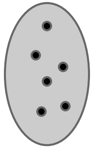
Model vs Optimization Meta Learning

Oriol Vinyals @OriolVinyalsML
Research Scientist, Deepmind

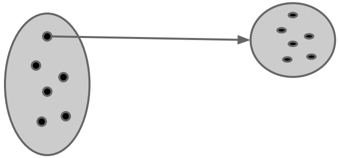
NIPS, December 2017

Definition of Meta Learning

- What is Meta Learning / Learning to Learn?
    - Go beyond train from samples from a single distribution.



    - Distribution over tasks, so model has to "learn to learn" when a new task is presented

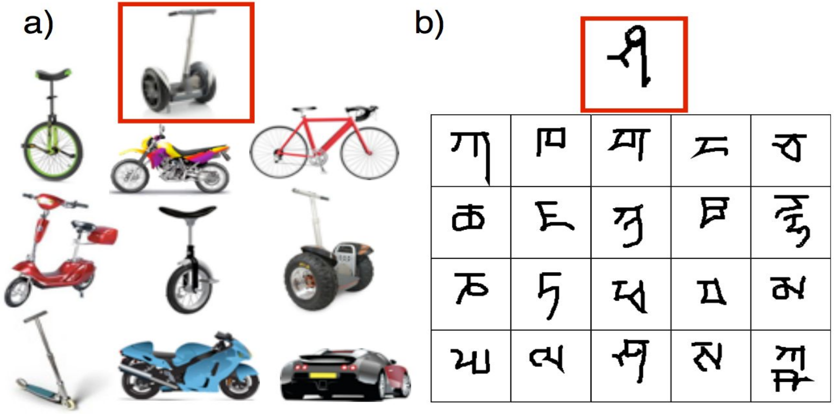


"... a system that improves or discovers a learning algorithm"

Hochreiter et al, '01

Datasets: Omniglot

- To make progress, we need datasets / metrics!



Lake et al, 2013, 2015

Datasets: Mini-ImageNet

- To make progress, we need datasets / metrics!

Vinyals et al, 2016

Datasets: Beyond

- To make progress, we need datasets / metrics!

Reinforcement learning

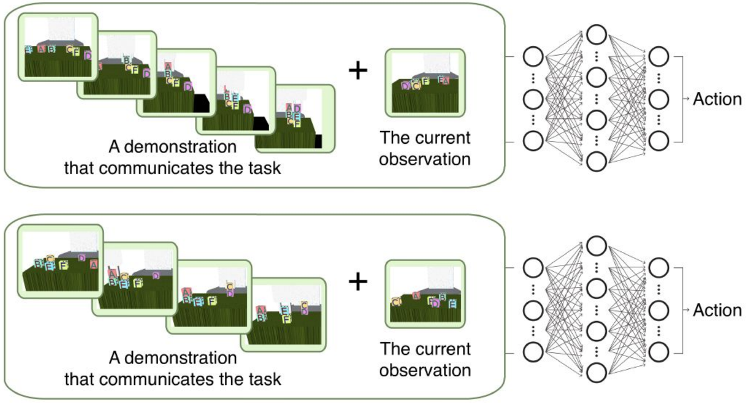
Given a small amount of experience

Solve a new task

Chelsea Finn, UC Berkeley

How? learn to learn many other tasks

fig. from Duan et al. '17



Yan Duan, Marcin Andrychowicz, Bradly Stadie, Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, Wojciech Zaremba (2017)

