Meta Learning with Memory-Augmented Neural Networks

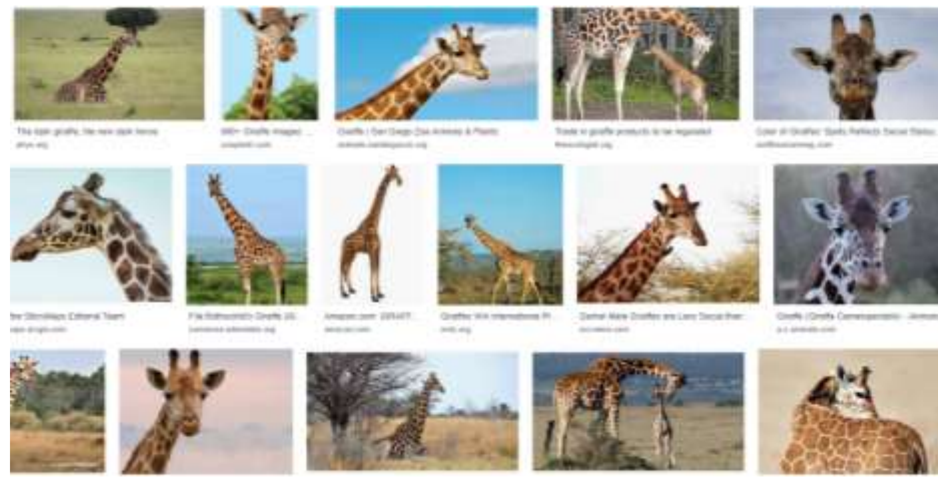
Santoro et al. (ICML 2016)

Jungsoo Park

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

Introduction

VS



Human

AI (Parametric Method)

Introduction

Building models capable of generalizing to new tasks

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{\mathcal{D}\sim p(\mathcal{D})}[\mathcal{L}_\theta(\mathcal{D})]$$

