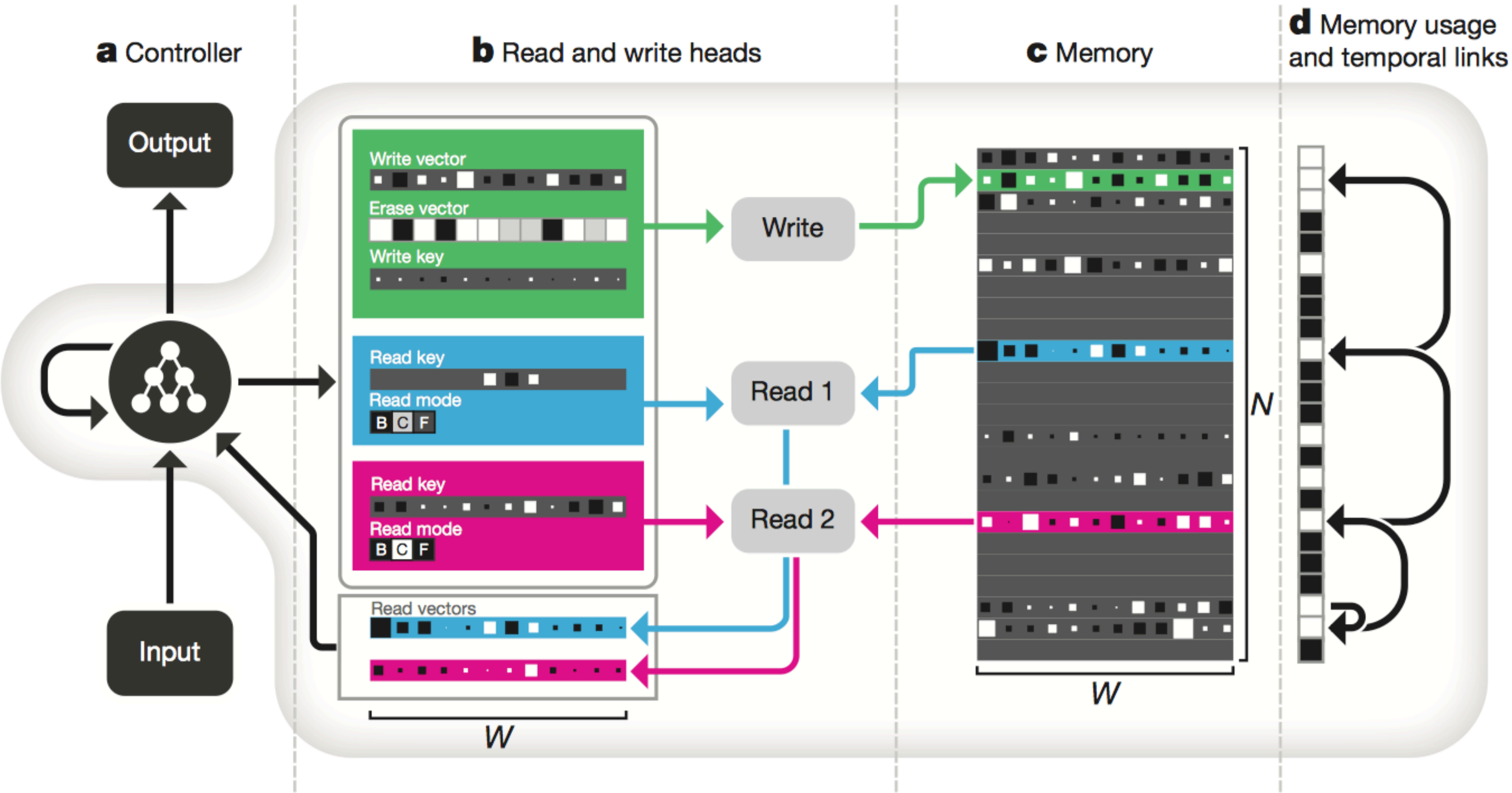
Deep SLAM Study

July 11, 2018

Dong-Won Shin

DNC architecture

- Differentiable attention mechanisms



Controller Network

- Basically, the controller network is deep LSTM architecture.

Input gate $\boldsymbol{i}_t^l = \sigma(W_{\boldsymbol{i}}^l[\chi_t; \boldsymbol{h}_{t-1}^l; \boldsymbol{h}_t^{l-1}] + \boldsymbol{b}_{\boldsymbol{i}}^l)$

Forget gate $\boldsymbol{f}_t^l = \sigma(W_{\boldsymbol{f}}^l[\chi_t; \boldsymbol{h}_{t-1}^l; \boldsymbol{h}_t^{l-1}] + \boldsymbol{b}_{\boldsymbol{f}}^l)$

State (long-term memory) $\boldsymbol{s}_t^l = \boldsymbol{f}_t^l \boldsymbol{s}_{t-1}^l + \boldsymbol{i}_t^l \tanh(W_{\boldsymbol{s}}^l[\chi_t; \boldsymbol{h}_{t-1}^l; \boldsymbol{h}_t^{l-1}] + \boldsymbol{b}_{\boldsymbol{s}}^l)$

