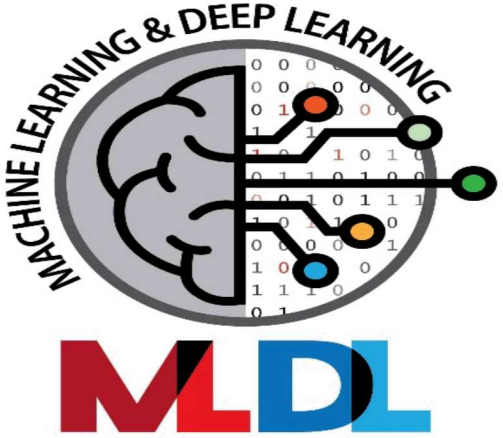


MLDL

Machine Learning and Deep Learning Conference 2018

SAND2018-8250PE

Survey of Few-Shot Learning Techniques

- Stephen Heck (6361), Phil Lewis (6362), Tim Draelos (6362)

Few-shot Learning: Omniglot Dataset

MNIST Dataset
10 Classes with > 6000 examples of each class

Omniglot is like “MNIST Transpose”
~ 6000 Classes with 20 examples of each class

Multiple alphabets: Japanese, Tibetan, Korean, Tengwar

Note: Few-shot learning does not mean few data samples!

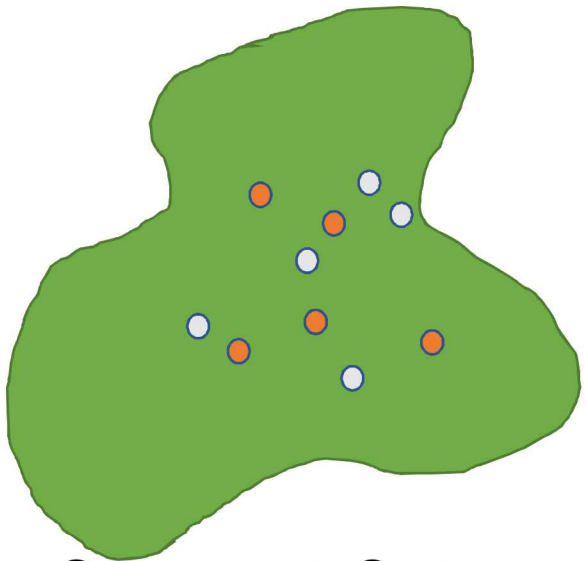
Omniglot Examples

Tibetan Character 2 (no rotations)

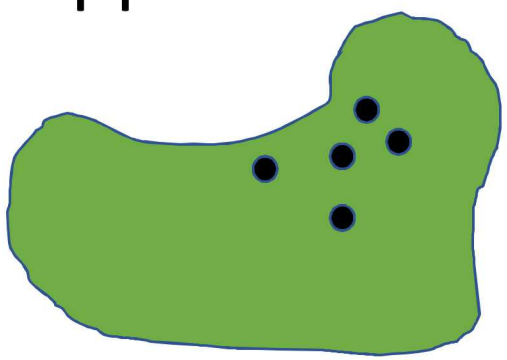
Tengwar Character 21 (no rotations)