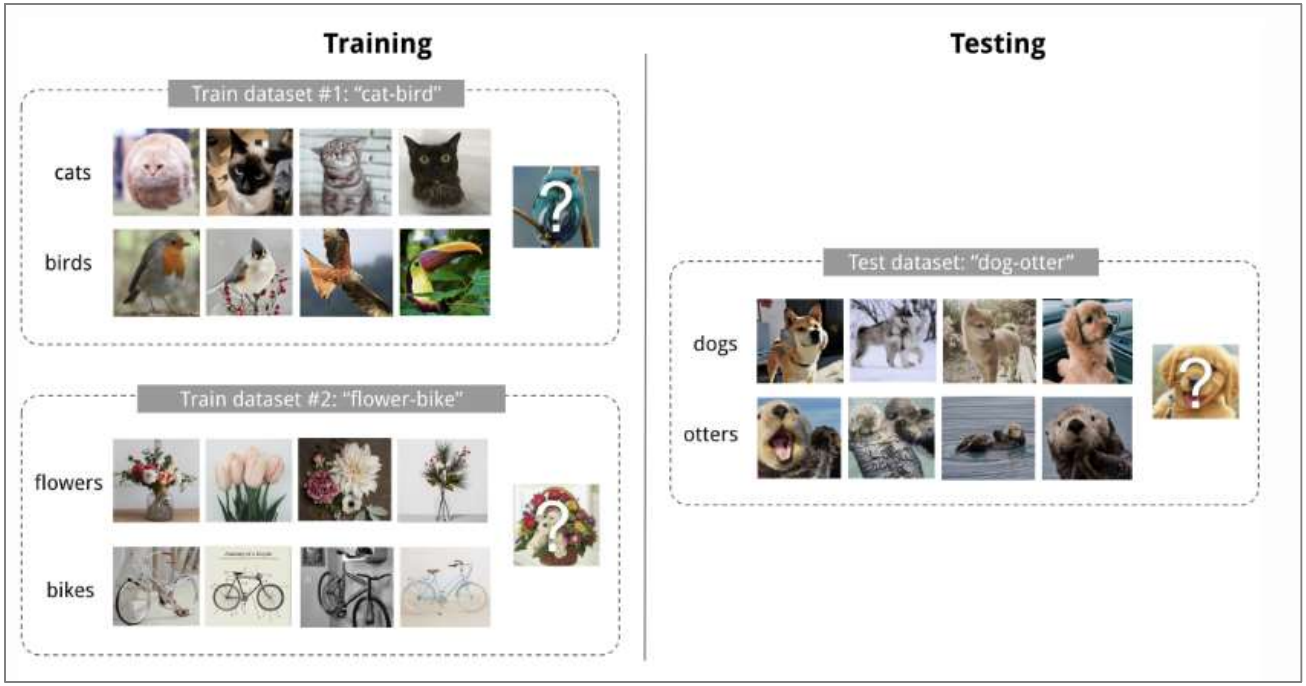


Fig1. Few-shot classification is an instantiation of meta-learning

Dataset

Omniglot

Consists of 1623 Characters from 50 different Languages written by 20 distinct people

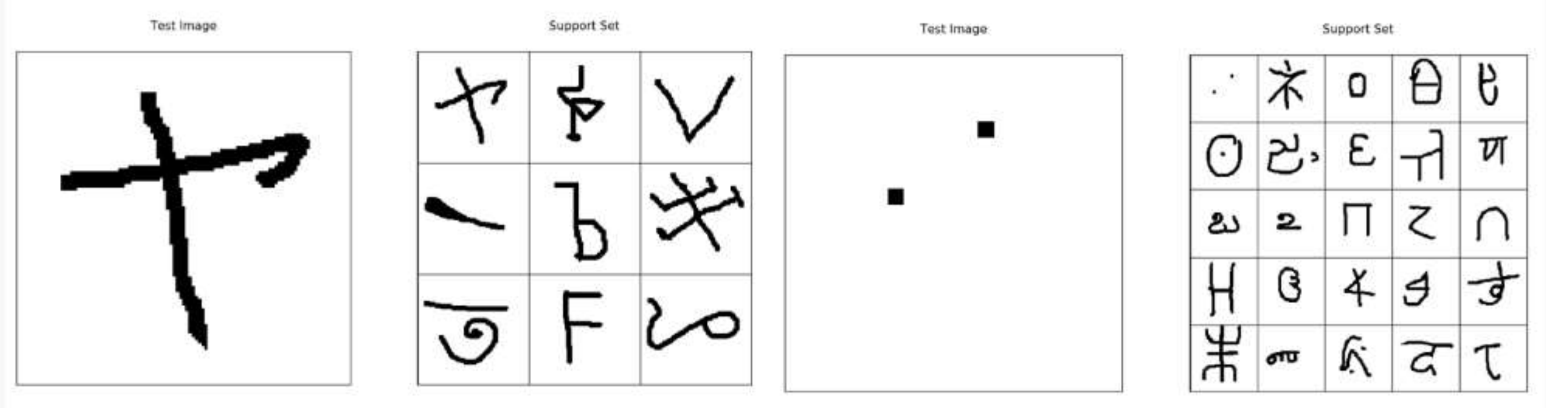
⇒ 20 samples for 1623 labels

Train : 1200 Classes Test : 423 Classes

Dataset

Omniglot

Evaluation(9-way, 25-way)



Baseline(Random Guess) 1/9, 1/25

Introduction

Training in the Same Way as Testing

Standard

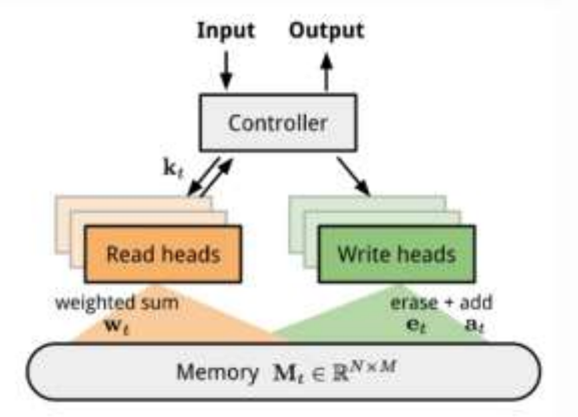
$$\theta^* = \arg\max_\theta \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}}[P_\theta(y|\mathbf{x})]$$
$$\theta^* = \arg\max_\theta \mathbb{E}_{B\subset\mathcal{D}}[\sum_{(\mathbf{x},y)\in B} P_\theta(y|\mathbf{x})]$$

