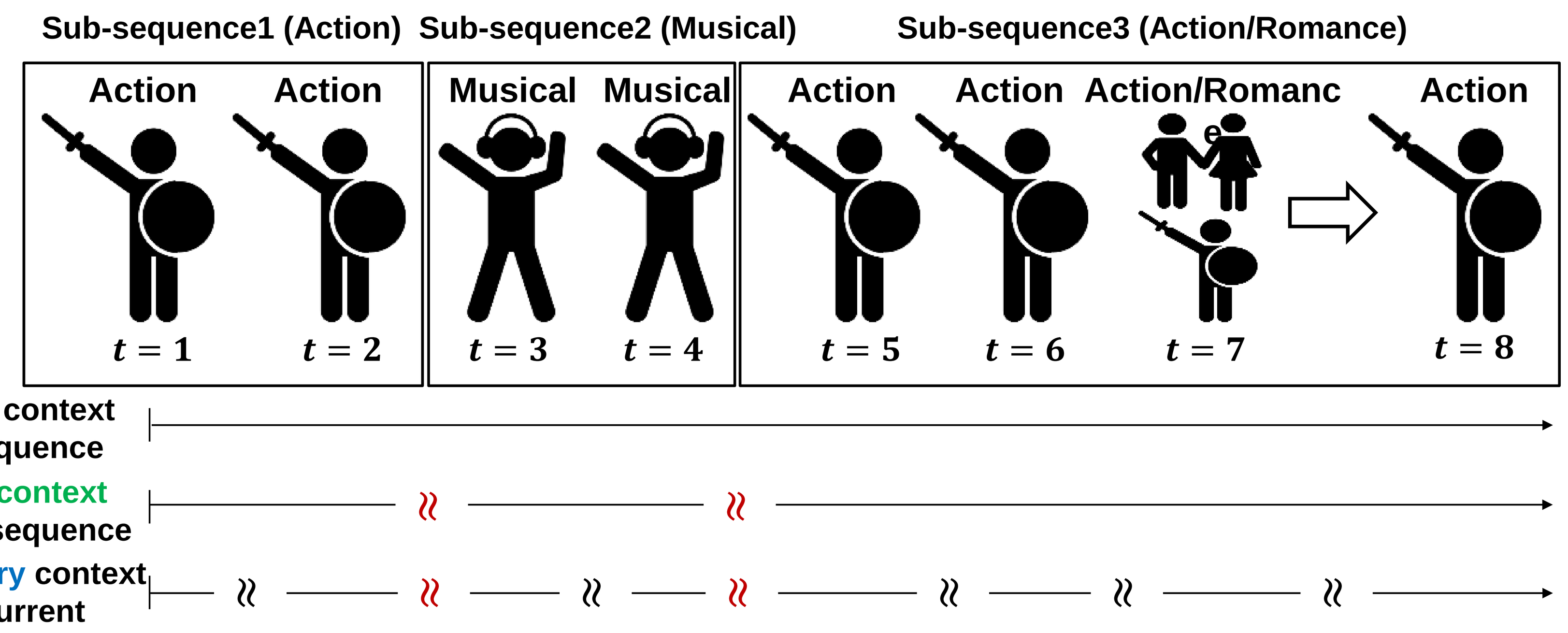
Hierarchical Context enabled Recurrent Neural Network for Recommendation

Kyungwoo Song*, Mingi Ji*, Sungrae Park, Il-Chul Moon
(* : Equal Contribution)
KAIST · APPLIED ARTIFICIAL INTELLIGENCE LAB · NAVER Clova

30-Second Summary

Question : Can we detect where the **user's interest changes**?
Our answer : Yes!
How? : **Interest drift assumption**
- "If the user's **local context** (for sub-sequence) and the current item are very different, the user's **temporary** interest drift occurs."

More specific : Hierarchical Context enabled Recurrent Neural Network (**HCRNN**)
- Incorporate the interest drift assumption
- Design **hierarchical contexts** (**global**, **local**, and **temporary**)
- Keep **local** and **temporary** contexts independently
- Introduce **interest drift gate** to capture the interest drift



Motivation

- A user history is a sequence of user orders or clicks, and the history represents the user's interest
- A long user history inevitably reflects the transitions of personal interests over time
- We can predict next item better if we include modeling on an interest drift of users.

