1) 1) 1)

1)(

. (21676295). 2018 ,"
"(2462018QZDX02) .
1966
.E-mail: liujw@cup.edu.cn. 1995 .E-mail: 714244712 @qq.com.
1963 E-mail: luoxl@cup.edu.cn.

network structure and training algorithm. Afterwords, we introduce the extended model of the memory network and its application in different fields and scenarios. Finally, the future research direction of the memory network is prospected.

Key words recurrent neural network; long short term memory network; memory network; neural turing machine; natural language processing

1

20 80 Recurrent Neural Network, RNN RNN

RNN

RNN

Hochreiter S Schmidhuber J 1997 Long Short-Term Memory LSTM [1] RNN

20 80

LSTM

RNN

LSTM RNN

2014 Weston J, Chopra S, Bordes A Memory Network MN [2]

Google DeepMind

Neural Turing Machine NTM [3]

Sukhbaatar S, Weston J, Fergus R

[4]

2017 Google

Transformer [5]

RNN

2

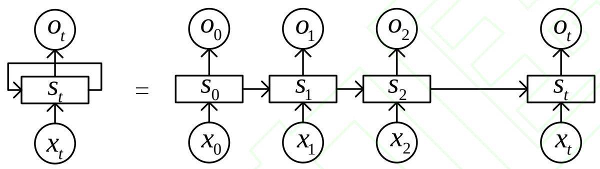
2014

[1] [3]

[2] [5][4]

2.1

1 RNN



1 RNN

RNN

x_t t x_1

s_t t

s_t

$$s_t \quad f(Ux_t \quad Ws_{t\ 1}) \tag{1}$$

3

f Ux_t $Ws_{t\ 1}$

tanh ReLU

$s\ _1$

$$o_t \quad soft\max(Vs_t) \tag{2}$$

o_t t

o_t t

V V

RNN

o_t t

s_t s_t

t

t s_t

s_t

RNN

U,V,W

RNN

RNN

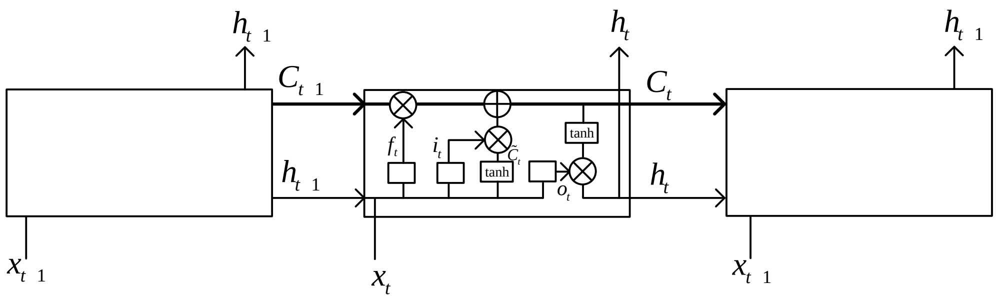
2.2

LSTM RNN

RNN

LSTM RNN LSTM

2



2 LSTM

LSTM 2 tanh

$\tilde{C}_t$

,

$i_t \quad (W_{t\ 1} \quad [h_{t\ 1}, x_t] \quad b_i)$ 4

LSTM

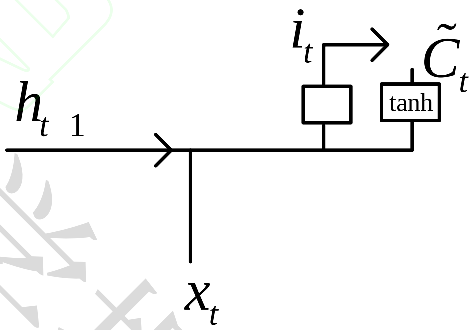
$\tilde{C}_t \quad \tanh(W_c \quad [h_{t\ 1}, x_t] \quad b_c)$ 5

sigmoid

Sigmoid 0 1

0

1 LSTM



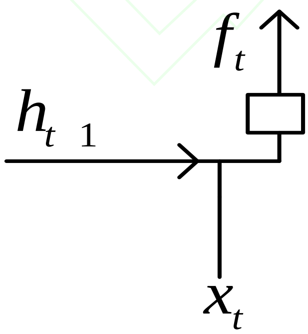
4 LSTM

LSTM

sigmoid

3 $h_{t\ 1}$ x_t f_t

$f_t \quad (W_{t\ 1} \quad [h_{t\ 1}, x_t]) \quad b_f$ 3



3 LSTM

4 sigmoid

$C_{t\ 1}$ 0 1

C_t 5 $C_{t\ 1}$

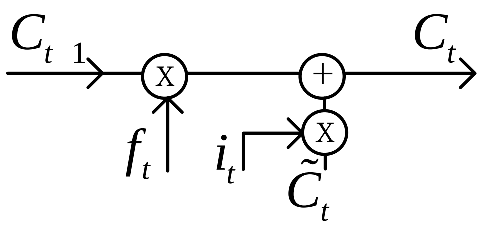
i_t f_t $C_{t\ 1}$ f_t

i_t $\tilde{C}_t$

f_t $C_{t\ 1}$ i_t $\tilde{C}_t$

C_t

$C_t \quad f_t \quad C_{t\ 1} \quad i_t \quad \tilde{C}_t$ 6



5 LSTM

C_t 6 sigmoid

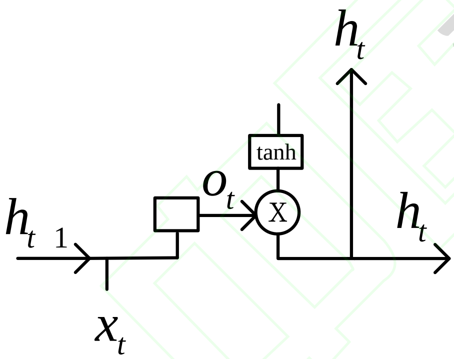
o_t C_t

tanh C_t -1 1

o_t h_t

$$o_t \quad (W_o \ [h_{t\ 1}, x_t] \ \ b_o) \qquad (7)$$

$$h_t \quad o_t \quad \tanh(C_t) \qquad (8)$$



6 LSTM

LSTM c_t

h_t t

$$i_t \quad f_t \quad o_t \quad \hat{c}_t \ \ \tanh \quad W \quad h_{t\ 1}, x_t \qquad (9)$$

$$c_t \quad f_t \quad c_{t\ 1} \quad i_t \quad \hat{c}_t \qquad (10)$$

$$h_t \quad o_t \quad \tanh(c_t) \qquad (11)$$

Greff K [6] LSTM

LSTM

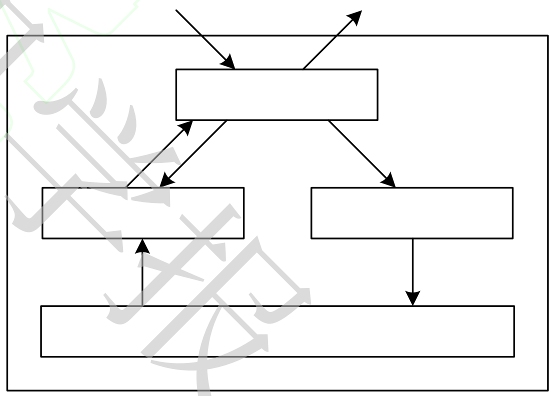
LSTM

2.3

2014 Google DeepMind [3]

1936 ——

NTM 7



7

NTM

NTM

1

M_t Î ¡ $^{N' M}$ t

N M

w_t Î ¡ ^{N} t read head

N w_t

$\sum_i w_t(i) = 1, 0 \le w_t(i) \le 1$ w_t N i t

$w_t(i)$ M N

M r_t $M_t(i)$

$M_t(i)$

$$r_t \leftarrow \sum_i w_t(i) M_t(i) \quad (12)$$

r_t w_t

2

LSTM

write head t w_t

e_t w_t e_t

M (0,1)

