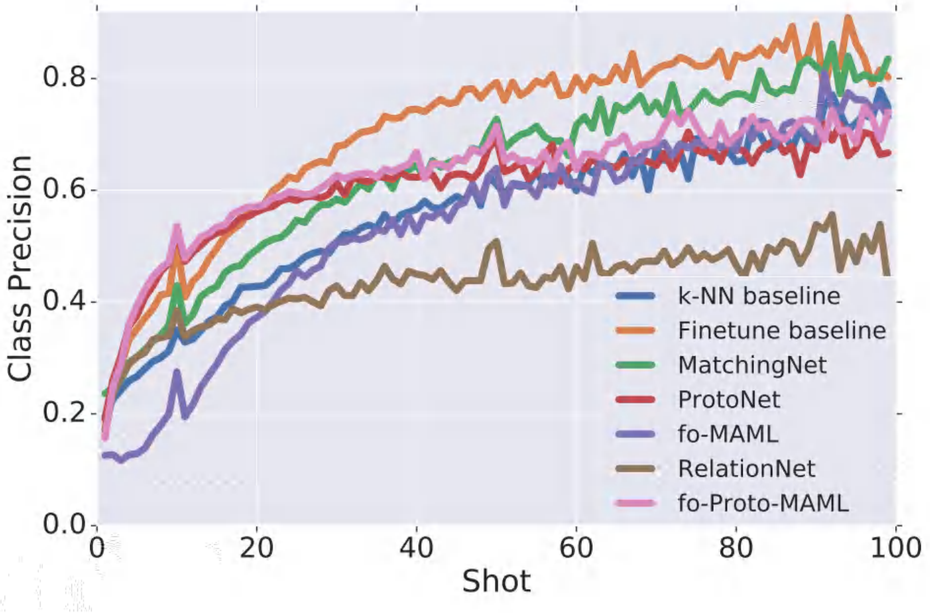
Train on ImageNet

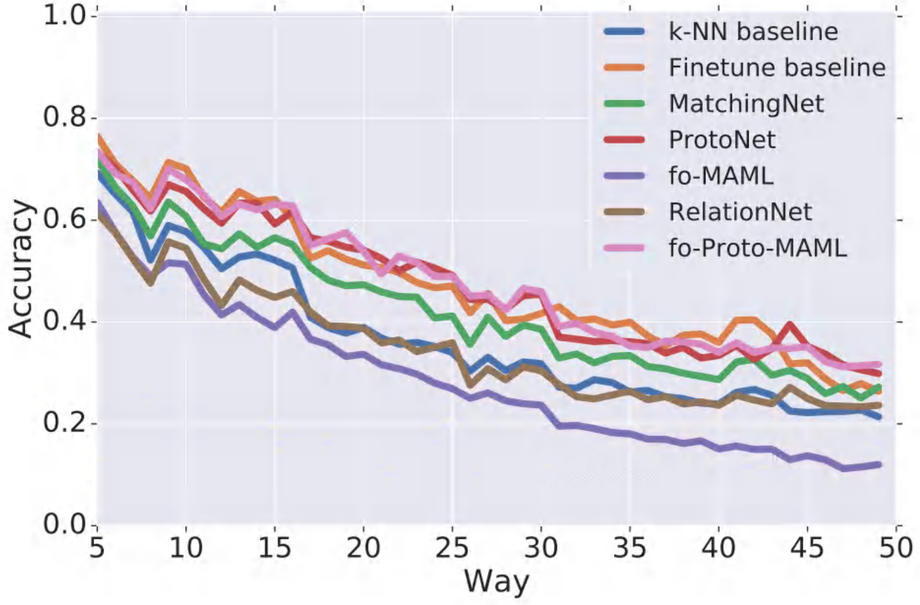
Test Source	k-NN	Finetune	MatchingNet	ProtoNet	fo-MAML	Relation Net	fo-Proto-MAML
ILSVRC	41.03	45.78	45.00	50.50	36.09	34.69	**51.01**
Omniglot	37.07	60.85	52.27	59.98	38.67	45.35	**63.00**
Aircraft	46.81	**68.69**	48.97	53.10	34.50	40.73	55.31
Birds	50.13	57.31	62.21	**68.79**	49.10	49.51	66.87
Textures	66.36	**69.05**	64.15	66.56	56.50	52.97	67.75
Quick Draw	32.06	42.60	42.87	48.96	27.24	43.30	**53.70**
Fungi	36.16	38.20	33.97	**39.71**	23.50	30.55	37.97
VGG Flower	83.10	85.51	80.13	85.27	66.42	68.76	**86.86**
Traffic Signs	44.59	**66.79**	47.80	47.12	33.23	33.67	51.19
MSCOCO	30.38	34.86	34.99	41.00	27.52	29.15	**43.41**
Avg. rank	5	2.5	4	2.4	6.7	5.8	**1.6**