Model Assumption

- The user's interest can be hierarchically ranging from general interest to a temporary (**global**, **local**, and **temporary**)
- Each hierarchical context have different abstract levels of information.
- **Interest drift assumption**

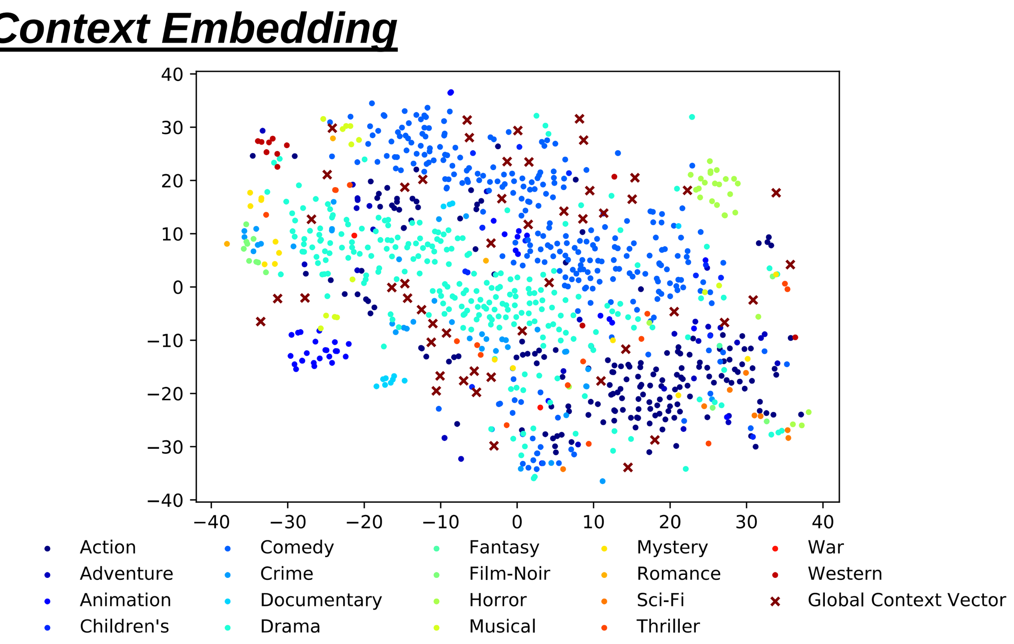
Results

1) Quantitative Results

	CiteULike				LastFM				MovieLens			
	R@3	R@20	M@3	M@20	R@3	R@20	M@3	M@20	R@3	R@20	M@3	M@20
POP	1.44	5.78	0.92	1.44	0.37	1.99	0.34	0.51	2.43	12.51	1.54	2.65
S-POP	1.26	4.99	0.79	1.23	0.87	3.65	0.55	0.87	2.27	12.23	1.42	2.52
Item-KNN	0.00	6.90	0.00	4.79	0.00	11.59	0.00	8.00	0.00	6.32	0.00	4.28
BPR-MF	0.49	3.15	0.27	0.60	0.82	2.15	0.59	0.73	1.69	8.93	1.07	1.91
LSTM4REC	7.07	23.33	4.93	6.82	15.29	24.75	12.68	13.95	8.52	32.80	5.63	8.45
GRU4REC	8.37	24.19	5.98	7.86	18.29	26.46	15.85	16.95	8.50	32.74	5.60	8.42
NARM	7.81	24.82	5.40	7.41	18.30	33.60	13.12	15.25	9.14	33.42	6.09	8.93
STAMP	5.09	21.93	3.25	5.22	9.29	19.84	6.62	8.01	3.95	20.52	2.65	4.47
HCRNN- 1	8.60	25.36	6.18	8.16	20.67*	34.40*	15.77	17.68*	9.23	33.78*	6.13	9.00
HCRNN- 2	8.83	25.10	6.41*	8.38*	20.78*	34.14*	16.20	18.08*	9.22	33.76*	6.14	9.01
HCRNN- 3	9.21*	25.42*	6.65*	8.61*	21.39*	34.72*	16.66*	18.52*	9.38*	33.67*	6.23*	9.08*
HCRNN-3 + Bi	**9.33***	**25.81***	**6.74***	**8.70***	**21.90***	**34.80***	**17.33***	**19.12***	**9.53***	**33.83***	**6.38***	**9.21***
Improvement(%)	11.47	3.99	12.71	10.69	19.67	3.57	9.34	12.80	4.27	1.23	4.76	3.14