M_{t-1}

$$\tilde{M}_t(i) \leftarrow M_{t-1}(i)[1 - w_t(i) e_t] \quad (13)$$

1 1 $M_{t-1}(i)$

$[1 - w_t(i) e_t]$

i

$w_t(i)$ e_t 1

0 $w_t(i)$

e_t 0

M a_t

$$M_t(i) \leftarrow \tilde{M}_t(i) + w_t(i) a_t \quad (14)$$

3

w_t NTM

8

$M_t(i)$ k_t

t k_t

M_t $M_t(i)$

w_t^c

$$w_t^c(i) \leftarrow \frac{\exp\left(b_t K[k_t, M_t(i)]\right)}{\sum_j \exp\left(b_t K[k_t, M_t(j)]\right)} \quad (15)$$

K

$$K[u, v] = \frac{u \cdot v}{\|u\| \cdot \|v\|} \quad (16)$$

b_t

k_t

M_t $M_t(i)$

0,1

g_t w_{t-1}

w_t^c

w_t^g

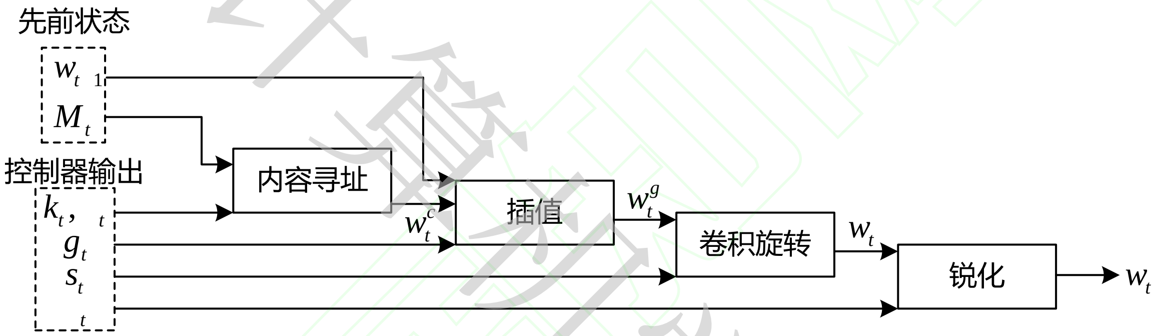
$$w_t^g \ ?\ g_t w_t^c \quad (1- g_t) w_{t-1} \qquad 17$$

N 0 N-1 s_t

w_t^g

$\tilde{w}_t(i)$

$$\tilde{w}_t(i) \ ?\ \sum_{j=0}^{N-1} w_t^g(j) s_t(i \quad j) \qquad 18$$



8 NTM

4

NTM LSTM

LSTM

M_t

sharpening

g_t

$$w_t(i) \leftarrow \frac{\tilde{w}_t(i)^{g_t}}{\sum_j \tilde{w}_t(i)^{g_t}} \qquad 19$$

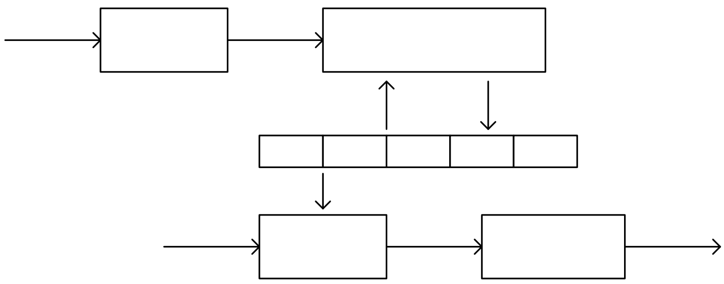
2.4

MN

Question Answers, QA

MN m I G

O R 9



9

1 x I

$I(x)$ $I(x)$ m

2 x

N k

k=1 N

$$o_1 \quad O_1(x,m) \quad \arg\max_{i\ 1,\dots,N} s_O(x,m_i) \qquad (20)$$

s_O o_1

m_{o_1}

k=2 x m_{o_1}

$$o_2 \quad O_2(x,m) \quad \arg\max_{i\ 1,\dots,N} s_O([x,m_{o_1}],m_i) \qquad (21)$$

o_2 m_{o_2}

$x, m_{o_1}, m_{o_2}, \cdots, m_{o_k}$

3 m_{o_k} r

r

$$r \quad \arg\max_{w\ W} s_R([x,m_{o_1},m_{o_2}],w) \qquad (22)$$

w W W s_R s_O

$$s(x,y) \quad {}_x(x)^T U^T U \quad {}_y(y) \qquad (23)$$

U n D D

n x y

D

MN

margin ranking loss

$$\sum_{\bar f\ m_{o_1}} \max(0, \quad s_O(x,m_{o_1}) \quad s_O(x,\bar f))$$

$$\sum_{\bar f\ m_{o_2}} \max(0, \quad s_O([x,m_{o_1}],m_{o_2}) \quad s_O([x,m_{o_1}],\bar f\,))$$

$$\sum_{\bar r\ r} \max(0, \quad s_R([x,m_{o_1},m_{o_2}],r) \quad s_R([x,m_{o_1},m_{o_2}],\bar r)) \qquad (24)$$

$\bar f, \bar f, \bar r$

0

MN Fader

QA [7] 14M

MN

QA WikiAnswers

Bordes [8]

F1

MN

k=1

Fader 0

20 10

100

Bordes

20 200

13 6

1 MN

Bordes

MN

QA

1 MN QA

	F1
Fader [7]	0.54
Bordes [8]	0.73
MN	0.72
MN	0.82

QA :

1 5

RNN LSTM

RNN LSTM

[9]

MN

100 0.01 0.1

10 2

2 MN QA

q

k v

$\sqrt{d_k}$ d_k

softmax

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (25)$$

q k v

-

q k

v

h

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1,\text{K},\text{head}_h)W^O$$
$$\text{head}_i=\text{Attention}(QW_i^Q,KW_i^K,VW_i^V) \quad (26)$$

W^O,W_i^Q,W_i^K,W_i^V

Position Embedding

RNN

$$PE_{(pos,2i)}=\sin(pos/10000^{2i/d_{model}})$$
$$PE_{(pos,2i+1)}=\cos(pos/10000^{2i/d_{model}}) \quad (27)$$

pos i

d_{model}

$pos+k$ pos

$$\sin(pos+k)=\sin(pos)\cos(k)+\cos(pos)\sin(k)$$
$$\cos(pos+k)=\cos(pos)\cos(k)-\sin(pos)\sin(k) \quad (28)$$