Output gate $\boldsymbol{o}_t^l = \sigma(W_{\mathbf{o}}^l[\chi_t; \boldsymbol{h}_{t-1}^l; \boldsymbol{h}_t^{l-1}] + \boldsymbol{b}_{\mathbf{o}}^l)$

Hidden (short-term memory) $\boldsymbol{h}_t^l = \boldsymbol{o}_t^l \tanh(\boldsymbol{s}_t^l)$

Read vector $\boldsymbol{r}_{t-1}^1, \ldots, \boldsymbol{r}_{t-1}^R$ from read heads at the previous time step

Interface vector $\xi_t = W_\xi[\boldsymbol{h}_t^1; \ldots; \boldsymbol{h}_t^L]$

Output vector $\boldsymbol{y}_t = \boldsymbol{v}_t + W_r[\boldsymbol{r}_t^1; \ldots; \boldsymbol{r}_t^R]$,where $\boldsymbol{v}_t = W_y[\boldsymbol{h}_t^1; \ldots; \boldsymbol{h}_t^L]$

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Interface Vector

- Interface parameters

$$\boldsymbol{\xi}_t = \left[\boldsymbol{k}_t^{\mathrm{r},1}; \ldots ; \boldsymbol{k}_t^{\mathrm{r},R}; \hat{\beta}_t^{\mathrm{r},1}; \ldots ; \hat{\beta}_t^{\mathrm{r},R}; \boldsymbol{k}_t^{\mathrm{w}}; \hat{\beta}_t^{\mathrm{w}}; \hat{\boldsymbol{e}}_t; \boldsymbol{v}_t; \hat{f}_t^1; \ldots ; \hat{f}_t^R; \hat{g}_t^{\mathrm{a}}; \hat{g}_t^{\mathrm{w}}; \hat{\boldsymbol{\pi}}_t^1; \ldots ; \hat{\boldsymbol{\pi}}_t^R\right]$$

- R read keys $\{\boldsymbol{k}_t^{\mathrm{r},i} \in \mathbb{R}^W; 1 \leq i \leq R\}$;
- R read strengths $\{\beta_t^{\mathrm{r},i} = \text{oneplus}(\hat{\beta}_t^{\mathrm{r},i}) \in [1,\infty) ; 1 \leq i \leq R\}$;
- the write key $\boldsymbol{k}_t^{\mathrm{w}} \in \mathbb{R}^W$;
- the write strength $\beta_t^{\mathrm{w}} = \text{oneplus}(\hat{\beta}_t^{\mathrm{w}}) \in [1,\infty)$;
- the erase vector $\boldsymbol{e}_t = \sigma(\hat{\boldsymbol{e}}_t) \in [0,1]^W$;
- the write vector $\boldsymbol{v}_t \in \mathbb{R}^W$;
- R free gates $\{f_t^i = \sigma(\hat{f}_t^i) \in [0,1]; 1 \leq i \leq R\}$;
- the allocation gate $g_t^{\mathrm{a}} = \sigma(\hat{g}_t^{\mathrm{a}}) \in [0,1]$;
- the write gate $g_t^{\mathrm{w}} = \sigma(\hat{g}_t^{\mathrm{w}}) \in [0,1]$; and
- R read modes $\{\boldsymbol{\pi}_t^i = \text{softmax}(\hat{\boldsymbol{\pi}}_t^i) \in \mathcal{S}_3 ; 1 \leq i \leq R\}$.

Reading and Writing to Memory

- Read operation

$$\mathbf{r}_t^i = \underline{M_t^\top}\,\underline{\mathbf{w}_t^{r,i}}$$

memory read weight

- Write operation

$$M_t = \underline{M_{t-1} \circ (E - \mathbf{w}_t^{w} \mathbf{e}_t^\top)} + \underline{\mathbf{w}_t^{w} \mathbf{v}_t^\top}$$

erasing writing

From the interface vector

- the erase vector $\mathbf{e}_t = \sigma(\hat{\mathbf{e}}_t) \in [0,1]^W$;
- the write vector $\mathbf{v}_t \in \mathbb{R}^W$;

Memory Addressing

- To determine where to write
  - Combination of content-based addressing and dynamic memory allocation
- To determine where to read
  - Combination of content-based addressing and temporal memory linkage

- Content-based addressing

$$\mathcal{C}(M,\boldsymbol{k},\beta)[i]=\frac{\exp\{\mathcal{D}(\boldsymbol{k},M[i,\cdot])\beta\}}{\sum_j \exp\{\mathcal{D}(\boldsymbol{k},M[j,\cdot])\beta\}}$$,where $\mathcal{D}(\boldsymbol{u},\boldsymbol{v})=\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{|\boldsymbol{u}||\boldsymbol{v}|}$