Prototypical Networks Overview

Prototypical Networks for Few-shot Learning, Snell, Swersky, Zemel



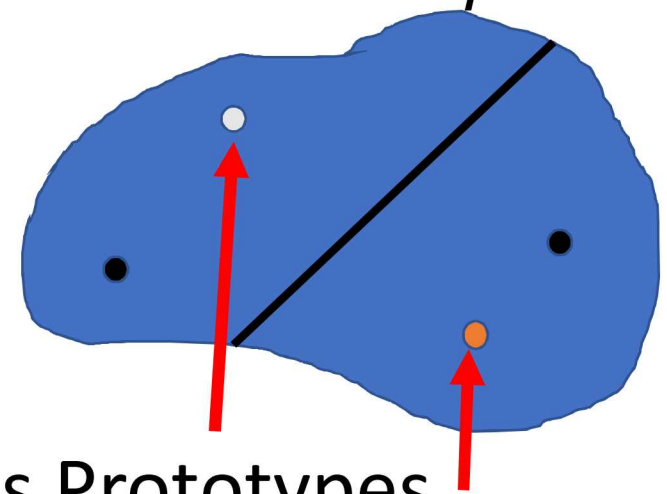
Support Set



Query Set

f_θ

$d_\varphi()$ partition



Class Prototypes

$$c_k = \frac{1}{n}\sum_{i=1}^{n} f_\theta(x_{k_i})$$

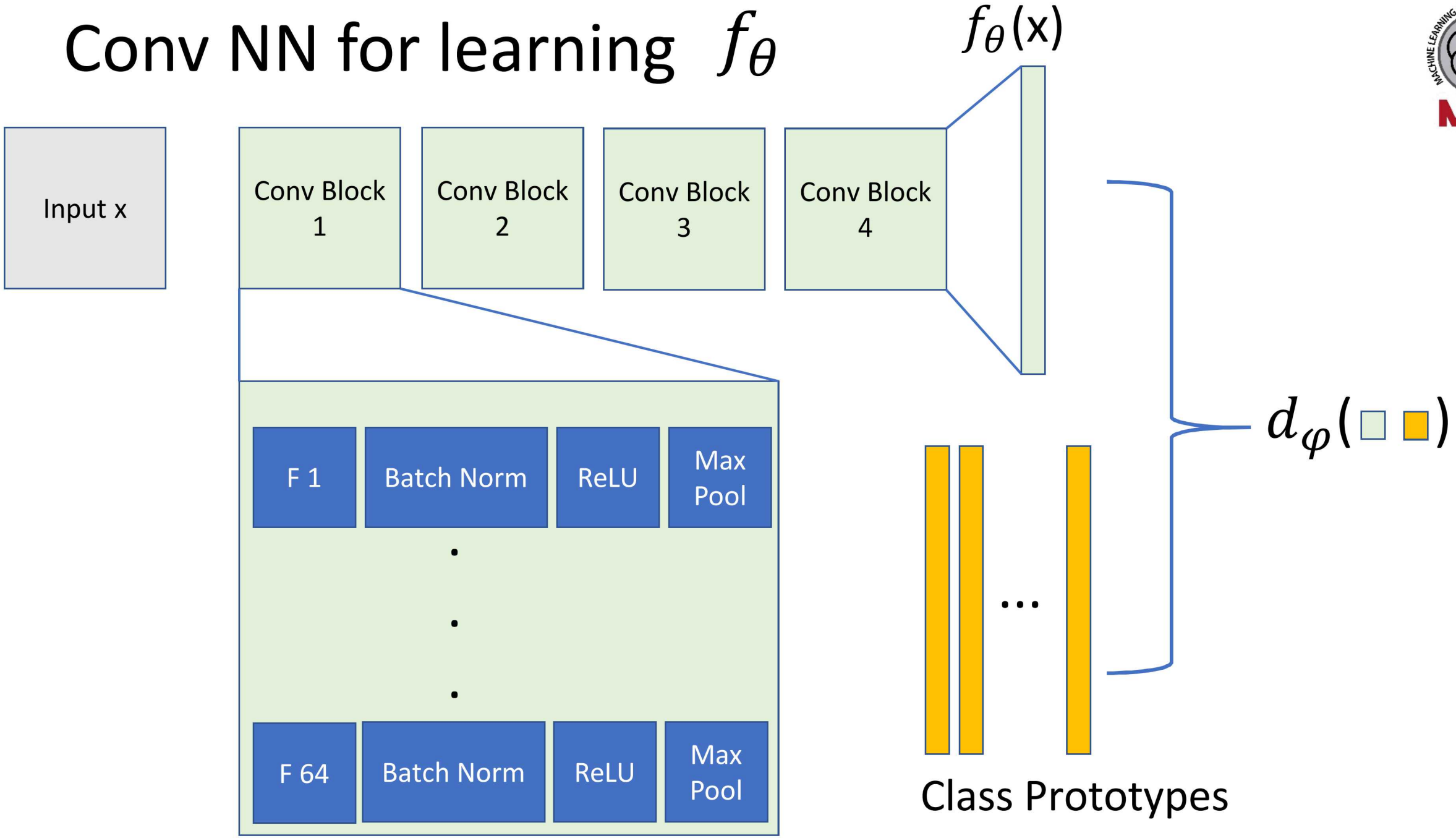
1. Compute Class Prototype for n-shot learning

$$c_k = \frac{1}{n}\sum_{i=1}^{n} f_\theta(x_{k_i})$$

2. Classify

$$class = \underset{k \in Classes}{\text{argmax}}\ softmax(d_\varphi(f_\theta(x), c_k))$$

Conv NN for learning f_θ



Learning episode

C = Select Random Sample of N Classes

For each class k:

S_k = Random Sample of N examples from Class k

Q_k = Random Sample of M examples from Class k not in S_k

Compute Class Prototype c_k

For each class k:

For Query Example in Q_k

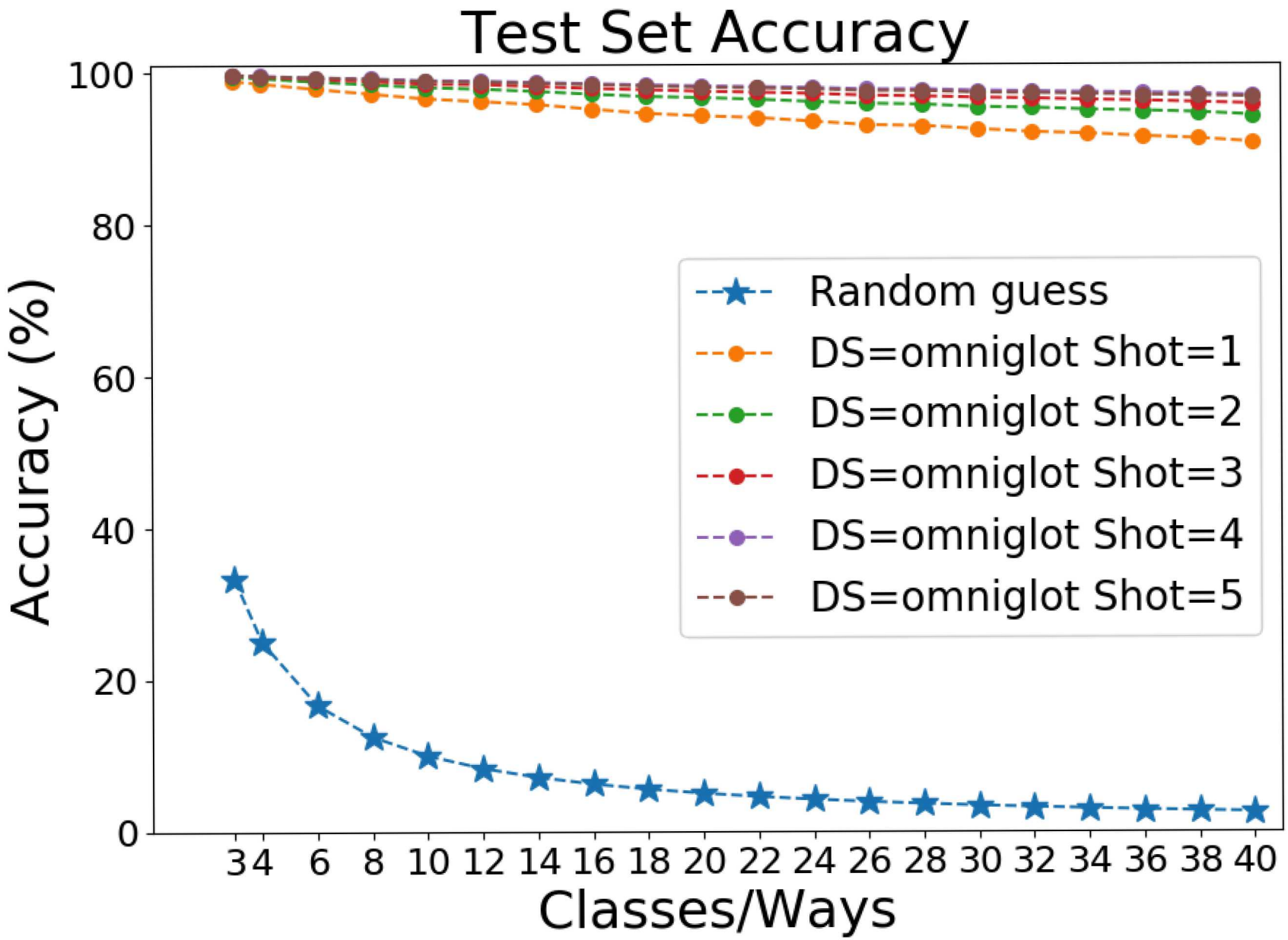
Update Loss/Neural net parameters θ

Notes:

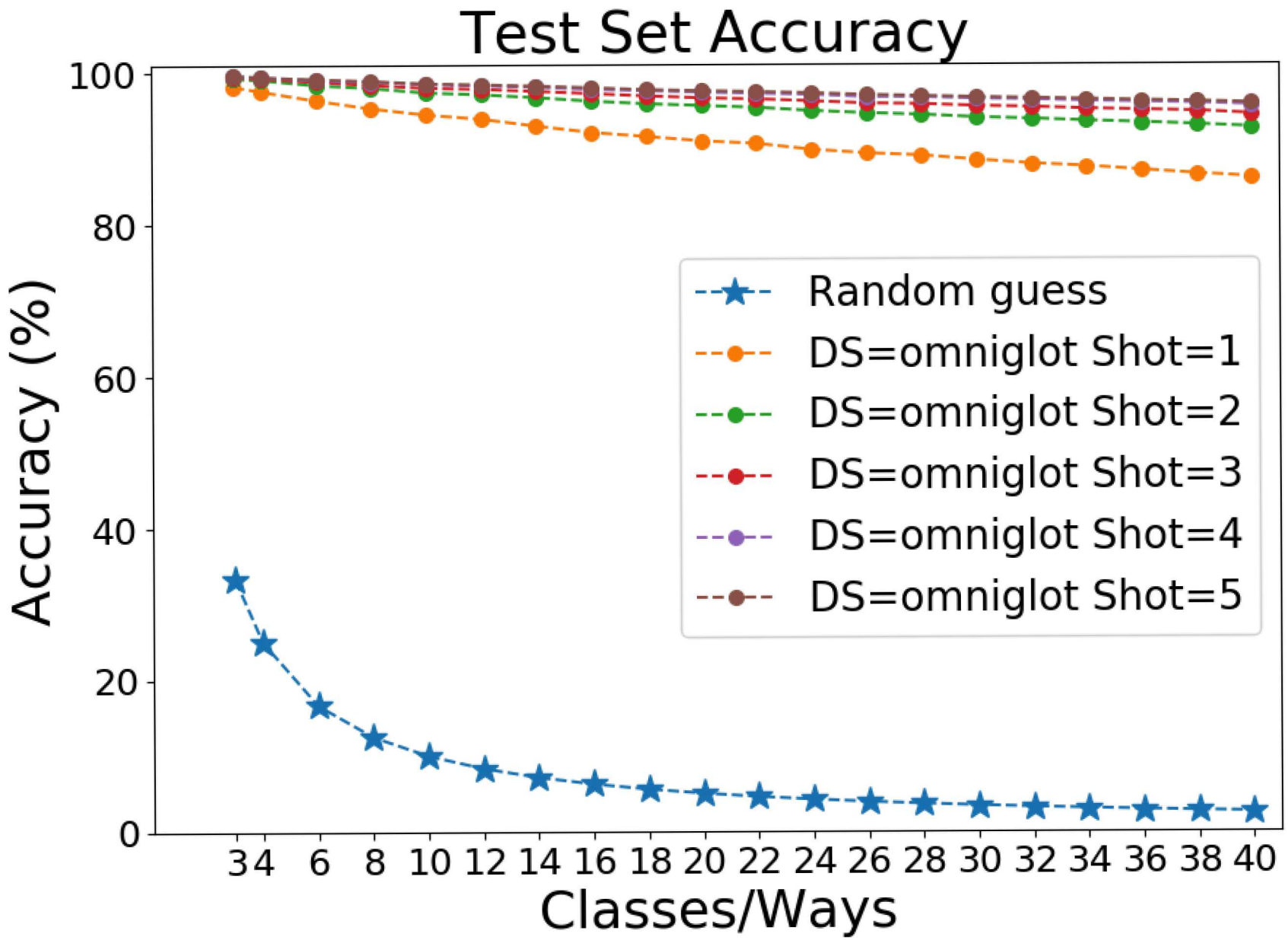
- 100 Learning episodes per Epoch
- Quit training when 200 Epochs have passed with no improvement on validation set

Datasets	Full		Reduced	
Train	# classes	4112	# classes	635
	# rotation classes	3084	# rotation classes	480
	# samples/classes	20	# samples/classes	20
	total samples	82240	total samples	12700
	# alphabets	33	# alphabets	33
Validation	# classes	688	# classes	98
	# rotation classes	516	# rotation classes	69
	# samples/classes	20	# samples/classes	20
	total samples	13760	total samples	1960
	# alphabets	5	# alphabets	5
Test Set	# classes	1692	# classes	1692
	# rotation classes	1269	# rotation classes	1269
	# samples/classes	20	# samples/classes	20
	total samples	33840	total samples	33840
	# alphabets	13	# alphabets	13