WMT 2014 - WMT 2014 -

Adam [10]

b_1=0.9 b_2=0.98 e=10^{-9}

$$l_{rate}=d_{\text{model}}^{-0.5}\ ?\min(step\_num^{-0.5},\ step\_num \times warmup\_steps^{-1.5}) \quad (29)$$

warmup_steps

warmup_steps 4000

Dropout [11]

$P_{drop}=0.1$

[12] $e_{ls}=0.1$

5

ByteNet WMT 2014 -

1.2

Dilated Convolution

1

16

Deep-Att+PosUnk WMT 2014 -

256 LSTM

512

GNMT+RL WMT 2014 -

WMT 2014 -

8 8

1024 LSTM

1024 2

MoE MoE 2048

ConvS2S WMT 2014 - 200

WMT 2014 - BLEU

512 3

512

0.99

0.1 0.25

MoE WMT 2014 - WMT 4

2014 -

GNMT

LSTM 3

3

	BLEU			
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet[13]	23.75			
Deep-Att + PosUnk[14]		39.2		1.0 10^{20}
GNMT + RL[15]	24.6	39.92	2.3 10^{19}	1.4 10^{20}
ConvS2S[16]	25.16	40.46	9.6 10^{18}	1.5 10^{20}
MoE[17]	26.03	40.56	2.0 10^{19}	1.2 10^{20}
Deep-Att + PosUnk Ensemble[14]		40.4		8.0 10^{20}
GNMT + RL Ensemble[15]	26.30	41.16	1.8 10^{20}	1.1 10^{21}
ConvS2S Ensemble[16]	26.36	**41.29**	7.7 10^{19}	1.2 10^{21}
Transformer (base model)	27.3	38.1	**3.3 10^{18}**	
Transformer (big)	**28.4**	**41.0**	2.3 10^{19}	

4

DTN[18]	2017	
WTN[19]	2017	
AAN[20]	2018	
BlendCNN[21]	2018	CNN
Action Transformer[22]	2018	
Universal Transformers[23]	2018	
Evolved Transformer[24]	2019	
Set Transformer[25]	2019	
Transformer-XL[26]	2019	

2.6

RNN LSTM NTM

MN

5

1 RNN

2 LSTM RNN

3 NTM RNN

4 MN NTM

NTM

MN

5

RNN	1986		
LSTM	1997		
NTM	2014		
MN	2014		
Transformer	2017		GPU

3

RNN LSTM NTM MN

RNN

LSTM MN

3.1 RNN

RNN

RNN RNN

RNN RNN

3.1.1 RNN

RNN

5 RNN

-

-

MN

GPU MN

RNN

Attention-based Memory Selection Recurrent Network AMSRN [49]

AMSRN LSTM

LSTM

1 LSTM LSTM

$\{x_1, x_2, \cdots, x_t, \cdots\}$ x_t N 1-of-N

LSTM d

h_t R^d t LSTM

$M_t \quad [h_0, h_1, \cdots, h_{t\ 1}]$

2

h_t d w_{h1} w_{h2} LSTM

$M_t \quad [h_0, h_1, \cdots, h_{t\ 1}]$ h_t

d k_t

$$k_t \quad W_{kh} h_t \quad b_k \qquad 30$$

W_{kh} $\square^{d\ d}$ b_k $\square^{d}$

k_t LSTM

$M_t \quad [h_0, h_1, \cdots, h_{t\ 1}]$ h_i e_{ti}

$$e_{ti} \quad (h_i \circ w_{h1}) \ k_t \qquad 31$$

$\circ$

$h_i \circ w_{h1}$ LSTM $M_t \quad [h_0, h_1, \cdots, h_{t\ 1}]$

h_t w_{h1}

$h_i \circ w_{h1}$ k_t

softmax e_{ti}

ti

$$ti \quad \frac{\exp(e_{ti})}{\sum_{i\ 0}^{t\ 1} \exp(e_{ti})} \qquad 32$$

w_{h2} h_i h_i

72 h_i r_t

r_t ti h_i

$$h_i \quad h_i \circ w_{h2} \qquad 33$$

$$r_t \quad \sum_{i\ 0}^{t\ 1} {}_{ti} h_i \qquad 34$$

r_t h_t

$$P_w \quad \mathrm{softmax}(W_{ph} h_t \quad W_{pr} r_t \quad b_p) \qquad 35$$

W_{ph}, W_{pr}, b_p

LSTM+	134.09	93.74	102.04
LSTM+ +	133.36	92.49	86.85
RMN	123.32	64.41	121.28
RMR	134.30	71.04	145.24

7

LSTM

LSTM

Penn Treebank

Switchboard

Gigaword

RMN Recurrent Memory Network RMR Recurrent Memory Recurrent [30] AMSRN

RMN RMR

RMN RMR

AMSRN

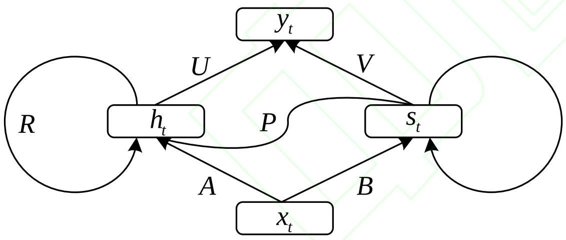
3.1.2 RNN

Mikolov T [31]

RNN

LSTM

11



11 RNN

h_t

$h_{t\ 1}$ x_t

$$h_t \quad (Ax_t \quad Rh_{t\ 1}) \qquad 36$$

RNN

s_t

$$s_t \quad (1 \quad)Bx_t \quad s_{t\ 1}$$
$$h_t \quad (Ps_t \quad Ax_t \quad Rh_{t\ 1}) \qquad 37$$
$$y_t \quad f(Uh_t \quad Vs_t)$$

$A \quad R \quad B \quad P \quad U \quad V$

0 1

0 1

Q t k

$s_k, k \quad t \quad 1,\cdots,t \quad K$

Penn Treebank Corpus Text8

LSTM RNN

LSTM

3.1.3 RNN

[32]