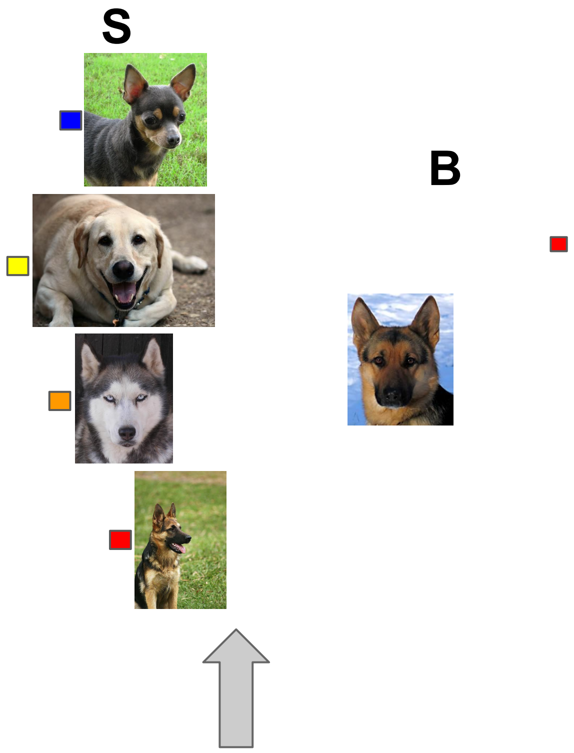
Training Setup: An “Episode”

1. Sample label set **L** from **T**

L =[Pinscher,
Golden Retriever,
Husky,
German Shepherd]

Training Setup: An “Episode”

1. Sample label set **L** from **T**
2. Sample a few images as support set **S** from **L**
3. Sample a few images as batch **B** from **L**



L =[Pinscher,
Golden Retriever,
Husky,
German Shepherd]

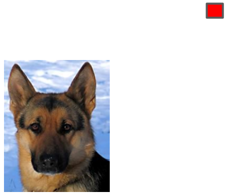
Oriol Vinyals, NIPS 17

Training Setup: An “Episode”

1. Sample label set **L** from **T**
2. Sample a few images as support set **S** from **L**
3. Sample a few images as batch **B** from **L**
4. Optimize batch, Go to 1

S

B



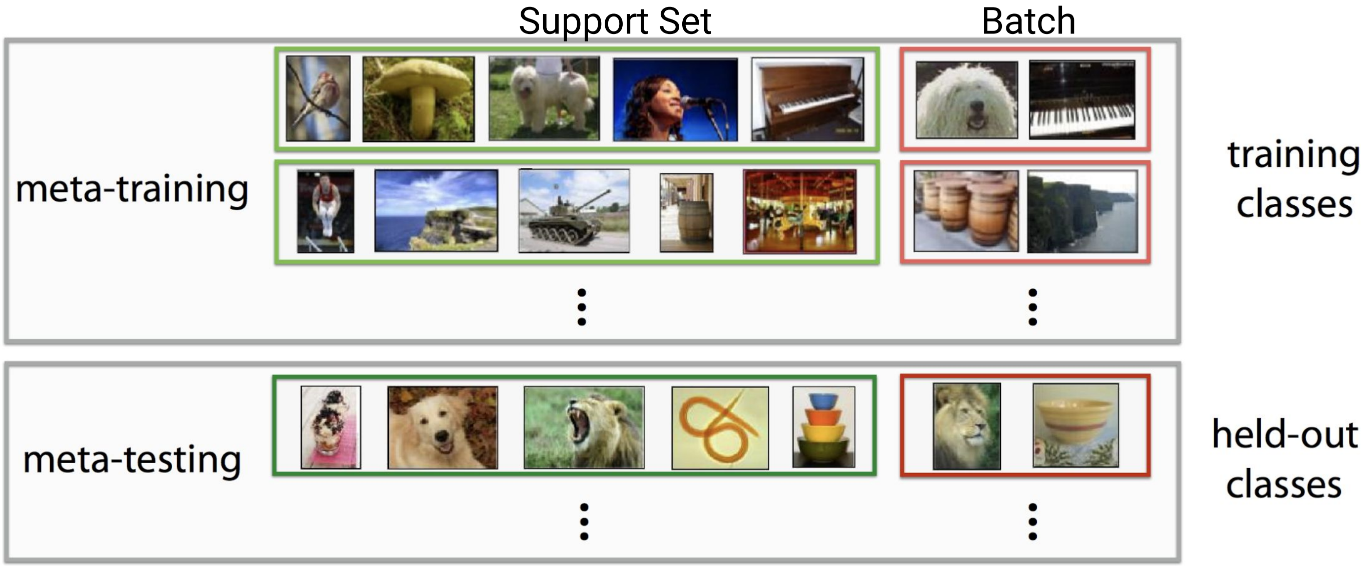
$$\theta = \arg\max_{\theta} E_{L\sim T}\left[E_{S\sim L,B\sim L}\left[\sum_{(x,y)\in B}\log P_{\theta}\left(y|x,S\right)\right]\right].$$