Full Dataset Results for 20-way training



Reduced Dataset Results for 20-way training



Full vs Reduced Datasets

Way/Classes

Shot	3	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40
1	98.1	97.5	96.3	95.2	94.4	93.8	92.9	92	91.5	**90.9**	90.5	89.7	89.2	88.9	88.3	87.8	87.4	86.9	86.4	**86**
1	98.8	98.5	97.8	97.1	96.5	96.1	95.7	95.1	94.5	**94.1**	93.9	93.4	92.9	92.8	92.3	91.9	91.7	91.4	91.1	**90.6**
2	99.2	99	98.3	97.9	97.3	97	96.6	96.1	95.8	95.5	95.2	94.8	94.5	94.2	93.9	93.7	93.4	93.1	92.9	92.5
2	99.4	99.2	98.8	98.4	98	97.8	97.4	97	96.7	96.5	96.3	96	95.7	95.6	95.3	95.1	94.9	94.7	94.5	94.1
3	99.4	99.2	98.8	98.3	97.9	97.7	97.5	97.1	96.8	96.5	96.3	96.1	95.8	95.6	95.4	95.2	95	94.8	94.6	94.3
3	99.6	99.4	99.1	98.8	98.4	98.4	98.1	97.8	97.6	97.4	97.2	97	96.8	96.7	96.5	96.4	96.2	96	95.8	95.6
4	99.6	99.4	99	98.7	98.4	98.2	98	97.7	97.5	97.3	97.1	96.9	96.6	96.6	96.3	96.2	96	95.8	95.7	95.4
4	99.7	99.5	99.3	99.2	98.9	98.8	98.6	98.4	98.2	98.1	98	97.8	97.6	97.6	97.4	97.3	97.2	97.1	96.9	96.7
5	99.6	99.4	99.1	98.8	98.5	98.3	98.1	97.9	97.6	**97.5**	97.3	97.1	96.9	96.7	96.5	96.4	96.2	96.1	95.9	**95.7**
5	99.7	99.5	99.3	99.1	98.8	98.7	98.5	98.3	98.1	**97.9**	97.8	97.7	97.4	97.4	97.1	97.1	96.9	96.8	96.7	**96.5**

= Reduced Dataset

Prototypical Networks Recap

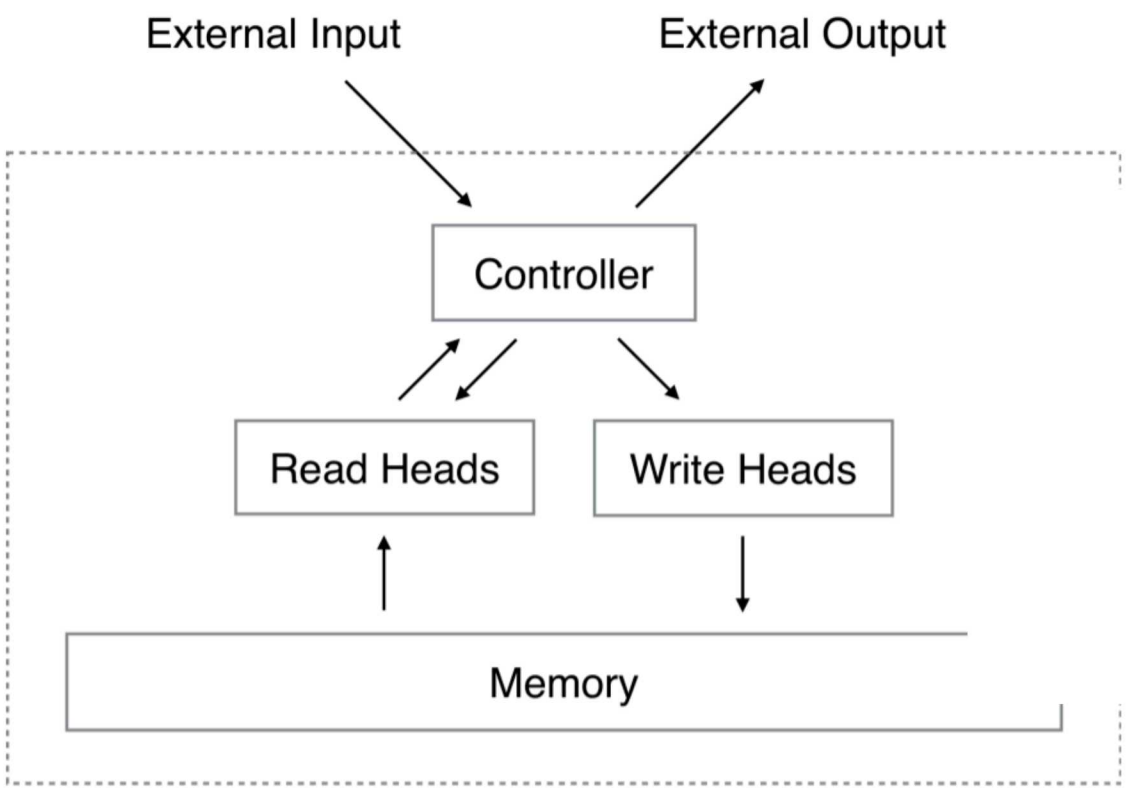
1. c_k is optimal w.r.t. training data when d is a Bregman divergence
    1. Euclidean Distance was used for these results