Train on all datasets

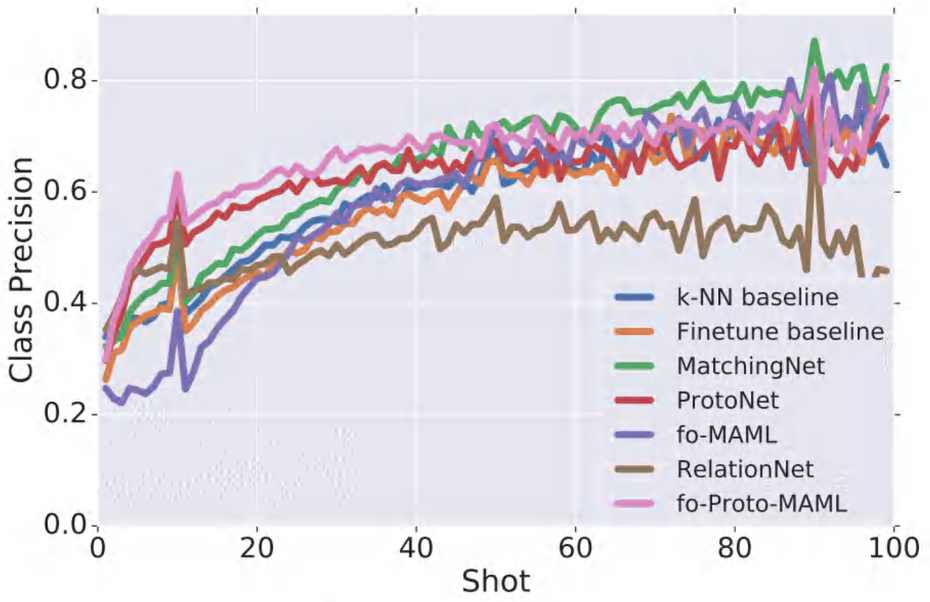
Test Source	k-NN	Finetune	MatchingNet	ProtoNet	fo-MAML	Relation Net	fo-Proto-MAML
ILSVRC	38.55	43.08	36.08	44.50	32.36	30.89	**47.85**
Omniglot	74.60	71.11	78.25	79.56	71.91	**86.57**	82.86
Aircraft	64.98	72.03	69.17	71.14	52.76	69.71	**74.24**
Birds	66.35	59.82	56.40	67.01	47.24	54.14	**69.97**
Textures	63.58	**69.14**	61.80	65.18	56.66	56.56	67.94
Quick Draw	44.88	47.05	60.81	64.88	50.50	61.75	**66.57**
Fungi	37.12	38.16	33.70	40.26	21.02	32.56	**41.99**
VGG Flower	83.47	85.28	81.90	86.85	70.93	76.08	**88.45**
Traffic Signs	40.11	**66.74**	55.57	46.48	34.18	37.48	52.32
MSCOCO	29.55	35.17	28.79	39.87	24.05	27.41	**41.29**
Avg. rank	4.6	3.3	4.5	2.5	6.5	5.2	**1.4**

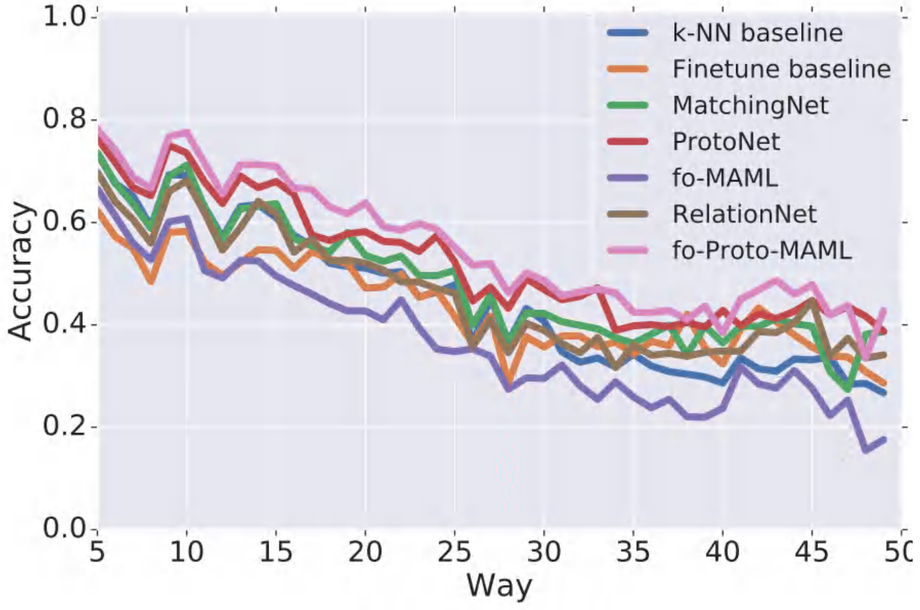
Trained on Imagenet





Trained on all datasets





Extensions and variations

Zero-shot classification

Train

Blue jay

From Wikipedia, the free encyclopedia

For other uses, see Blue jay (disambiguation).