L =[Pinscher,
Golden Retriever,
Husky,
German Shepherd]

Contrasting with Supervised Learning

$$\theta = \arg\max_{\theta} \left[E_{B} \left[\sum_{(x,y)\in B} \log P_{\theta}\left(y|x\right) \right]\right].$$

Contrasting with Supervised Learning



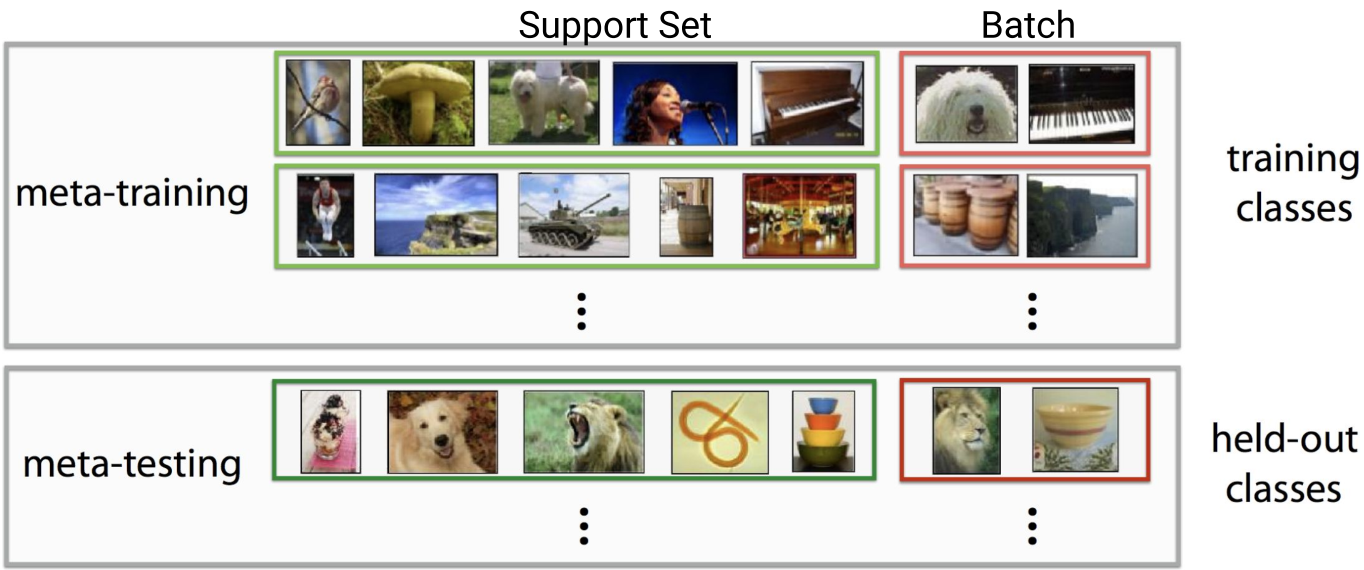
Chelsea Finn, Berkeley AI

diagram adapted from Ravi & Larochelle '17

$$\theta = \arg\max_{\theta} E_{L \sim T} \left[E_{S \sim L, B \sim L} \left[\sum_{(x,y) \in B} \log P_{\theta}\left(y | x, S\right) \right] \right].$$

Oriol Vinyals, NIPS 17

Contrasting with Supervised Learning



??????
??????