Modified

$$\theta = \arg\max_\theta E_{L\subset\mathcal{L}}[E_{S^L\subset\mathcal{D},B^L\subset\mathcal{D}}[\sum_{(x,y)\in B^L} P_\theta(x,y,S^L)]]$$

Recap

Neural Turing Machine (Graves et al. 2014)



Read

$$\mathbf{r}_i = \sum_{i=1}^{N} w_t(i)\mathbf{M}_t(i), \text{ where } \sum_{i=1}^{N} w_t(i) = 1, \forall i : 0 \le w_t(i) \le 1$$

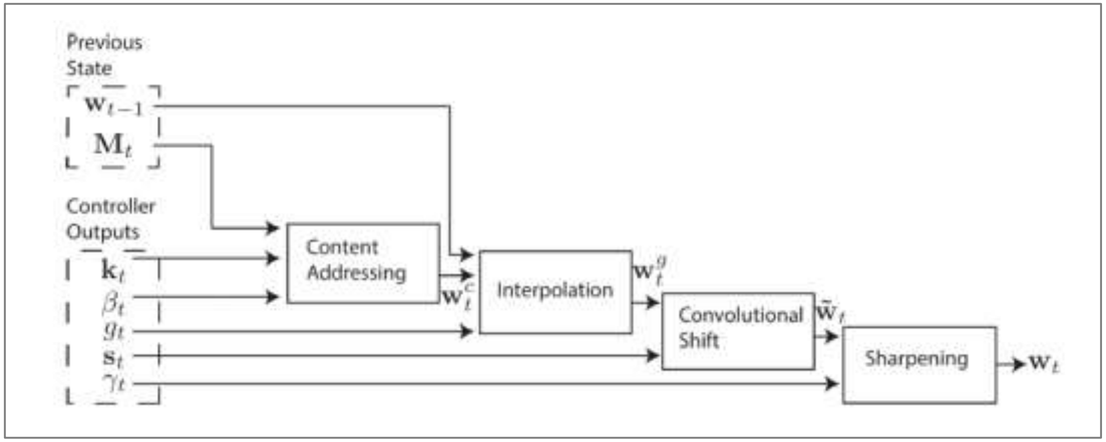
Write

$$\tilde{\mathbf{M}}_t(i) = \mathbf{M}_{t-1}(i)[\mathbf{1} - w_t(i)\mathbf{e}_t] \qquad \text{; erase}$$
$$\mathbf{M}_t(i) = \tilde{\mathbf{M}}_t(i) + w_t(i)\mathbf{a}_t \qquad \text{; add}$$

Recap

Neural Turing Machine (Graves et al. 2014)

Addressing Mechanism



Content Addressing

$$w_t^c(i) \leftarrow \frac{exp(\beta_t K[k_t, M_t(i)])}{\sum_j exp(\beta_t K[k_t, M_t(j)])}$$

Interpolation

$$w_t^g \leftarrow g_t w_t^c + (1 - g_t) w_{t-1}$$

Convolutional Shift

$$\tilde{w}_t(i) \leftarrow \sum_{j=0}^{N-1} w_t^g(j) s_t(i-j)$$

Sharpening

$$w_t(i) \leftarrow \frac{\tilde{w}_t(i)^{\gamma_t}}{\sum_j \tilde{w}_t(i)^{\gamma_t}}$$

Method

Memory-Augmented Model

- Memory encoding and retrieval in a NTM external memory is rapid, which leads to a suitable candidate for meta-learning of low-shot learning

Addressing Mechanism

- Pure content-based addressing mechanism is utilized which is suitable for the data independent of sequence.

Read

$$\mathbf{r}_i = \sum_{i=1}^{N} w_t^r(i)\mathbf{M}_t(i), \text{ where } w_t^r(i) = \text{softmax}\left(\frac{\mathbf{k}_t \cdot \mathbf{M}_t(i)}{\|\mathbf{k}_t\| \cdot \|\mathbf{M}_t(i)\|}\right)$$

Method

Least Recently Used Access

- Addressing mechanism for writing newly received information into memory operates like the cache replacement policy, in which the write heads prefer to write new contents to either the *least used* memory or the *recently used* ones.