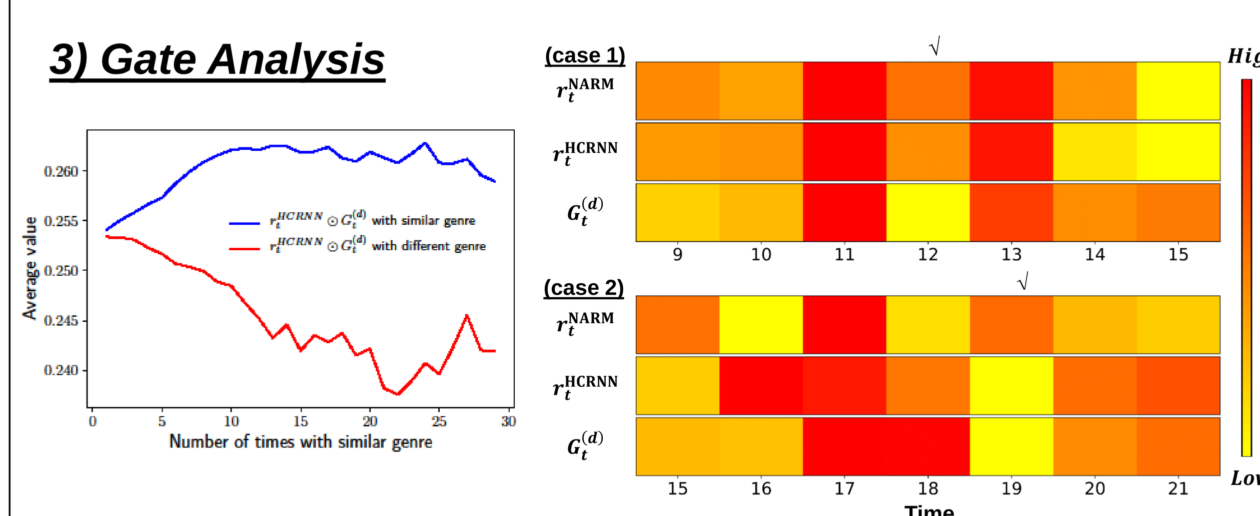
- HCRNN-1 > Baselines (NARM, STAMP)
  - necessity of hierarchical context
- HCRNN-3 > HCRNN-2, HCRNN-1
  - Interest drift assumption is experimentally justifiable.
- HCRNN-3+Bi > HCRNN-3
  - bi-channel attention with hierarchical contexts may improve the performance experimentally.

2) Context Embedding



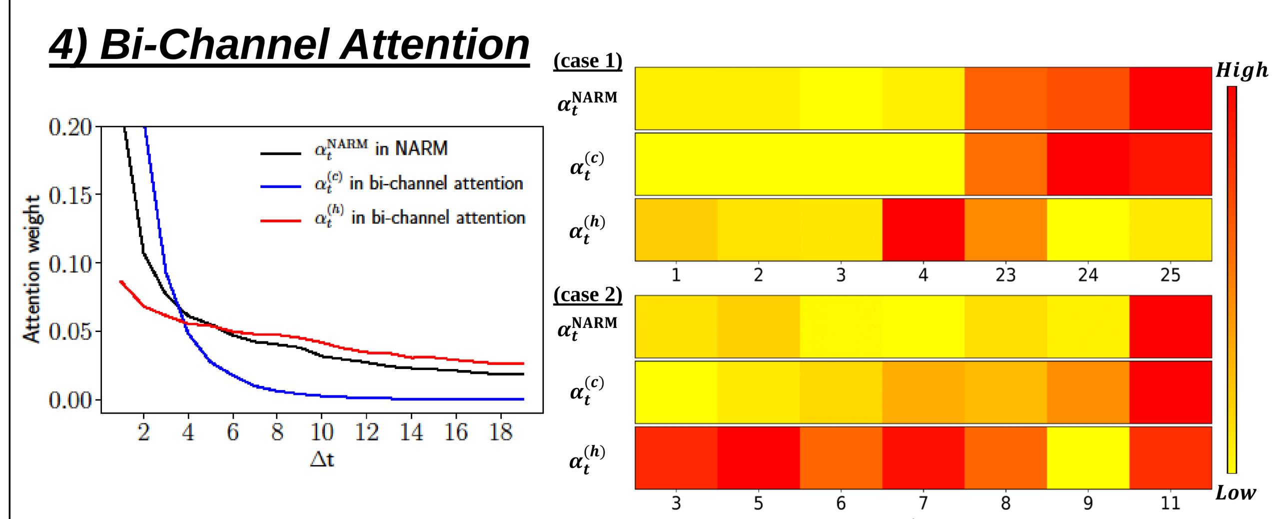
	Genre	Movie Title
$M^{(6)}_{global}$	Animation	Pinocchio, Yellow Submarine, Snow White and the Seven Dwarfs
$M^{(19)}_{global}$	Action	Star Trek: Generations, Predator, Butch Cassidy and the Sundance Kid
$M^{(31)}_{global}$	Horror	Scream, An American Werewolf in London, Dracula