The **blue jay** (*Cyanocitta cristata*) is a passerine bird in the family Corvidae, native to North America. It resides through most of eastern and central United States, although western populations may be migratory. Resident populations are also found in Newfoundland, Canada, while breeding populations can be found across southern Canada. It breeds in both deciduous and coniferous forests, and is common in residential areas. It is predominantly blue with a white chest and underparts, and a blue crest. It has a black, U-shaped collar around its neck and a black border behind the crest. Both sexes are similar in size and plumage, and plumage does not vary throughout the year. Four subspecies of the blue jay have been recognized.

The blue jay mainly feeds on nuts and seeds such as acorns, soft fruits, arthropods, and occasionally small vertebrates. It typically gleans food from trees, shrubs, and the ground, though it sometimes hawks insects from the air. Like squirrels, blue jays are known to hide nuts for later consumption.[2] It builds an open cup nest in the branches of a tree, which both sexes participate in constructing. The clutch can contain two to seven eggs, which are blueish or light brown with brown spots. Young are altricial, and are brooded by the female for 8–12 days after hatching. They may remain with their parents for one to two months.

The name "jay" derives from its noisy, garrulous nature and has been applied to other birds of the same family, which are also mostly gregarious.[3] It is sometimes called a "jaybird".[4]

Test query

Test annotation

American robin

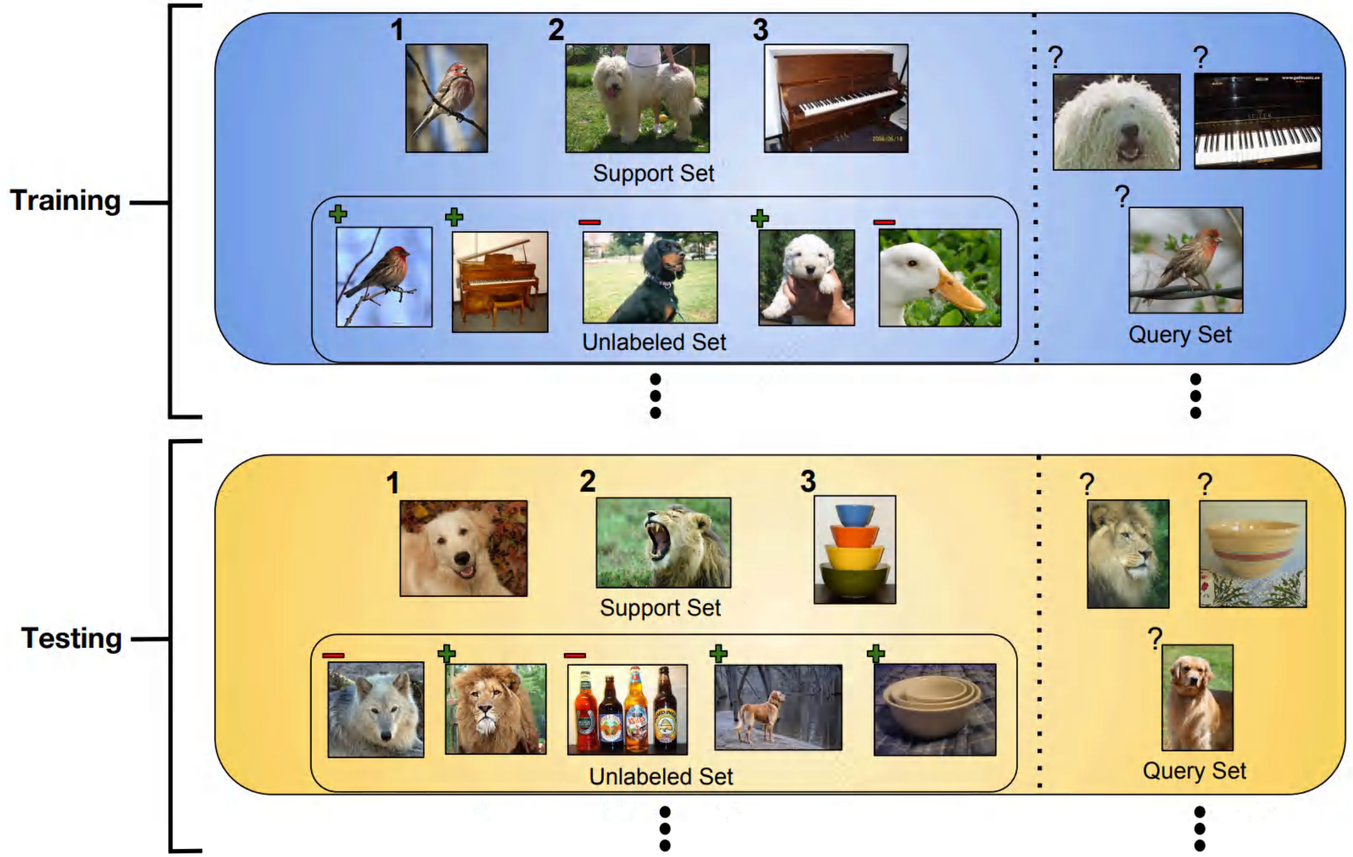
From Wikipedia, the free encyclopedia

The **American robin** (*Turdus migratorius*) is a migratory songbird of the true thrush genus and Turdidae, the wider thrush family. It is named after the European robin[2] because of its reddish-orange breast, though the two species are not closely related, with the European robin belonging to the Old World flycatcher family. The American robin is widely distributed throughout North America, wintering from southern Canada to central Mexico and along the Pacific Coast. It is the state bird of Connecticut, Michigan, and Wisconsin.