1

L θ

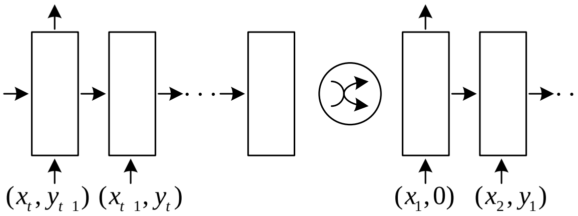
12

$$L(D; \quad) \qquad D \quad x_i, y_i \ {}_{i\ 1}^{t} \qquad p(D)$$

$$^{*} \quad \arg\min \quad E_{D \sim p(D)}[L(D; \quad)] \qquad 38$$

t

$y_{t\ 1}$ x_t



12

2

LSTM

, Least Recently Used Access LRUA

LRUA

w_t^u

$w_{t\ 1}^u$ w_t^r w_t^w

w_t^u

$$w_t^u \quad w_{t\ 1}^u \quad w_t^r \quad w_t^w \qquad (39)$$

w_t^r

k_t M_t

$$K(k_t, M_t(i)) \quad \frac{k_t \ \ M_t(i)}{\|k_t\|\|M_t(i)\|} \qquad (40)$$

$$w_t^r(i) \quad \frac{\exp(K(k_t, M_t(i)))}{{}_j\exp(K(k_t, M_t(i)))} \qquad (41)$$

sigmoid $w_{t\ 1}^r$

$w_{t\ 1}^{lu}$

$$w_t^w \quad (\ \ ) w_{t\ 1}^r \quad (1 \quad (\ \ )) w_{t\ 1}^{lu} \qquad (42)$$

$w_{t\ 1}^{lu}$ $w_t^u(i)$

w_t^w 1 0

k_t

$$M_t(i) \quad M_{t\ 1}(i) \quad w_t^w(i) k_t \qquad (43)$$

RNN

RNN Wang J RNN [33]

[34] RNN

g $\|g\|$

$\|g\| > v$

$$g \leftarrow \frac{gv}{\|g\|} \quad (44)$$

v g

$\nabla_{h^{(t)}} L$

Pascanu R [35]

$$\Omega = \sum_t \left(\frac{\left\| (\nabla_{h^{(t)}} L) \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \right\|}{\|\nabla_{h^{(t)}} L\|} - 1 \right)^2 \quad (45)$$

RNN

3.2 LSTM

LSTM

LSTM

LSTM

LSTM LSTM

LSTM LSTM

3.2.1 LSTM

LSTM

LSTM

N_h LSTM

$O(N_h^2)$ LSTM

Danihelka I

Associative Long Short-Term Memory

ALSTM [36] LSTM Holographic

Reduced Representations HRR -

HRR -

$r \quad (a_r[1]e^{i\,r[1]}, a_r[2]e^{i\,r[2]}, \cdots)$

-

$$y \quad r \quad x \quad (a_r[1]a_x[1]e^{i(\,r[1]\quad x[1])}, a_r[2]a_x[2]e^{i(\,r[2]\quad x[2])}, \cdots) \quad (46)$$

r_1, r_2, r_3

$$h = \begin{bmatrix} h_{real} \\ h_{imaginary} \end{bmatrix} \quad (48)$$

$h \quad \square^{N_h}, h_{real}, h_{imaginary} \quad \square^{N_h/2}$

LSTM

$\hat{r}_i, \hat{r}_o$

$$\hat{g}_f, \hat{g}_i, \hat{g}_o, \hat{r}_i, \hat{r}_o \quad W_{xh}x_t \quad W_{hh}h_{t\ 1} \quad b_h \quad (49)$$

$$\hat{u} \quad W_{xu}x_t \quad W_{hu}h_{t\ 1} \quad b_u \quad (50)$$

0 1

$$\mathrm{bound}(h) \quad \begin{bmatrix} h_{real} / d \\ h_{imaginary} / d \end{bmatrix} \quad (51)$$

$$d \quad \max(1, \sqrt{h_{real} \circ h_{real} \quad h_{imaginary} \circ h_{imaginary}})$$

$$\square^{N_h/2} \quad (52)$$

/ ∘

d 1

u

r_i r_o

$$\begin{aligned} u & \quad \mathrm{bound}(\hat{u}) \\ r_i & \quad \mathrm{bound}(r_i) \\ r_o & \quad \mathrm{bound}(r_o) \end{aligned} \quad (53)$$

$r_i \quad \square^{N_h}$

$r_o \quad \square^{N_h}$

u g_i

$s \quad \{1, \ldots, N_{copies}\}$

-

$$r_{i,s} \quad \begin{bmatrix} P_s & 0 \\ 0 & P_s \end{bmatrix} r_i \quad (54)$$

$$c_{s,t} \quad g_f \circ c_{s,t\ 1} \quad r_{i,s} \quad (g_i \circ u) \quad (55)$$

$r_{i,s}$ $P_s \quad \square^{N_h/2 \ N_h/2}$

s

$$r \quad u \quad \begin{bmatrix} r_{real} \circ u_{real} \quad r_{imaginary} \circ u_{imaginary} \\ r_{real} \circ u_{imaginary} \quad r_{imaginary} \circ u_{real} \end{bmatrix} \quad (56)$$

$r_{o,s}$ $r_{i,s}$

$$r_{o,s} \quad \begin{bmatrix} P_s & 0 \\ 0 & P_s \end{bmatrix} r_o \quad (57)$$

$$h_t \quad g_o \circ \mathrm{bound}(\frac{1}{N_{copies}} \sum_{s\ 1}^{N_{copies}} r_{o,s} \quad c_{s,t}) \quad (58)$$

3.2.2 LSTM

Zhang X [37]

Tree Long Short-Term Memory Networks

TLSTM LSTM

TLSTM

$D(w)$

w

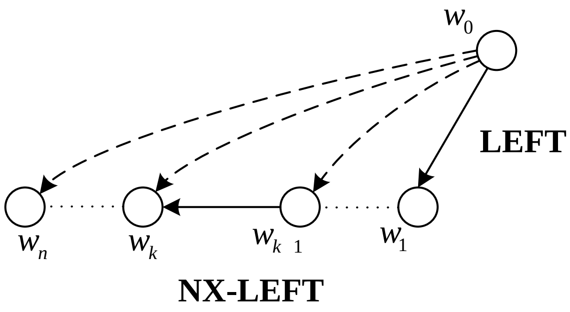
w_0 $_1$, $_2$, , $_n$

LEFT w_0

w_k w_0 $w_{k\ 1}$ w_k

NX-LEFT 13

RIGHT NX-RIGHT



13

T

S

T

breadth-first search BFS

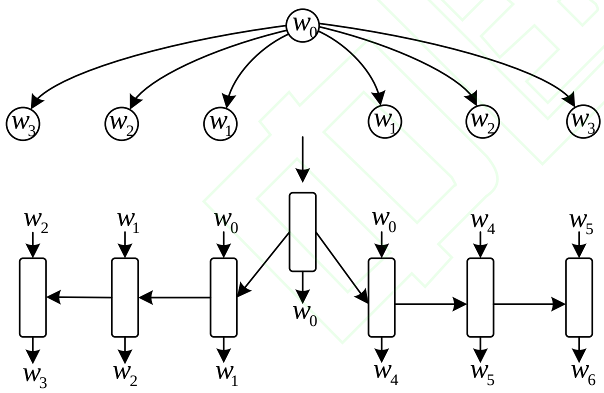
ROOT

14

w

T S

$$P(S\,|\,T) \qquad P(w\,|\,D(w)) \qquad (59)$$
$$_{w\ \ BFS(T)\backslash ROOT}$$