Lookup key Key strength

From the interface vector

- R read keys $\{\boldsymbol{k}_t^{r,i}\in\mathbb{R}^W;1\le i\le R\}$;
- R read strengths $\{\beta_t^{r,i}=\text{oneplus}(\hat{\beta}_t^{r,i})\in[1,\infty);1\le i\le R\}$;
- the write key $\boldsymbol{k}_t^{w}\in\mathbb{R}^W$;
- the write strength $\beta_t^{w}=\text{oneplus}(\hat{\beta}_t^{w})\in[1,\infty)$;

Memory Addressing

- Dynamic memory allocation
  - To allow the controller to free and allocate memory as needed
  - Retention vector: how much each location will not be freed by the free gates

$$\psi_t = \prod_{i=1}^{R} \left(\mathbf{1} - f_t^i \mathbf{w}_{t-1}^{\mathrm{r},i}\right)$$

  - Usage vector: which locations have been used so far

$$\boldsymbol{u}_t = (\boldsymbol{u}_{t-1} + \boldsymbol{w}_{t-1}^{\mathrm{w}} - \boldsymbol{u}_{t-1} \circ \boldsymbol{w}_{t-1}^{\mathrm{w}}) \circ \boldsymbol{\psi}_t$$

  - Allocation weighting: to provide new locations for writing

$$\boldsymbol{a}_t[\phi_t[j]] = (1 - \boldsymbol{u}_t[\phi_t[j]]) \prod_{i=1}^{j-1} \boldsymbol{u}_t[\phi_t[i]]$$

  - Content weighting: from content-based addressing

$$\boldsymbol{c}_t^{\mathrm{w}} = \mathcal{C}(M_{t-1}, \boldsymbol{k}_t^{\mathrm{w}}, \beta_t^{\mathrm{w}})$$

  - Write weighting

$$\boldsymbol{w}_t^{\mathrm{w}} = g_t^{\mathrm{w}}\left[g_t^{\mathrm{a}} \boldsymbol{a}_t + (1 - g_t^{\mathrm{a}}) \boldsymbol{c}_t^{\mathrm{w}}\right]$$

From the interface vector

- R free gates $\{f_t^i = \sigma(\hat{f}_t^i) \in [0,1];\ 1 \le i \le R\}$;
- the allocation gate $g_t^{\mathrm{a}} = \sigma(\hat{g}_t^{\mathrm{a}}) \in [0,1]$;
- the write gate $g_t^{\mathrm{w}} = \sigma(\hat{g}_t^{\mathrm{w}}) \in [0,1]$; and

Memory Addressing

- Temporal memory linkage
  - To keep track of consecutively modified memory locations
  - Precedence weighting
    - $p_t[i]$: the degree to which location *i* was the last one written to

$$\boldsymbol{p}_0 = \boldsymbol{0}$$

$$\boldsymbol{p}_t = \left(1 - \sum_i \boldsymbol{w}_t^{\mathrm{w}}[i]\right)\boldsymbol{p}_{t-1} + \boldsymbol{w}_t^{\mathrm{w}}$$

  - Link matrix
    - $L_t[i,j]$: the degree to which location *i* was the location written to after location *j*

$$L_0[i,j] = 0 \quad \forall\, i,j$$

$$L_t[i,i] = 0 \quad \forall\, i$$

$$L_t[i,j] = (1 - \boldsymbol{w}_t^{\mathrm{w}}[i] - \boldsymbol{w}_t^{\mathrm{w}}[j])L_{t-1}[i,j] + \boldsymbol{w}_t^{\mathrm{w}}[i]\boldsymbol{p}_{t-1}[j]$$

  - Forward and backward weighting

$$\boldsymbol{f}_t^i = L_t \boldsymbol{w}_{t-1}^{\mathrm{r},i}$$

$$\boldsymbol{b}_t^i = L_t^\top \boldsymbol{w}_{t-1}^{\mathrm{r},i}$$

Memory Addressing

- Temporal memory linkage
  - Content weighting: from content-based addressing

$$\boldsymbol{c}_t^{r,i} = \mathcal{C}(M_t, \boldsymbol{k}_t^{r,i}, \beta_t^{r,i})$$

  - Read weighting

$$\boldsymbol{w}_t^{r,i} = \boldsymbol{\pi}_t^i[1]\boldsymbol{b}_t^i + \boldsymbol{\pi}_t^i[2]\boldsymbol{c}_t^{r,i} + \boldsymbol{\pi}_t^i[3]\boldsymbol{f}_t^i$$

From the interface vector

- R read modes $\{\boldsymbol{\pi}_t^i = \text{softmax}(\hat{\boldsymbol{\pi}}_t^i) \in \mathcal{S}_3\,;\, 1 \leq i \leq R\}$.
- R read keys $\{\boldsymbol{k}_t^{r,i} \in \mathbb{R}^W;\, 1 \leq i \leq R\}$;
- R read strengths $\{\beta_t^{r,i} = \text{oneplus}(\hat{\beta}_t^{r,i}) \in [1,\infty)\,;\, 1 \leq i \leq R\}$;
- the write key $\boldsymbol{k}_t^w \in \mathbb{R}^W$;
- the write strength $\beta_t^w = \text{oneplus}(\hat{\beta}_t^w) \in [1,\infty)$;

Thank you