[3] According to some sources, the American robin ranks behind only the red-winged blackbird (and just ahead of the introduced European starling and the not-always-naturally-occurring house finch) as the most abundant extant land bird in North America.[4] It has seven subspecies, but only *T. m. confinis* of Baja California Sur is particularly distinctive, with pale gray-brown underparts.

The American robin is active mostly during the day and assembles in large flocks at night. Its diet consists of invertebrates (such as beetle grubs, earthworms, and caterpillars), fruits, and berries. It is one of the earliest bird species to lay its eggs, beginning to breed shortly after returning to its summer range from its winter range. The robin's nest consists of long coarse grass, twigs, paper, and feathers, and is smeared with mud and often cushioned with grass or other soft materials. It is among the earliest birds to sing at dawn, and its song consists of several discrete units that are repeated.

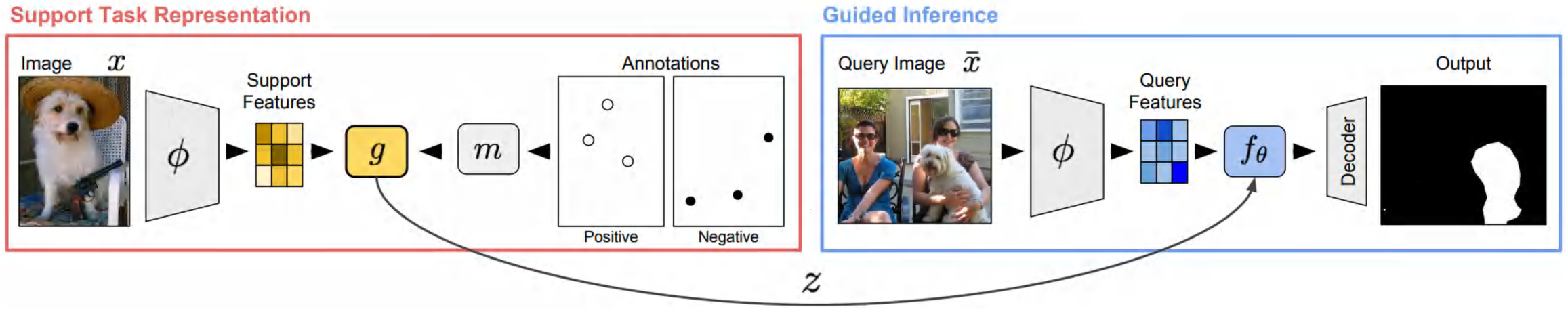
The adult robin's main predators are hawks, domestic cats, and snakes. When feeding in flocks, it can be vigilant, watching other birds for reactions to predators. Brown-headed cowbirds (*Molothrus ater*) lay eggs in robin nests (see brood parasite), but the robins usually reject the eggs.

Semi-supervised few-shot learning



- Meta-Learning for Semi-Supervised Few-Shot Classification
  Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, Richard S. Zemel (2018)

Few-shot segmentation



- Few-Shot Segmentation Propagation with Guided Networks
  Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei Efros, Sergey Levine (2018)

Conclusions

- Few-shot classification is still largely unsolved
  - Lots of techniques, but baselines are still extremely strong by comparison
- Perhaps time to revisit Bayesian techniques
- Data augmentation, self-supervision, generative models are also currently under-explored
- Lots of work being done to develop robust and memory efficient meta-learners
- Important to think about real-world applications and environments