14 TLSTM

$D(w)$ <

\> LSTM

LSTM LEFT LSTM

NX-LEFT RIGHT NX-RIGHT 15

LSTM

t $1,n$ $\langle w_t, z_t \rangle$ $D(w_t)$

t t w

z_t LEFT,RIGHT,NX-LEFT,NX-RIGHT

LSTM

W H

t $H[:,t]$

$D(w)$

$$x_t \quad W_e \ e(w_t)$$
$$h_t \quad \mathrm{LSTM}^{z_t}(x_t, H[:,t\,])$$
$$H[:,t\,] \quad h_t \qquad (60)$$
$$y_t \quad W_{ho} \ h_t$$

W_e $\square^{s\ |V|}$

W_{ho} $\square^{|V|\ d}$ V

s d

H $\square^{d\ (n\ 1)}$

$e(w_t)$ w_t LSTM^{z_t}

LSTM t x_t

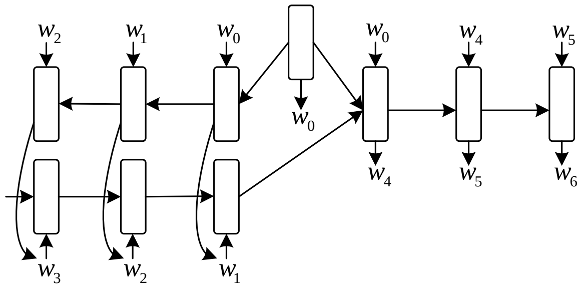
h_t h_t

LSTM

$D(w_t)$ softmax

$$P(w_t\,|\,D(w_t)) \quad \frac{\exp(y_t, w_t)}{{}_{k\ 1}^{|V|}\exp(y_t, k\,)} \qquad (61)$$

TLSTM



15

w_0

$p_j^t \quad \text{soft}\max(u^T \tanh(W \quad _j \quad W_x x_t \quad W_{\tilde{}} \tilde{}_{t\ 1}))$

72

$\tilde{}_t \quad \sum_{j\ 1}^{m} p_j^t \quad _j$

73

r_t

$r_t \quad (W_r \ [\tilde{}_t, x_t])$

74

c_t

h_t

$c_t \quad r_t \quad \tilde{}_t \quad f_t \quad \tilde{c}_t \quad i_t \quad \hat{c}_t$

75

$h_t \quad o_t \quad \tanh(c_t)$

76

Penn Treebank

RNN LSTM LSTM

Stanford Sentiment Treebank

LSTM

LSTM

T-CNN

3.2.4 LSTM

[40]

k -

$S \quad \{(x_i, y_i)\}_{i\ 1}^{k} \quad c_S(\hat{x})$

$\hat{x} \quad \hat{x} \quad \hat{y}$

$S \quad c_S(\hat{x})$

$P(\hat{y} \mid \hat{x}, S) \quad P(\hat{y} \mid \hat{x}, S)$

1

$\hat{x} \quad \hat{y} \quad \sum_{i\ 1}^{k} a(\hat{x}, x_i) y_i$

$P(\hat{y} \mid \hat{x}, S)$

$c_S(\hat{x}) \quad x_i, y_i$

$a \quad \hat{x}$

$x_i \quad a$

c softmax

$a(\hat{x}, x_i) \quad e^{c(f(\hat{x}), g(x_i))} / \sum_{j\ 1}^{k} e^{c(f(\hat{x}), g(x_j))}$

77

c

2

$g(x_i, S)$

LSTM $g(x_i, S) \quad \vec{h}_i \quad \overleftarrow{h}_i \quad g(x_i)$

LSTM S

$g(x_i) \quad x_i$

$\vec{h}_i, \vec{c}_i \quad \text{LSTM}(g(x_i), \vec{h}_{i\ 1}, \vec{c}_{i\ 1})$

$\overleftarrow{h}_i, \overleftarrow{c}_i \quad \text{LSTM}(g(x_i), \overleftarrow{h}_{i\ 1}, \overleftarrow{c}_{i\ 1})$

78

$h_i \quad c_i$ LSTM

$\overleftarrow{h} \quad i \quad |S|$

3

$\hat{x} \quad f(\hat{x}, S)$

LSTM attLSTM(, ,)

$$f(\hat{x},S) \quad \mathrm{attLSTM}(f(\hat{x}),g(S),K) \qquad (79)$$

f g LSTM

K LSTM

$g(S)$ S x_i

k

$$\hat{h}_k, c_k \quad \mathrm{LSTM}(f(\hat{x}),[h_{k\ 1},r_{k\ 1}],c_{k\ 1})$$
$$h_k \quad \hat{h}_k \quad f(\hat{x})$$
$$r_{k\ 1} \quad \sum_{i\ 1}^{|S|} a(h_{k\ 1},g(x_i))g(x_i) \qquad (80)$$
$$a(h_{k\ 1},g(x_i)) \quad \mathrm{softmax}(h_{k\ 1}^{T}g(x_i))$$

$\mathrm{LSTM}(x,h,c)$ x h

c a

$g(S)$ $r_{k\ 1}$ $h_{k\ 1}$

LSTM $\mathrm{attLSTM}(f(\hat{x}),g(S),K)$ h_K

3.3 MN

MN

MN

MN

MN

Dynamic Memory Networks DMN -

Key-Value Memory Networks KV-MemNN

Hierarchical Memory Network

HMN

Hierarchical Attentive Memory HAM

MN DMN

KV-MemNN

HMN

3.3.1

MN

[4]