Chelsea Finn, Berkeley AI

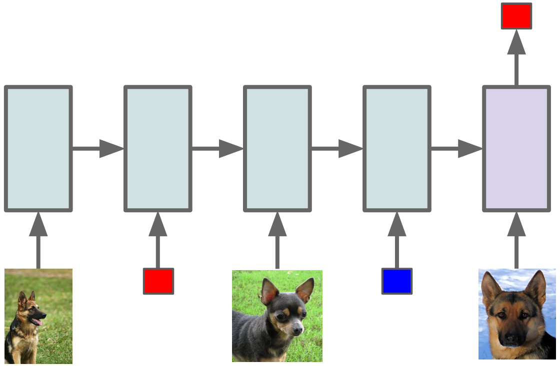
diagram adapted from Ravi & Larochelle '17

$$\theta = \arg\max_{\theta} E_{L\sim T}\left[E_{S\sim L, B\sim L}\left[\sum_{(x,y)\in B} \log P_{\theta}\left(y|x,S\right) \right]\right].$$

Oriol Vinyals, NIPS 17

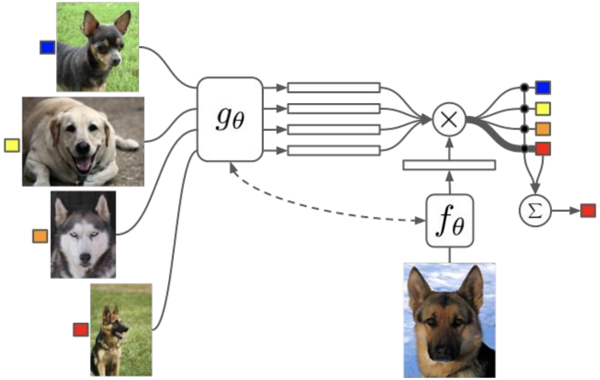
Meta Learning Models Taxonomy

Model Based



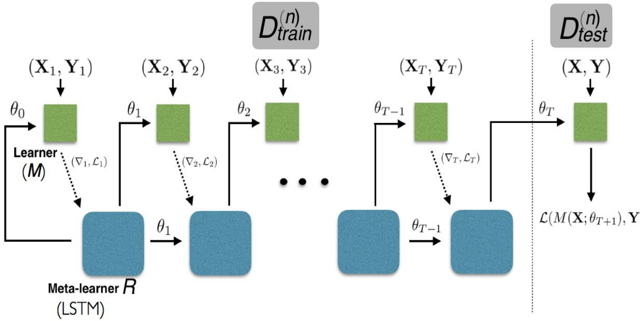
- Santoro et al. ’16
- Duan et al. ’17
- Wang et al. ‘17
- Munkhdalai & Yu ‘17
- Mishra et al. ‘17

Metric Based



- Koch ’15
- Vinyals et al. ‘16
- Snell et al. ‘17
- Shyam et al. ‘17
- Sung et al. ‘17