3) Gate Analysis



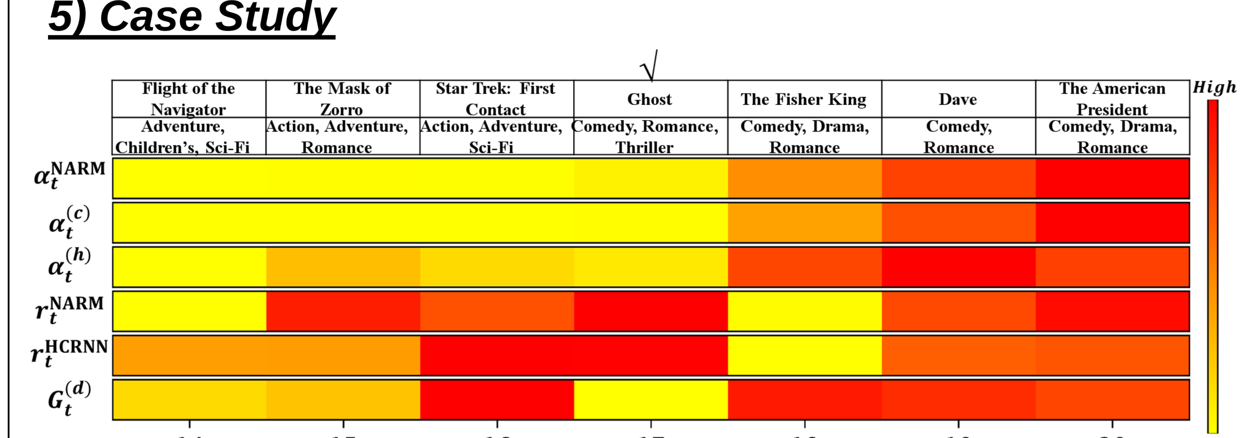
- If the genre of the current input is different with previous items, $r_t \odot G^{(d)}_t$ has a smaller value compared to the opposite situation.

4) Bi-Channel Attention



- The bi-channel attentions distinguishes the attentions
  - $\alpha^{(c)}_t$ focuses on the neighbor attention (short-term)
  - $\alpha^{(h)}_t$ reads out through the whole sequence (long-term)

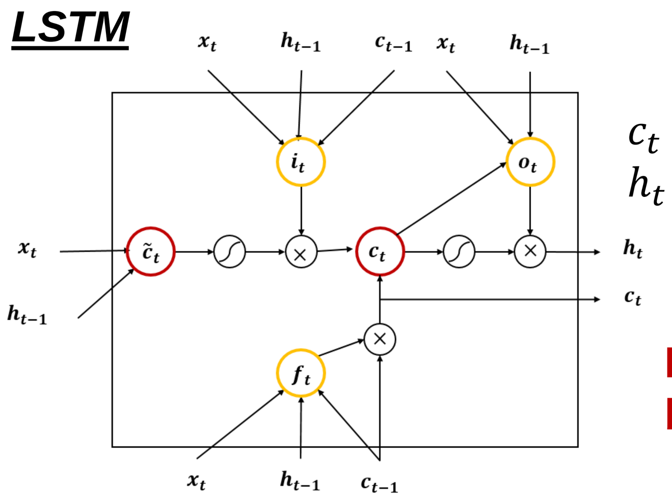
5) Case Study



- $G^{(d)}_{t=17}$ has a relatively small value
  - This small value is caused by the selection of different category items to the previous sub-sequence at t=16.