$x_1,\cdots,x_n$ q a

x

x q

a

1 19

x_i

m_i c_i

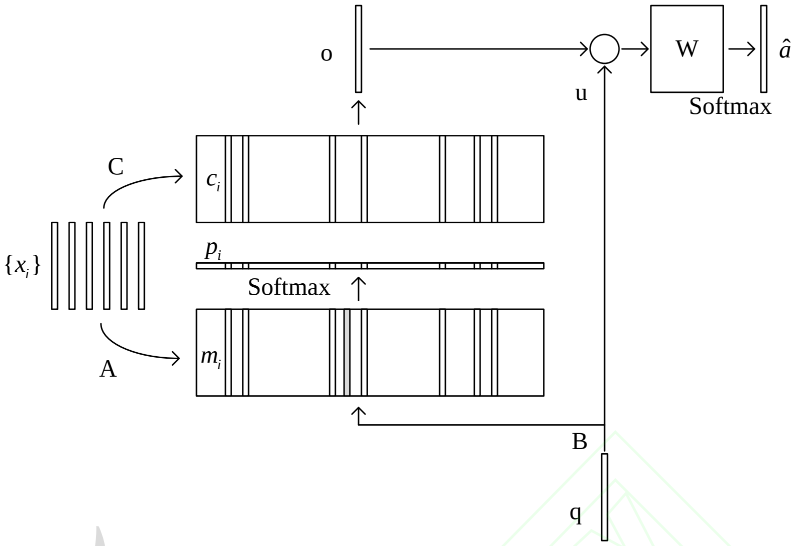
m_i u

p_i p_i

c_i o

o u W

$\hat{a}$



19

2

20

$u^{k\ 1}$

o^k u^k H

$u^{k\ 1}$

$u^{k\ 1}$ Hu^k o^k 81

$\{x_i\}$

W

$\hat{a}$

$\hat{a}$ $\mathrm{soft\,max}(Wu^{k\ 1})$ $\mathrm{soft\,max}(W(o^k\ \ u^k))$

82

Bag of Words **BoW**

x_{ij} i j

i

x_i $\{x_{i1}, x_{i2}, \ldots, x_{in}\}$

m_i $_j Ax_{ij}, c_i$ $_j Cx_{ij}$

u $_j Bq_j$

Position Encoding **PE**

m_i $_j l_j$ Ax_{ij}

l_{kj} $(1\ \ j/J)$ $(k/d)(1\ \ 2j/J)$ 83

l_j l_{kj}

J d

T_A T_C

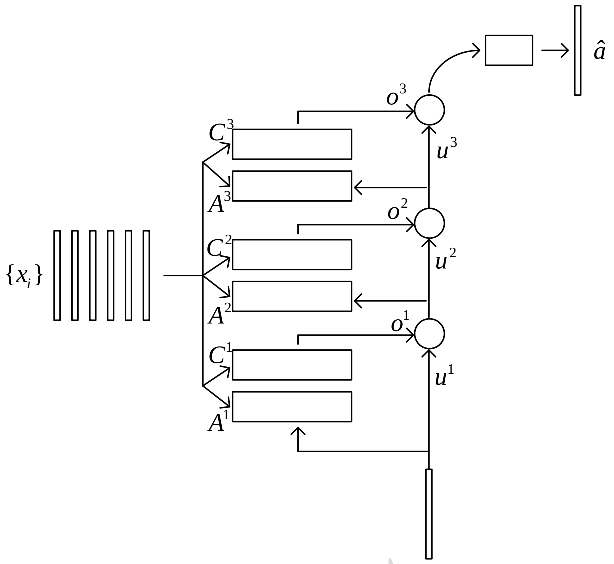
m_i c_i

m_i $_j l_j$ Ax_{ij} $T_A(i)$ 84

c_i $_j Cx_{ij}$ $T_C(i)$ 85

, $\hat{a}$

a



20

Penn Treebank Text8

RNN LSTM Structurally Constrained Recurrent Nets, SCRN

3.3.2

[41]

$u^{k\ 1}$ u^k o^k

$u^{k\ 1}$

$$T^k(u^k) \quad (W_T^k u^k \quad b_T^k) \qquad 86$$

$$u^{k\ 1} \quad o^k \ T^k(u^k) \quad u^k \ (1 \quad T^k(u^k)) \qquad 87$$

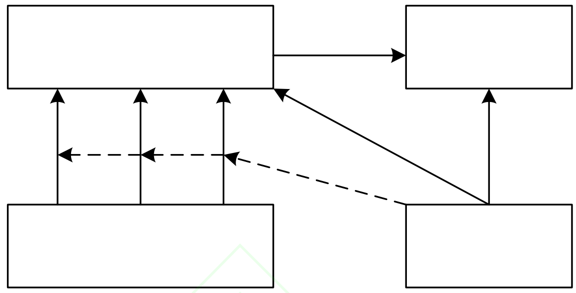
W_T^k, b^k k

T^k k

3.3.3

QA Kumar [42]

21



21 DMN

1

T_I

$w_1, \ldots, w_{T_I}$ Gated Recurrent Unit GRU

t x_t h_t GRU

$$\begin{aligned} z_t &\quad (W^{(z)}x_t \quad U^{(z)}h_{t\ 1} \quad b^{(z)}) \\ r_t &\quad (W^{(r)}x_t \quad U^{(r)}h_{t\ 1} \quad b^{(r)}) \\ \tilde{h}_t &\quad \tanh(Wx_t \quad r_t \circ Uh_{t\ 1} \quad b^{(h)}) \\ h_t &\quad z_t \circ h_{t\ 1} \quad (1 \quad z_t) \circ \tilde{h}_t \end{aligned} \qquad 88$$

$\circ$

$W^{(z)}, W^{(r)}, W \quad \square^{n_H \ n_I} \qquad U^{(z)}, U^{(r)}, U \quad \square^{n_H \ n_H}$

n GRU

$$h_t \quad GRU(x_t, h_{t\ 1})$$

RNN

GRU

T_C T_I c_t

t

T_C

2

T_Q

t

$q_t \quad GRU(L[w_t^Q], q_{t-1})$ L w_t^Q

t

$q \quad q_{T_Q}$

3

c

q m^{i-1}

$g_t^i \quad G(c_t, m^{i-1}, q)$ c_t

g_t^i c m

q

$$z(c,m,q) \quad [c,m,q,c\circ q,c\circ m,|c \quad q|,|c \quad m|,$$
$$c^T W^{(b)} q, c^T W^{(b)} m] \quad 89$$

$z(c,m,q)$

$G(c,m,q)$

$$G(c,m,q) \quad (W^{(2)} \tanh(W^{(1)} z(c,m,q) \quad b^{(1)}) \quad b \quad 90$$

GRU $c_1,\cdots,c_{T_C}$

GRU g_t^i

GRU t

$h_t^i \quad g_t^i GRU(c_t, h_{t-1}^i) \quad (1 \quad g_t^i) h_{t-1}^i$

$e^i \quad h_{T_C}^i$

e^i

$m^i \quad GRU(e^i, m^{i-1})$ T_M

m^{T_M}

4

GRU

$a_0 \quad m^{T_M}$ $y_t \quad \mathrm{softmax}(W^{(a)} a_t)$

a_t $a_t \quad GRU([y_{t-1}, q], a_{t-1})$

GRU

y_t

Facebook bAbI

20 93.6%

93.3%

Stanford Sentiment Treebank (SST)

97.5%

3.3.4 -

QA Miller [43]