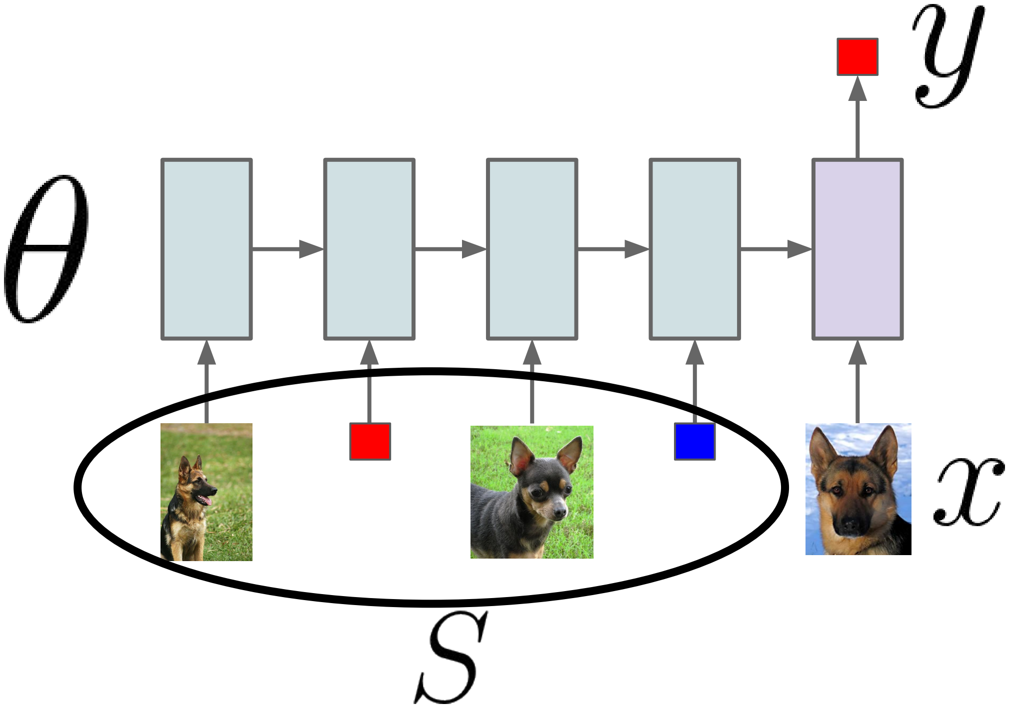
Optimization Based



- Schmidhuber ’87, ’92
- Bengio et al. ’90, ‘92
- Hochreiter et al. ’01
- Li & Malik ‘16
- Andrychowicz et al. ’16
- Ravi & Larochelle ‘17
- Finn et al. ‘17