Related Works

LSTM



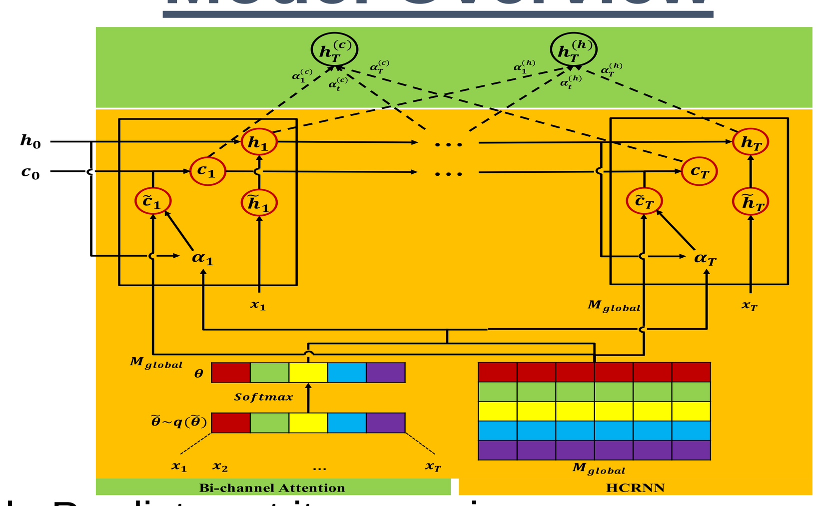
$c_t = f_t \odot c_{t-1} + i_t \odot \sigma_c(\tilde{c}_t)$
$h_t = o_t \odot \sigma_h(c_t)$

Direct connection between c_t and h_t

Sequential Recommendation
- NARM [CIKM-17] : Focus on long-term interest
- STAMP [KDD-18] : Focus on short-term interest
- **HCRNN (ours)** : Focus on interest drift with long-term and short-term interest modeling

There are no studies which capture user's interest change with hierarchical context modeling

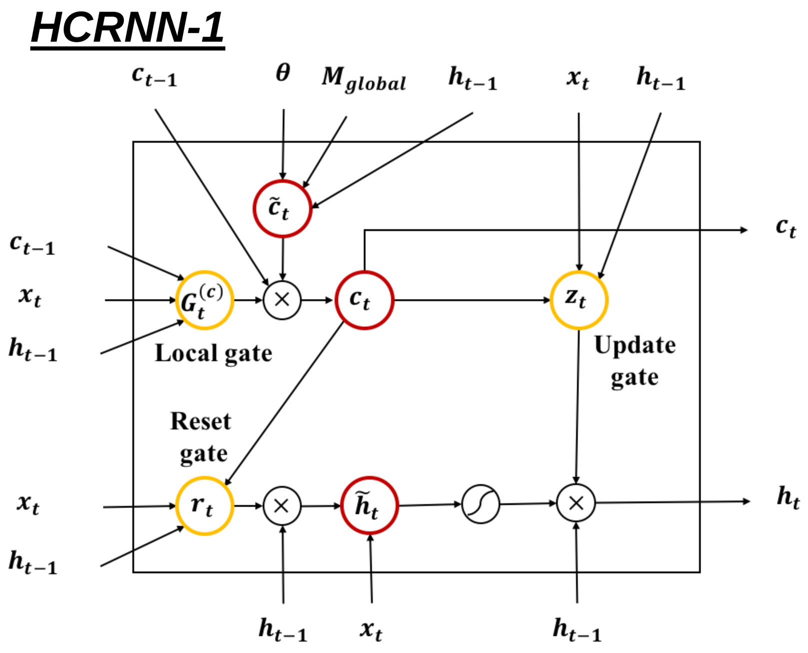
Model Overview



- Goal : Predict next item y_T given $x_{1:T}$
- x_t :Current item embedding
- θ : Global context proportion for $x_{1:T}$
- M_{global} : Global context memory
- c_t : Local context (generated by global context, **not current item**)
- h_t : Temporary context (generated by previous temporary context and current item, **not local context**)