KV-MemNN 22

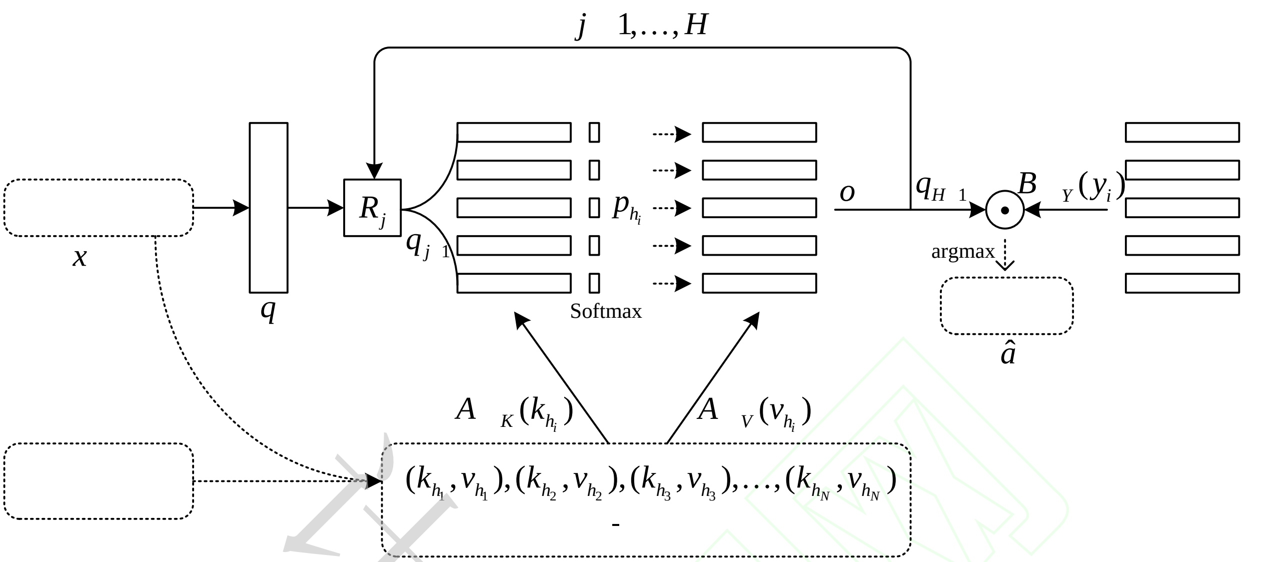
\- Key-value

key

value

KV-MemNN x

$(k_1,v_1),\ldots,(k_M,v_M)$



22 KV-MemNN

1 Key Hashing

N

$(k_{h_1},v_{h_1}),\ldots,(k_{h_N},v_{h_N})$ 1000

2 Key Addressing

Softmax k_{h_i} x

$$p_{h_i} = \text{Softmax}(A_X(x) \cdot A_K(k_{h_i})) \quad (93)$$

D A

$d' \times D$

3 Value Reading

$$o = \sum_i p_{h_i} A_V(v_{h_i})$$

x

$$q = A_X(x)$$

o q

$q_2 = R_1(q + o)$ R_1 $d \times d$

j hop R_j

q

$$p_{h_i} = \text{softmax}(q_{j+1}^T A_K(k_{h_i}))$$

H

$$\hat{a} = \text{argmax}_{i=1,\ldots,C}\,\text{softmax}(q_{H+1}^T B_Y(y_i)) \quad (91)$$

$y_i \in Y$

3.3.5

Chandar

Maximum Inner Product Search

MIPS

Hierarchical Memory Network HMN [44]

MN HMN

1 HMN

hash

MIPS

MIPS

MIPS

2 HMN

MIPS

HMN

MIPS MIPS

HMN K-MIPS

$\{x_1,\dots,x_n\}$ q K-MIPS

$\arg\max_i^{(K)} q^T x_i$ $\arg\max^{(K)}$ K

$q^T x_i$ $\{x_1,\dots,x_n\}$

q

HMN K-MIPS

MIPS

softmax

$$C \quad \arg\max^{(K)} h(q) M^T$$
$$R_{out} \quad \mathrm{softmax}^{(K)}(h(q)M^T) \qquad (92)$$
$$\mathrm{softmax}(h(q)M[C]^T)$$

$h(q) \quad \mathbb{R}^d$ C K

MIP $M \quad \mathbb{R}^{N \; d}$

N $M[C]$ M

$M[C]$ C

K-MIPS Auvolat

[45]

K-MIPS MIPS

MIPS Maximum Cosine Similarity Search MCSS

$$\arg\max_i^{(K)} \frac{q^T x_i}{\|q\|\|x_i\|} \quad \arg\max_i^{(K)} \frac{q^T x_i}{\|x_i\|} \qquad (93)$$

x_i MCSS

MIPS

MIPS MCSS

P Q

MIPS MCSS

$$P(x) \quad x, 1/2 \quad \|x\|_2^2, 1/2 \quad \|x\|_2^4, \dots, 1/2 \quad \|x\|_2^{2^m}$$
$$Q(x) \quad x, 0, 0 \dots, 0 \qquad (94)$$

q MIPS

$$\arg\max_i^{(K)} q^T x_i \quad \arg\max_i^{(K)} \frac{Q(q)^T P(x_i)}{\|Q(q)\|_2 \; \|P(x_i)\|_2} \qquad (95)$$

MIPS MCSS K

[46]

K K

HMN

1 K

mini-batch

GPU

2

3

K

3.3.6 MN

(n)

n n

(n^2)

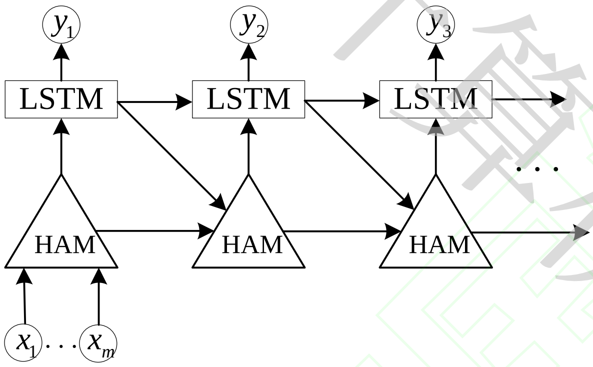
Andrychowicz Hierarchical Attentive Memory

HAM [47]

$(\log n)$

(n)

HAM HAM LSTM 23



23 HAM

x_1, x_2, L

y_1, y_2, L x_i, y_i Î $\{0,1\}^b$ b

n n 2

HAM n

$|V| = 2n - 1$ L Ì V

$l(e)$ $r(e)$

e e Î $V \setminus L$

HAM LSTM

HAM

1 EMBED (x_i)

x_i

0 JOIN

$$h_e = \mathrm{JOIN}(h_{l(e)}, h_{r(e)})$$

2

SEARCH p

1-p

a

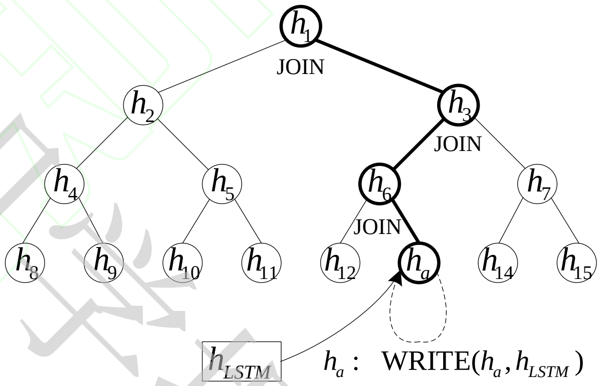
3 a h_a

LSTM y_t Î $\{0,1\}^b$ sigmoid

y_t

4 24

a WRITE



24 HAM

EMBED ^{b} ® ^{d} JOIN

^{d} ^{d} ? ^{d} SEARCH

^{d} ^{l} [0,1] WRITE

^{d} ^{l} ? ^{d}

Multilayer Perceptrons MLPs

-

[48] x y

A

y

$$L \quad \log p(y|x,\ ) \quad \log \quad p(A|x,\ )p(y|A,x,\ )_{A}$$

96

$$F \quad _{A} \quad p(A|x,\ )p(y|A,x,\ ) \quad L$$ 97

$$F \quad _{A} \quad p(A|x,\ )[\ \log p(y|A,x,\ )$$

$$\log p(y|A,x,\ ) \quad \log p(A|x,\ )]$$ 98

LSTM HAM

LSTM LSTM

FIFO

LSTM

99% 25%

LSTM HAM 0.04%

2 4 LSTM

LSTM

HAM 2.48%

$Q(\log n)$

3.3.7 MN

SimpleQuestions [50]