Adapted from Finn ‘17

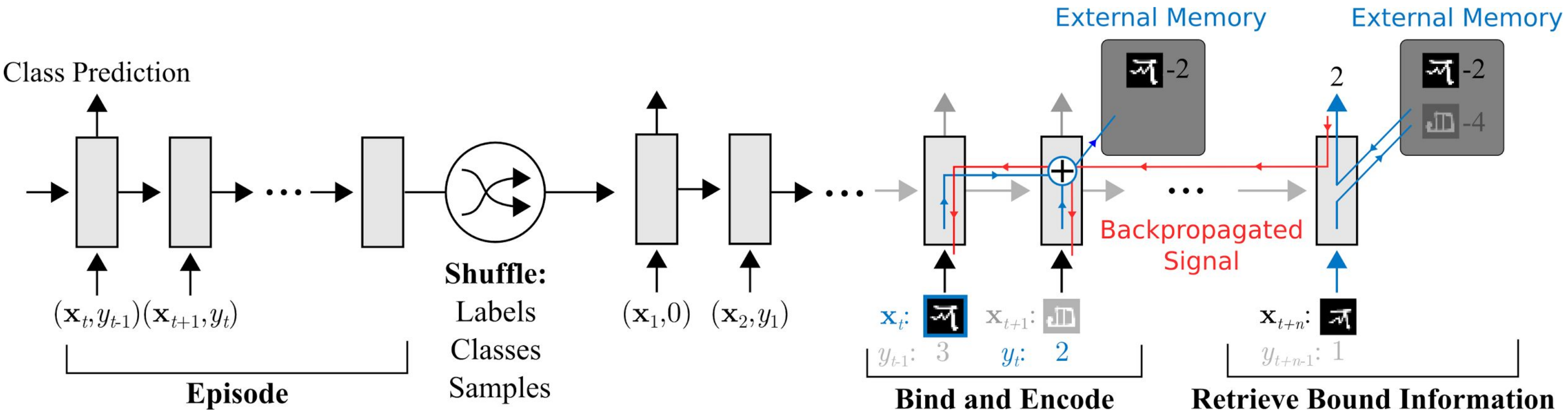
Model Based Meta Learning



$$P_\theta(y|x, S) = f_\theta(x, S)$$

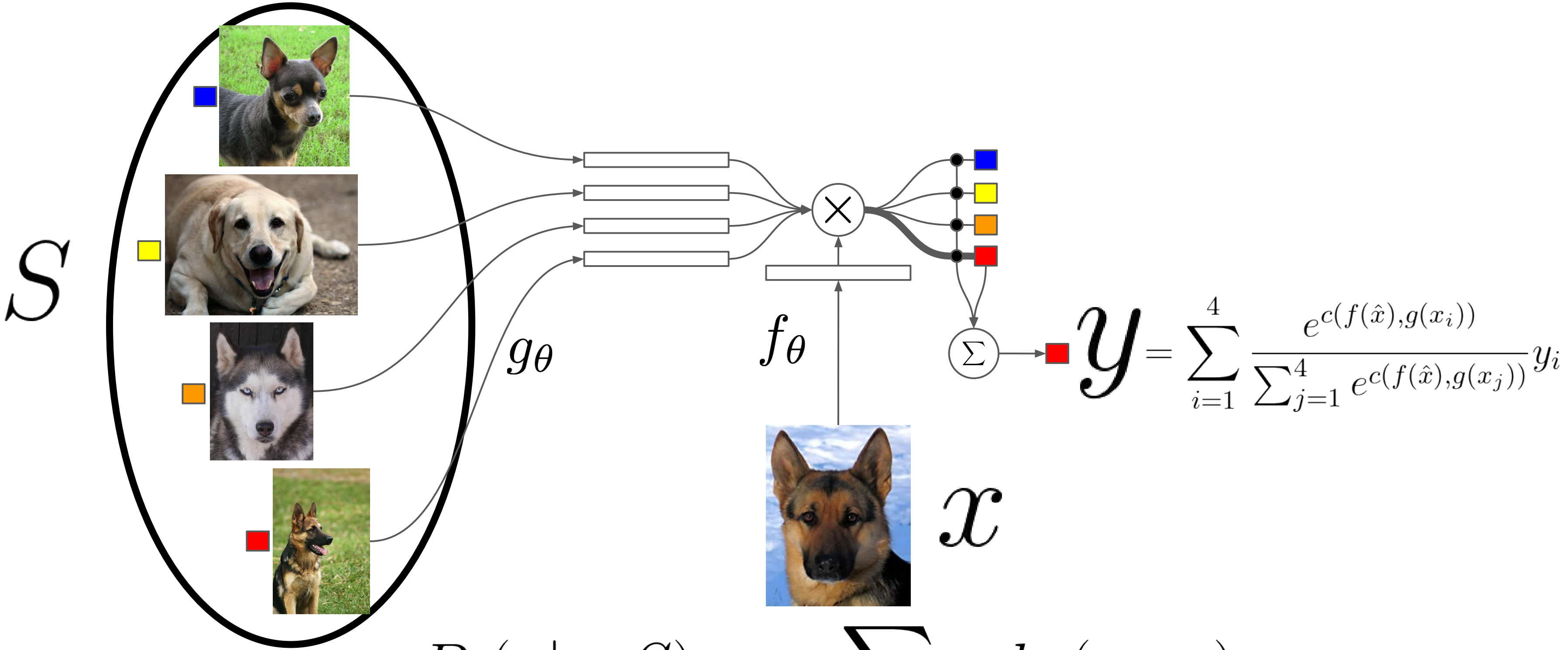
Model Based Meta Learning

Santoro et al, ICML 2016



Slide Credit: Adam Santoro

