Learning to Remember More with Less Memorization

Hung Le, Truyen Tran and Svetha Venkatesh

{lethai, truyen.tran, svetha.venkatesh}@deakin.edu.au

DEAKIN UNIVERSITY

Introduction

Current RAM-like memory models maintain memory accessing every timesteps, thus they do not effectively leverage the short-term memory held in the controller. We hypothesize that this scheme of writing is suboptimal in memory utilization and introduces redundant computation. To validate our hypothesis, we derive a theoretical bound on the amount of information stored in a RAM-like system and formulate an optimization problem that maximizes the bound.

Measurement of memorization

- RNN State transition: $h_t = \Phi(h_{t-1}, x_t)$
- $c_{i,t} = \left\|\frac{\partial h_t}{\partial x_i}\right\|$ measures h_t's memorization of x_i
  - E.g.: if $\left\|\frac{\partial h_t}{\partial x_i}\right\|$=0. h_t is constant w.r.t x_i , h_t contains no information on x_i (no memorization)
- To measure h_t's memorization of the whole sequence, we take average:

$$I_\lambda = \frac{\sum_{t=1}^{T} c_{i,T}}{T} = c_{T,T}\frac{\sum_{t=1}^{T}\lambda^{T-t}}{T}$$

where $\lambda \in \mathbb{R}^+$.

- Extend to MANNs:

Theorem 2:. With any D chosen writes at timesteps 1 $\leq K_1 < K_2 < \ldots < K_D < T$, there exist λ, C $\in \mathbb{R}^+$ such that the lower bound on the average contribution of a sequence of length T with respect to a MANN having D memory slots can be quantified as the following:

$$I_\lambda = C\frac{\sum_{i=1}^{D+1} f_\lambda(l_i)}{T}$$

where $l_i = \begin{cases} K_1; i = 1 \\ K_i - K_{i-1}; D \geq i > 1 \\ T - K_D; i = D+1 \end{cases}, f_\lambda(x) = \begin{cases} \frac{1-\lambda^x}{1-\lambda}, \lambda \neq 1 \\ x, \quad \lambda = 1 \end{cases}$

- Optimal solution to maximize I_λ:

$$l_1 = l_2 = \cdots = l_{D+1} = \frac{T}{D+1}$$

Uniform Writing (UW)

- Approximate the optimal solution by using discrete uniform writing rule:

$$M_t = \begin{cases} f_w(o_t, M_{t-1}) \; if \; t = k\left\lfloor\frac{T}{D+1}\right\rfloor, k \in \mathbb{N}^+ \\ M_{t-1} \end{cases}$$

where M_t, o_t, f_w, T, D are the memory, controller output, writing function, sequence length and number of memory slots, respectively.

- Assumptions: This writing policy works well if timesteps are equally important and the task is to remember all of them to produce outputs (i.e., in copy task). However, in reality, timesteps are not created equal and a good model may need to ignore unimportant or noisy timesteps.

Cached Uniform Writing (CUW)

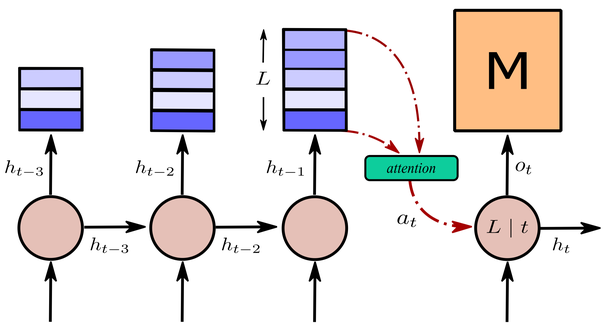

***Figure 1**: Writing mechanism in Cached Uniform Writing. During non-writing intervals, the controller hidden states are pushed into the cache. When the writing time comes, the controller attends to the cache, chooses suitable states and accesses the memory. The cache is then emptied.*

- The intervals between writes are equal with length
- After $\left\lfloor\frac{T}{L}\right\rfloor$ writes, all memory slots should be filled and the model has to learn to overwrite
- Attention discriminates timesteps