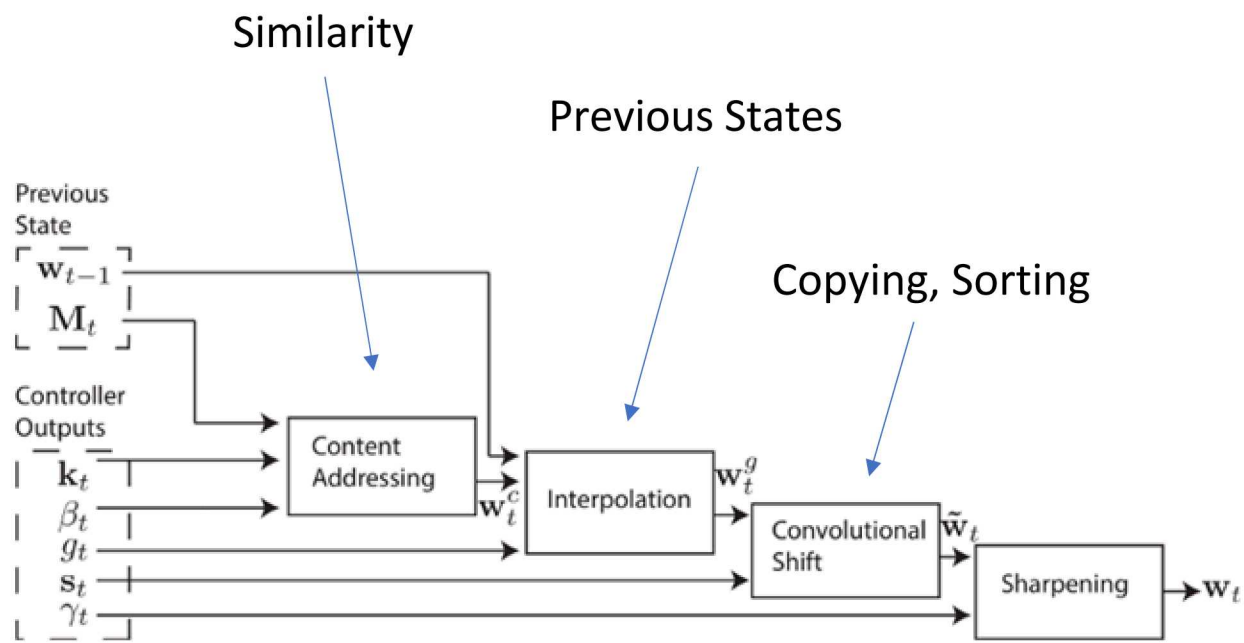
2. Versatile:
    1. Can train with (k-shot, n-way) and apply to NOT (k-shot, n-way) problem
    2. Prototypical Network don’t require retraining for new classes*