Methodology

HCRNN-1



$$\tilde{\theta} \sim q(\tilde{\theta}) = \mathcal{N}(\tilde{\theta}; \mu(x_{1:T}), \mathrm{diag}(\sigma^2(x_{1:T}))) \quad (11)$$
$$\theta \sim \mathrm{softmax}(\tilde{\theta}) \quad (12)$$
$$\alpha^{(k)}_t = \mathrm{softmax}(v_\theta^T \sigma(h_{t-1}W_{h\alpha} + (\theta^{(k)} M^{(k)}_{global})W_{\theta\alpha})) \quad (13)$$

Proportion for sequence (Variational Encoder)
Attention weight
$\theta^{(k)} \uparrow \Rightarrow M^{k}_{global}$ is important

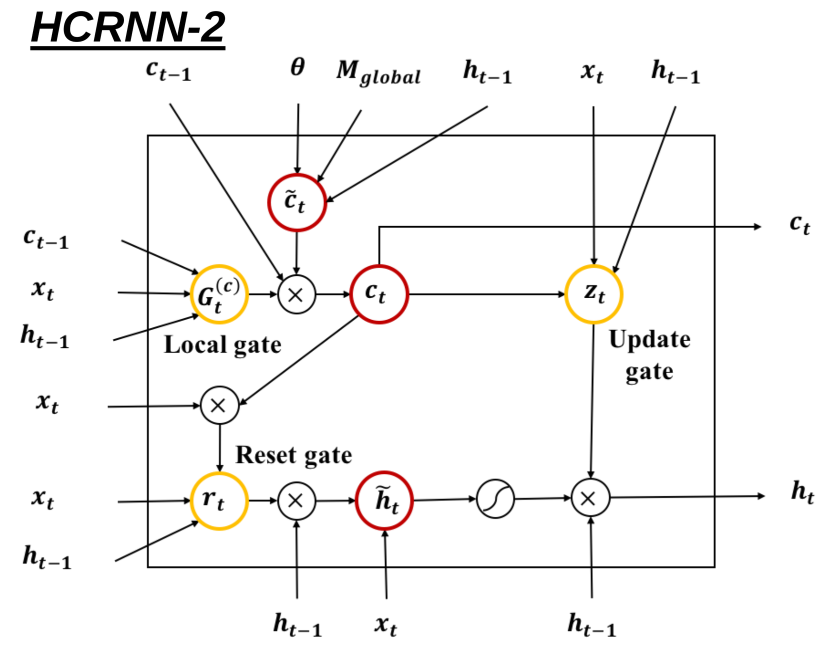
$$\tilde{c}_t = \sum_{k=1}^{K} \alpha^{(k)}_t M^{(k)}_{global} \quad (14)$$
$$G^{(c)}_t = \sigma_l(x_t W_{xl} + h_{t-1}W_{hl} + c_{t-1}W_{cl} + b_l) \quad (15)$$
$$c_t = (1 - G^{(c)}_t) \odot c_{t-1} + G^{(c)}_t \odot \tilde{c}_t \quad (16)$$

Generation of local context with local context gate $G^{(c)}_t$

$$z_t = \sigma_z(x_t W_{xz} + h_{t-1}W_{hz} + c_t W_{cz} + b_z) \quad (17)$$
$$\boxed{r_t = \sigma_r(x_t W_{xr} + h_{t-1}W_{hr} + c_t W_{cr} + b_r) \quad (18)}$$
$$\tilde{h}_t = (r_t \odot h_{t-1})W_{hh} + x_t W_{xh} + b_h \quad (19)$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \sigma_h(\tilde{h}_t) \quad (20)$$

Generation of temporary context h_t (separation with local context)