Write

$$\mathbf{M}_t(i) \leftarrow \mathbf{M}_{t-1}(i) + w_t^w(i)\mathbf{k}_t, \forall i$$

Usage Weights

$$\mathbf{w}_t^u = \gamma \mathbf{w}_{t-1}^u + \mathbf{w}_t^r + \mathbf{w}_t^w$$

Least-Used Weights

$$w_t^{lu}(i) = \begin{cases} 0 & \text{if } w_t^u(i) > m(\mathbf{w}_t^u, n) \\ 1 & \text{if } w_t^u(i) \leq m(\mathbf{w}_t^u, n) \end{cases}$$

Write Weights Update

$$\mathbf{w}_t^w \leftarrow \sigma(\alpha)\mathbf{w}_{t-1}^r + (1 - \sigma(\alpha))\mathbf{w}_{t-1}^{lu}$$

Method

Least Recently Used Access

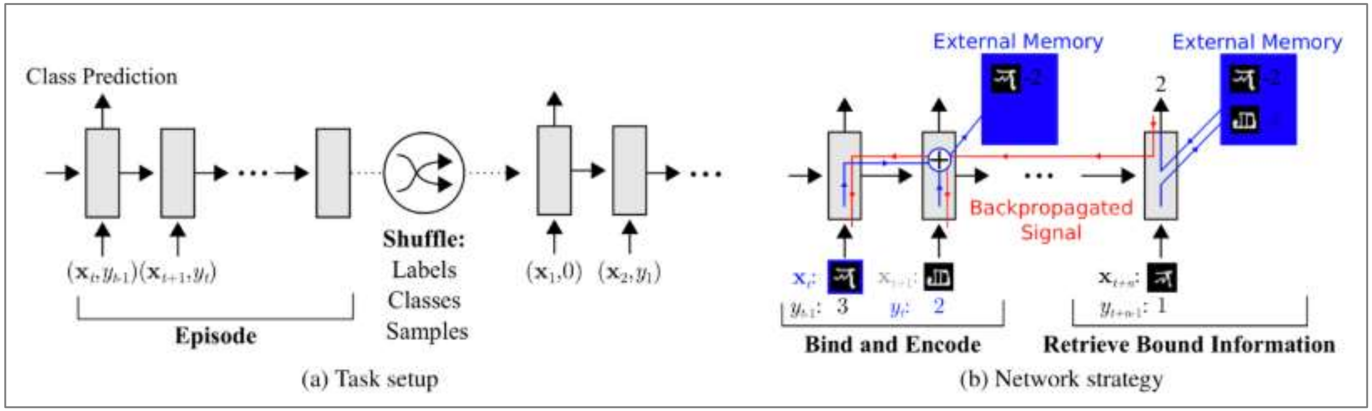
Motivation

- Rarely Used Locations : to preserve frequently used information
- Last Used Locations : update of the memory with newer, possibly more relevant information

Method

Meta-Learning Task Methodology

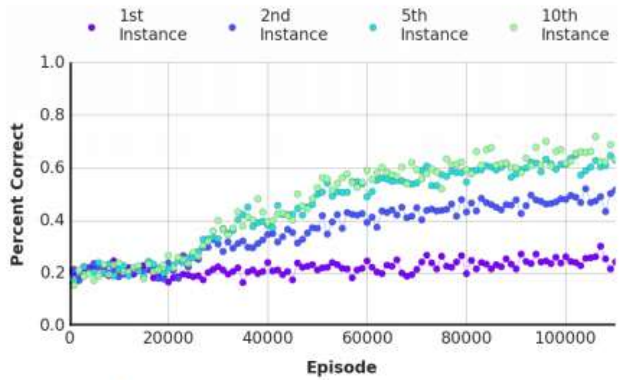
- Training in a way that the memory can encode and capture the information of new tasks fast and any stored representation is easily accessible