3. 1-shot degrades as
    1. Way goes up
    2. Training data set size goes down

1. Few-shot does not mean few data samples!

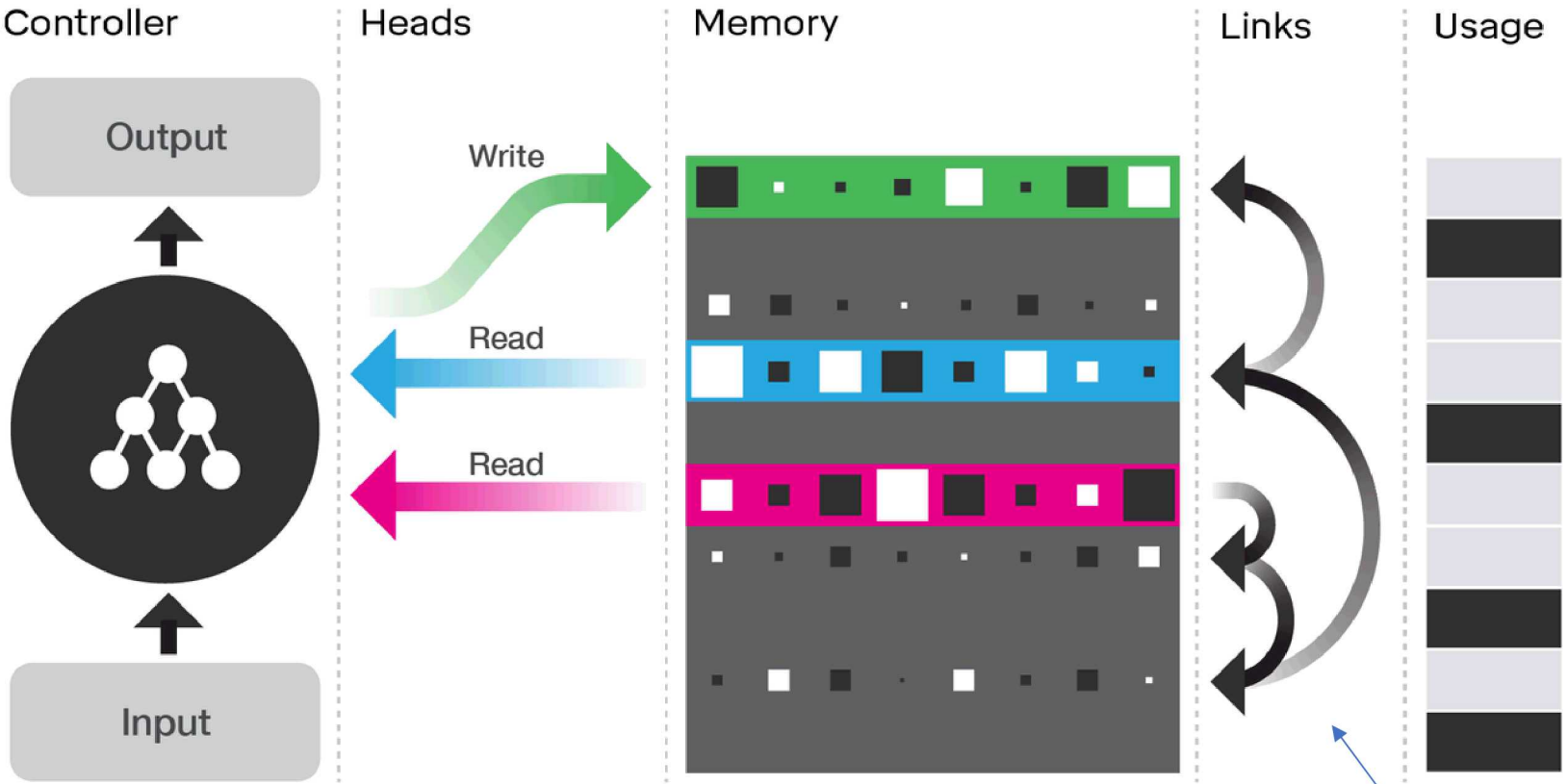
Neural Turing Machines





(Graves et al 2014)

Differentiable Neural Computers

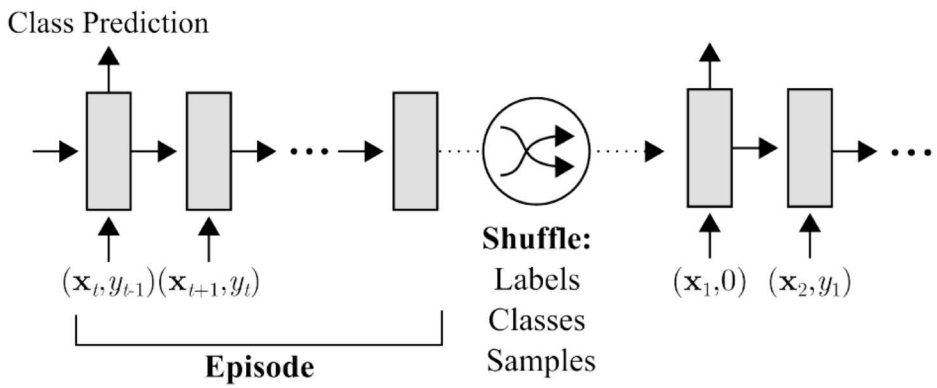


(Graves et al 2016)