HCRNN-2

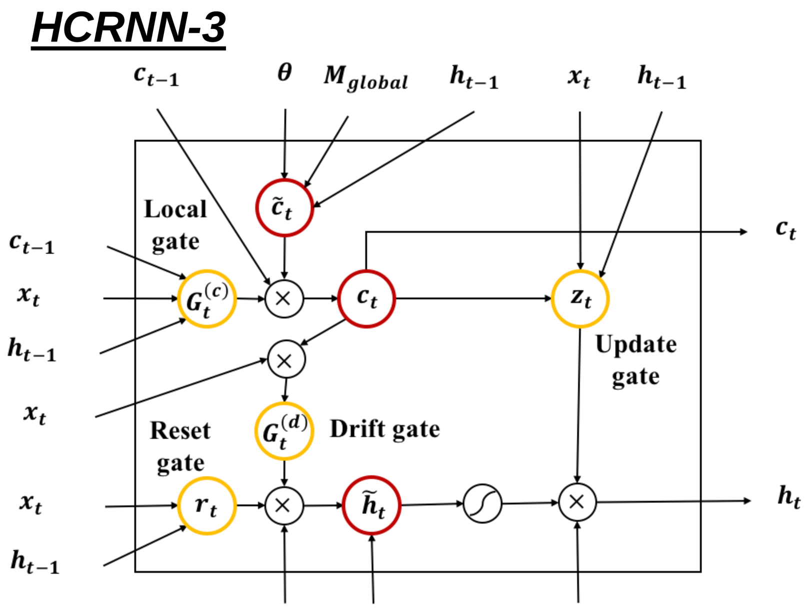


$$\boxed{r_t = \sigma_r(x_t W_{xr} + h_{t-1}W_{hr} + c_t W_{cr} + b_r) \quad (18)}$$
$$\tilde{h}_t = (r_t \odot h_{t-1})W_{hh} + x_t W_{xh} + b_h \quad (19)$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \sigma_h(\tilde{h}_t) \quad (20)$$

$$\boxed{r_t = \sigma_r(x_t W_{xr} + h_{t-1}W_{hr} + (x_t \odot c_t)W_d + b_r) \quad s.t. W_d \geq 0 \quad (21)}$$

Interest drift assumption
$x_t \odot c_t \downarrow \implies r_t \downarrow \implies h_t$ focus on the current input instead of h_{t-1}

HCRNN-3



$$r_t = \sigma_r(x_t W_{xr} + h_{t-1}W_{hr} + c_t W_{cr} + b_r) \quad (18)$$
$$\tilde{h}_t = (r_t \odot h_{t-1})W_{hh} + x_t W_{xh} + b_h \quad (19)$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \sigma_h(\tilde{h}_t) \quad (20)$$

$$\boxed{r_t = \sigma_r(x_t W_{xr} + h_{t-1}W_{hr} + (x_t \odot c_t)W_d + b_r) \quad s.t. W_d \geq 0 \quad (21)}$$

Sigmoid function is not sharp $\Rightarrow r_t$ in Eq. 21 : 0.47 (± 0.03)
It is hard to incorporate the interest drift

$$G^{(d)}_t = \sigma_d((x_t \odot c_t)W_d + b_d) \quad s.t. W_d \geq 0 \quad (22)$$
$$r_t = \sigma_r(x_t W_{xr} + h_{t-1}W_{hr} + b_r) \quad (23)$$
$$\tilde{h}_t = (r_t \odot (G^{(d)}_t \odot h_{t-1}))W_{hh} + x_t W_{xh} + b_h \quad (24)$$

Introduce the interest drift gate ($G^{(d)}_t$) to make h_t focus on the current input

HCRNN-3+Bi

$$\alpha^{(c)}_{tj} = \mathrm{softmax}\left(\frac{(c_t W^{(1)}_{c\alpha})(c_j W^{(2)}_{c\alpha})^T}{\sqrt{|H|}}\right) \quad (25)$$
$$\alpha^{(h)}_{tj} = \mathrm{softmax}(v_h^T \sigma(h_t W^{(1)}_{h\alpha} + h_j W^{(2)}_{h\alpha})) \quad (26)$$
$$h^{(c)}_t = \sum_j \alpha^{(c)}_{tj} h_j \quad \text{and} \quad h^{(h)}_t = \sum_j \alpha^{(h)}_{tj} h_j \quad (27)$$
$$\hat{y}_t = \mathrm{softmax}(W^T_{emb} W_B [h_t, h^{(c)}_t, h^{(h)}_t]) \quad (28)$$

$\alpha^{(c)}_t$: attention based on the local context (Short-term dependency)

$\alpha^{(h)}_t$: attention based on the temporary context (Long-term dependency)

Inference

Variational inference by optimizing the evidence lower bound (ELBO)

$$\log p(y_{1:T}|c_{1:T}, h_{1:T}) = \log \int p(\tilde{\theta}) \prod_{t=1} p(y_t|\tilde{\theta}, c_t, h_t) d\tilde{\theta}$$
$$\geq \sum_{t=1}^{T} E_{q(\tilde{\theta})}[\log p(y_t|\tilde{\theta}, c_t, h_t)] - \mathrm{KL}[(q(\tilde{\theta})||p(\tilde{\theta}))]$$