Metric Based Meta Learning

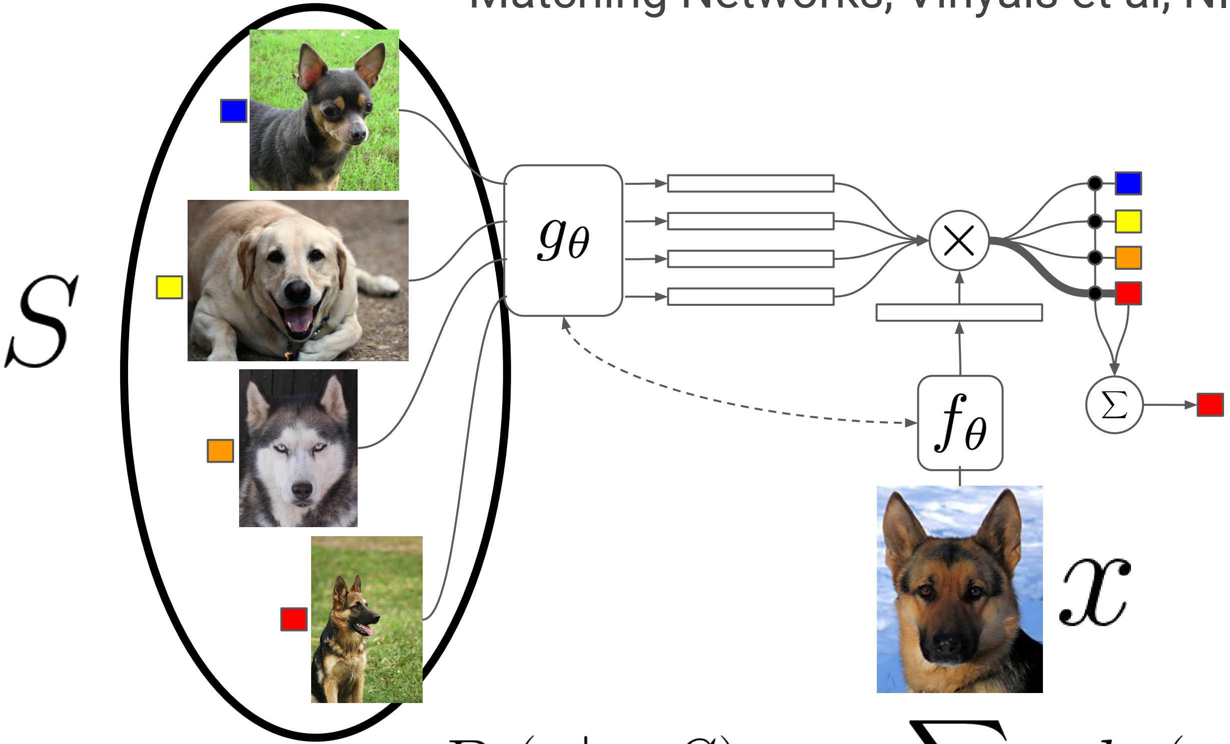


$$P_\theta(y|x, S) = \sum_{(x_i, y_i) \in S} k_\theta(x, x_i) y_i$$

Oriol Vinyals, NIPS 17

Metric Based Meta Learning

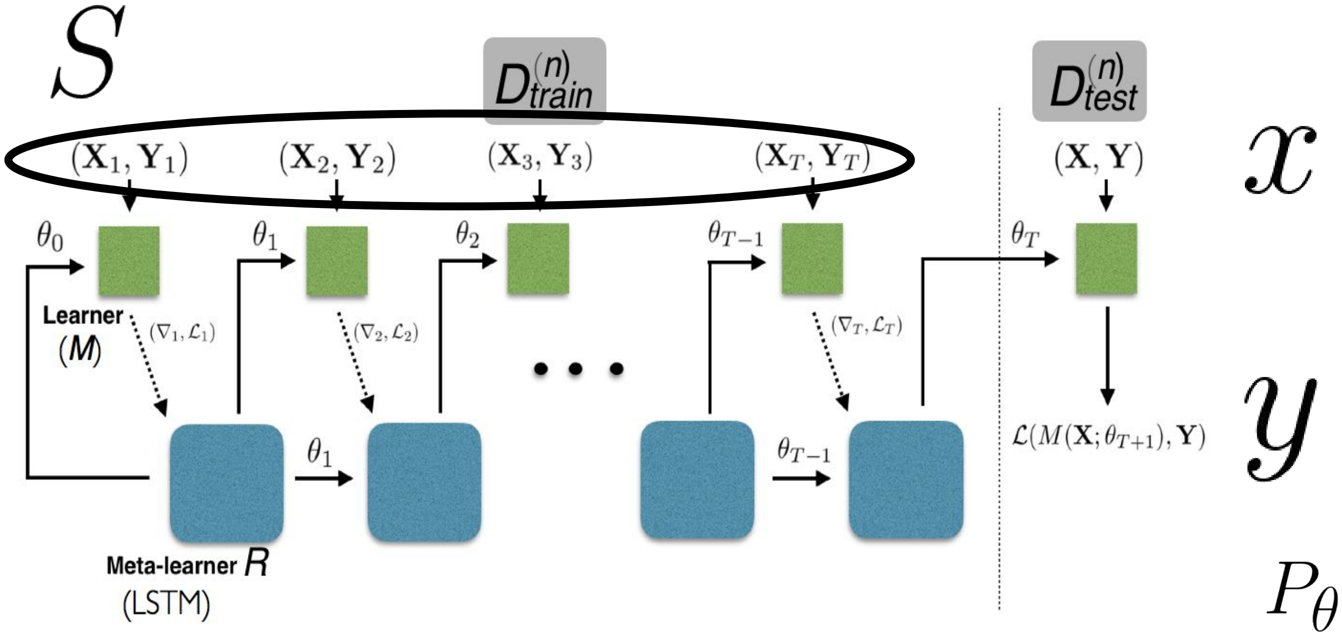
Matching Networks, Vinyals et al, NIPS 2016