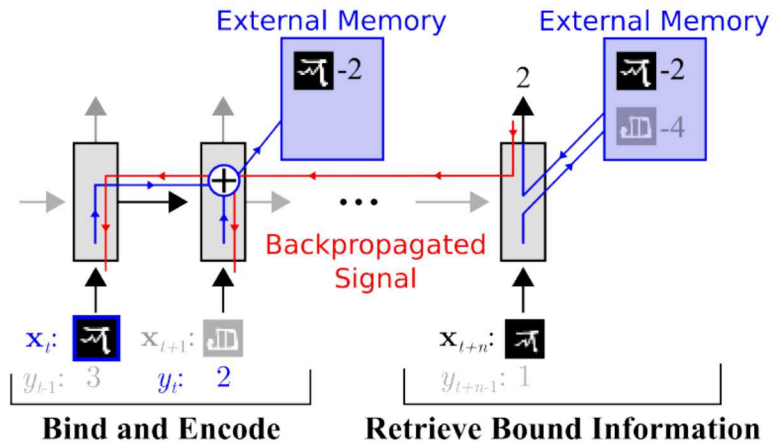
Dynamic Memory

More Flexible Temporal Associations
(No More Convolutional Shift)

Least Recently Used Access and One Shot



(a) Task setup



(b) Network strategy

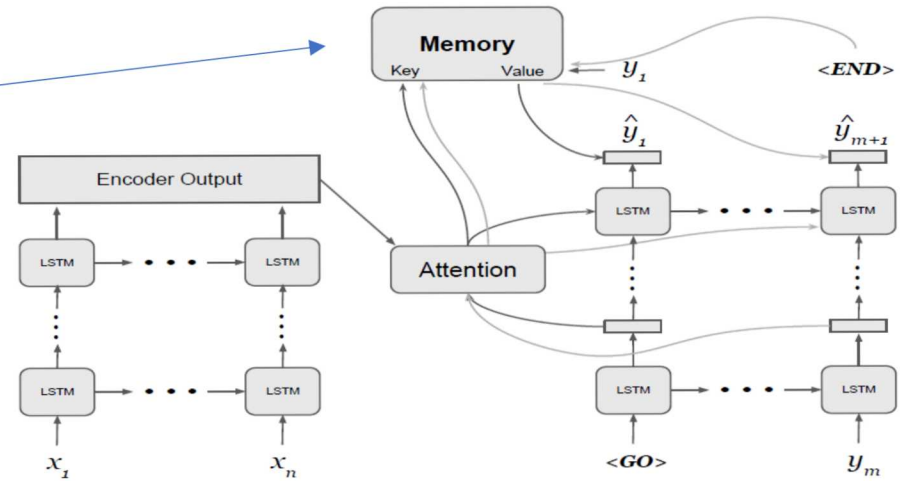
Model is learning deduction.
Is this few-shot or meta-learning?

Model	Instance (% Correct) 1st	2nd	3rd	4th	5th	10th
Human	34.5	57.3	70.1	71.8	81.4	92.4
Feedforward	24.4	19.6	21.1	19.9	22.8	19.5
LSTM	24.4	49.5	55.3	61.0	63.6	62.5
MANN	**36.4**	**82.8**	**91.0**	**92.6**	**94.9**	**98.1**

(Santoro et al 2016)

Nearest Neighbor Memory and Few-shot

Very Long Term Memory
For Rare Events



Traditional Few-Shot

Model	5-way 1-shot	5-way 5-shot	20-way 1-shot	20-way 5-shot
Pixels Nearest Neighbor	41.7%	63.2%	26.7%	42.6%
MANN (no convolutions)	82.8%	94.9%	–	–
Convolutional Siamese Net	96.7%	98.4%	88.0%	96.5%
Matching Network	98.1%	98.9%	93.8%	98.5%
ConvNet with Memory Module	98.4%	99.6%	95.0%	98.6%

(Kaiser et al 2017)

More Info and Results

https://gitlab.sandia.gov/deeplearning/low-shot

References

- One-shot Learning with Memory-Augmented Neural Networks (2016, Santoro, Bartunov, Google DeepMind)
- Learning to Remember Rare Events (2017, Kaiser, Nachum, Roy, Bengio)
- Hybrid Computing Using a Neural Network with Dynamic External Memory (2016, Graves, Wayne, Reynolds, Harley)
- https://codeburst.io/understanding-few-shot-intelligence-as-a-meta-learning-problem-7823a4cd4a0c
- http://rylanschaeffer.github.io/content/research/neural_turing_machine/main.html
- https://rylanschaeffer.github.io/content/research/one_shot_learning_with_memory_augmented_nn/main.html
- https://medium.com/@jonathan_hui/neural-turing-machines-a-fundamental-approach-to-access-memory-in-deep-learning-b823a31fe91d