Results

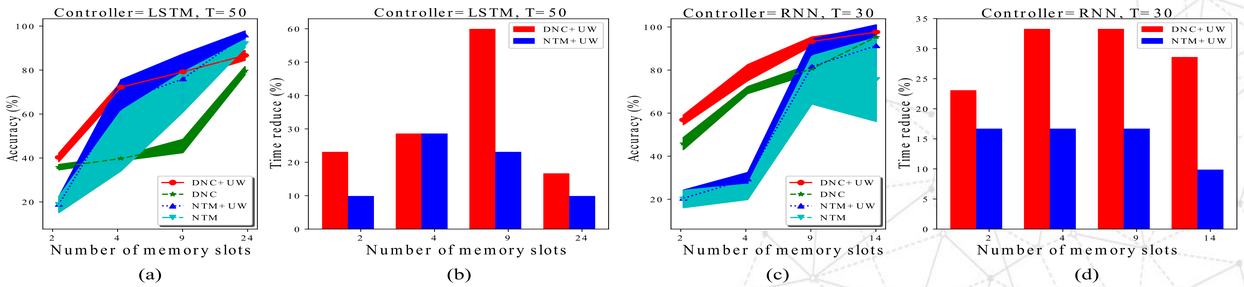

***Figure 2:** The accuracy (%) and computation time reduction (%) with different memory types and number of memory slots. The controllers/sequence lengths/memory sizes are chosen as LSTM/50/{2, 4, 9, 24} (a&b) and RNN/30/{2, 4, 9, 14} (c&d), respectively*

Model	Copy		Reverse		Add		Max	
	L=50	L=100	L=50	L=100	L=50	L=100	L=50	L=100
DNC	68.0	44.2	65.0	54.1	83.8	22.3	59.5	27.4
DNC+RW	47.6	37.0	70.8	50.1	83.0	22.7	59.7	36.5
DNC+UW	**97.7**	**69.3**	**100**	**79.5**	84.8	50.9	71.7	66.2
DNC+CUW	83.8	55.7	93.3	55.4	**94.4**	**60.1**	**82.3**	**70.7**

***Table 2:** Test accuracy (%) on synthetic tasks. MANNs have 4 memory slots*

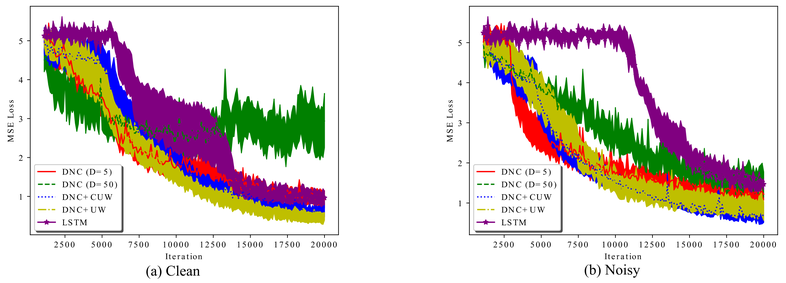

***Figure 3**: Learning curves of models in clean (a) and noisy (b) sinusoid regression experiment.*

Model	MNIST	pMNIST
r-LSTM	98.4	95.2
Dilated-GRU	**99.2**	94.6
DNC	98.1	94.0
DNC+UW	98.6	95.6
DNC+CUW	99.1	**96.3**

***Table 3:** Test accuracy (%) on MNIST, pMNIST*

Model	AG	IMDb	Yelp P.	Yelp F.	DBP	Yah. A.
D-LSTM	-	-	92.6	59.6	98.7	73.7
Skim-LSTM	93.6	91.2	-	-	-	-
Region Embedding	92.8	-	**96.4**	64.9	98.9	73.7
DNC+UW	93.7	**91.4**	**96.4**	65.3	**99.0**	74.2
DNC+CUW	**93.9**	91.3	**96.4**	**65.6**	**99.0**	**74.3**

***Table 4:** Document classification accuracy (%) on several datasets*