$$y = \sum_{i=1}^{4} \frac{e^{c(f(\hat{x},S),g(x_i,S))}}{\sum_{j=1}^{4} e^{c(f(\hat{x},S),g(x_j,S))}} y_i$$

$$P_\theta(y|x,S) = \sum_{(x_i,y_i)\in S} k_\theta(x, x_i, S) y_i$$

Optimization Based Meta Learning



$$P_\theta(y|x, S) = f_{\theta(S)}(x)$$

$$\theta(S) = g_{\theta_g}(\theta_0, \{\nabla_{\theta_0} L(x_i, y_i)\}_{(x_i, y_i) \in S})$$

Figure Credit: Hugo Larochelle

Examples of Optimization Based Meta Learning

Finn et al, 17

$$\theta = \theta_0 - \eta \sum_{(x_i, y_i) \in S} \nabla_{\theta_0} L(x_i, y_i)$$

Ravi et al, 17

$$\theta_t = f_t \odot \theta_{t-1} + i_t \odot \nabla_{\theta_{t-1}} L(x_t, y_t)$$

$$P_\theta(y|x, S) = f_{\theta(S)}(x)$$

$$\theta(S) = g_{\theta_g}(\theta_0, \{\nabla_{\theta_0} L(x_i, y_i)\}_{(x_i, y_i) \in S})$$

Progress on Mini-ImageNet

	Model	FT	5-way Acc.	
			1-shot	5-shot
Metric	MATCHING NETS [38]	N	$43.56 \pm 0.84\%$	$55.31 \pm 0.73\%$
Model	META NETS [26]	N	$49.21 \pm 0.96\%$	-
Optim	META-LEARN LSTM [28]	N	$43.44 \pm 0.77\%$	$60.60 \pm 0.71\%$
Optim	MAML [10]	Y	$48.70 \pm 1.84\%$	$63.11 \pm 0.92\%$
Metric	PROTOTYPICAL NETS [35]	N	$49.42 \pm 0.78\%$	$\mathbf{68.20 \pm 0.66\%}$
Metric	RELATION NET (NAIVE)	N	$\mathbf{51.38 \pm 0.82\%}$	$\mathbf{67.07 \pm 0.69\%}$
Model	TCML [25]	N	$55.71 \pm 0.99\%$	$68.88 \pm 0.92\%$
Metric	RELATION NET (DEEPER)	N	$\mathbf{57.02 \pm 0.92\%}$	$\mathbf{71.07 \pm 0.69\%}$

Table from Sung et al, 17

Summing Up

Model Based

$$P_\theta(y|x,S) = f_\theta(x,S)$$

Metric Based

$$P_\theta(y|x,S) = \sum_{(x_i,y_i)\in S} k_\theta(x,x_i)y_i$$

Optimization Based

$$P_\theta(y|x,S) = f_{\theta(S)}(x)$$

$$\theta(S) = g_{\theta_g}(\theta_0, \{\nabla_{\theta_0} L(x_i,y_i)\}_{(x_i,y_i)\in S})$$

Future Work

- Combining Model / Metric / Optimization based approaches
  - Reed et al, 2017
- Meta-Meta-Meta… learning
  - Tasks need to be related / from same distribution
- What are the right inductive biases?
  - Spatial invariance → convolution
  - Temporal sequences → recurrence
  - Learning → gradients?

Thanks!! Questions??

@OriolVinyalsML

NIPS, December 2017