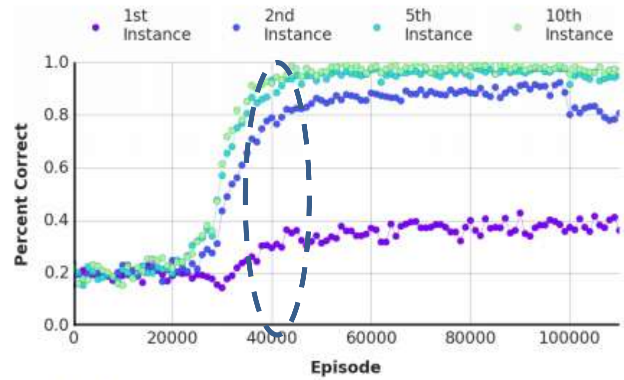
- Classes, Lables, Samples are shuffled for every episodes
- By this training setup, the MANN model must uphold the information of a newdataset until the true label is presented

Experiment

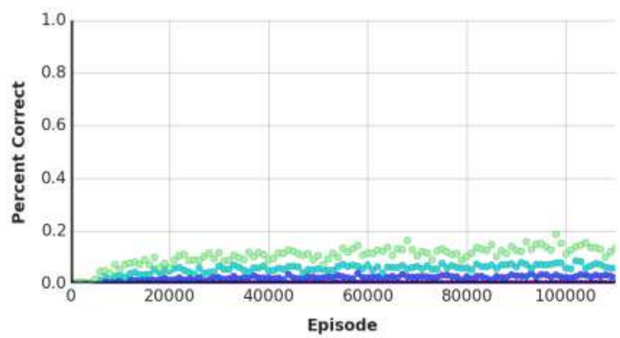
Adaptation Ability



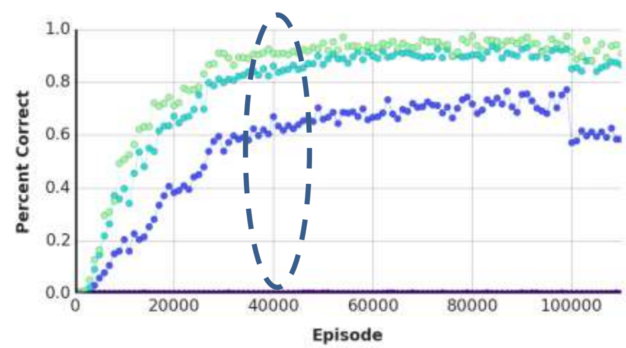
(a) LSTM, five random classes/episode, one-hot vector labels



(b) MANN, five random classes/episode, one-hot vector labels



(c) LSTM, fifteen classes/episode, five-character string labels



(d) MANN, fifteen classes/episode, five-character string labels

Experiment

Comparison with Baselines

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**

Model	Controller	# of Classes	Instance (% Correct) 1st	2nd	3rd	4th	5th	10th
kNN (raw pixels)	–	5	4.0	36.7	41.9	45.7	48.1	57.0
kNN (deep features)	–	5	4.0	51.9	61.0	66.3	69.3	77.5
Feedforward	–	5	0.0	0.2	0.0	0.2	0.0	0.0
LSTM	–	5	0.0	9.0	14.2	16.9	21.8	25.5
MANN	Feedforward	5	0.0	8.0	16.2	25.2	30.9	46.8
MANN	LSTM	5	0.0	**69.5**	**80.4**	**87.9**	**88.4**	**93.1**
kNN (raw pixels)	–	15	0.5	18.7	23.3	26.5	29.1	37.0
kNN (deep features)	–	15	0.4	32.7	41.2	47.1	50.6	60.0
Feedforward	–	15	0.0	0.1	0.0	0.0	0.0	0.0
LSTM	–	15	0.0	2.2	2.9	4.3	5.6	12.7
MANN (LRUA)	Feedforward	15	0.1	12.8	22.3	28.8	32.2	43.4
MANN (LRUA)	LSTM	15	0.1	**62.6**	**79.3**	**86.6**	**88.7**	**95.3**
MANN (NTM)	LSTM	15	0.0	35.4	61.2	71.7	77.7	88.4