QA

QA

1

Freebase

Reverb

s

r o k

$y \quad (s,r,\{o_1,\ldots,o_k\})$ $f(y)$ ^{N_S}

N_S $f(y)$

1

1/k

q $g(q)$ ^{N_V} N_V

q 1

0

$y \quad (s,r,o)$

$h(y)$ $^{N_S \ N_V}$ s o

r

2

3

n

n

n

n

W_V $^{d \ N_V}$

W_S $^{d \ N_S}$ q y

$$S_{QA}(q,y) \quad \cos(W_V g(q), W_S f(y))$$ 99

W_V W_S

y

W_{VS} $^{d \ (N_V \ N_S)}$

$$S_{RVB}(q,y) \quad \cos(W_V g(q), W_{VS} h(y))$$ 100

W_V W_S

4

WebQuestions SimpleQuestions
Reverb

WebQuestions
41.3 F1 41.2
SimpleQuestions
62-63
86 SimpleQuestions

Reverb
67%

3.3.8 DMN
DMN
Xiong [51]

DMN
QA QA

1 QA
DMN GRU

GRU

GRU
Xiong GRU

$[w_1^j, \ldots, w_{M_i}^j]$

f_i

$$f_i \quad \sum_{j\ 1}^{M} l_j \circ w_j^i \qquad (101)$$

$$l_{jd} \quad (1 \quad j/M) \quad (d/D)(1 \quad 2j/M) \qquad (102)$$

l_j l_{jd}

d
D M

f_i GRU

$$\overrightarrow{f_i} \quad GRU_{fwd}(f_i, \overrightarrow{f_{i\ 1}})$$
$$\overleftarrow{f_i} \quad GRU_{bwd}(f_i, \overleftarrow{f_{i\ 1}}) \qquad (103)$$
$$\overleftrightarrow{f_i} \quad \overleftarrow{f_i} \quad \overrightarrow{f_i}$$

2 QA

QA
VGG-19
CNN

q

F
F GRU $\overleftrightarrow{F}$

3

QA
$\overleftrightarrow{F} \quad \overleftrightarrow{f_1}, \cdots, \overleftrightarrow{f_N}$

g_i^t $\overleftrightarrow{f_i}$

$\overleftrightarrow{f_i}$ q

g_i^t

$$z_i^t \quad [\overleftrightarrow{f_i} \circ q; \overleftrightarrow{f_i} \circ m^{t\ 1}; |\overleftrightarrow{f_i} \quad q|; |\overleftrightarrow{f_i} \quad m^{t\ 1}|]$$
$$Z_i^t \quad W^{(2)} \tanh(W^{(1)} z_i^t \quad b^{(1)}) \quad b^{(2)}$$
$$g_i^t \quad \frac{\exp(Z_i^t)}{\sum_{k\ 1}^{M_i} \exp(Z_i^t)} \qquad (104)$$

$\overleftrightarrow{f_i}$ i $m^{t\ 1}$
∘

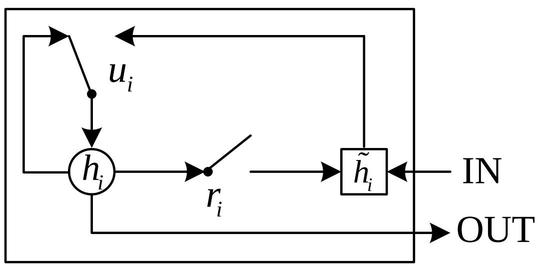
g_i^t

GRU c_t

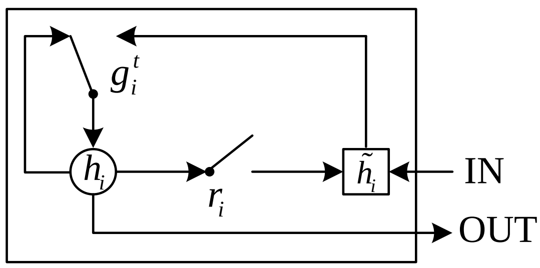
g_i^t c_t $\sum_{i=1}^{N} g_i^t \overleftrightarrow{}_i$ GRU

GRU

$\overleftrightarrow{F}$



25 GRU



GRU

/

g_i^t GRU u_i

25 GRU

$$h_i = g_i^t \circ \tilde{h}_i + (1 - g_i^t) \circ h_{i-1} \quad 105$$

3.3.9 KV-Mn≥C /PB53.184 Tc 13.08 0 Td(KV

h_N c_t

c_t m^{t-1}

m^t

$$m^t = GRU(c^t, m^{t-1}) \quad 106$$

GRU q

ReLU m^t

$$m^t = RuLU(W^t[m^{t-1}; c^t; q] + b) \quad 107$$

$W^t \in \mathbb{R}^{n_H \times n_H}, b \in \mathbb{R}^{n_H}, n_H$

Ramachandran DMN Dynamic Memory Tensor Network DMTN [52]

[53]

DMN 80% Facebook bAbI 20% Visual Question Answering VQA VQA

$(k_1, v_1), \cdots, (k_T, v_T)$

Key CNN

I_i RNN

h_i^e

k_i h_i^e

Value I_i

v_i (I_i)

- (k_i, v_i)

26

a_i

LSTM $_{t\ 1}(K)$

$h_t^k \quad f^k(\ _{t\ 1}(K), h_{t\ 1}^k) \qquad f^k$

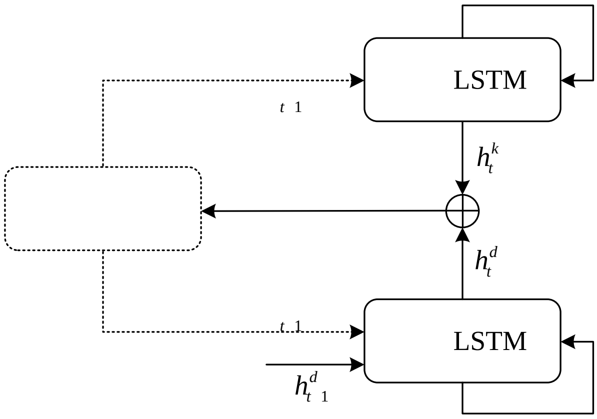
t q $h_{t\ 1}^d$

h_t^k

$q \quad W_k h_t^k \quad W_d h_{t\ 1}^d$ 108

q i k_i

tanh i k_i



26

$e_i^t \quad w_t \tanh(q \quad U_a k_i)$ 109

W_k, W_d, w_t, U_a

softmax

$_i^t \quad \exp\{e_i^t\} / \sum_{j\ 1}^{N} \exp\{e_j^t\}$ 110

-

$_t(K) \quad \sum_{i\ 1}^{T} \ _i^{(t)} k_i, \quad _t(V) \quad \sum_{i\ 1}^{T} \ _i^{(t)} v_i$ 111

$_i^t$

$_t(K)$

$_t(V)$

3 LSTM

LSTM

RNN h_t c_t

LSTM

x_t f_t

[55]
o_t

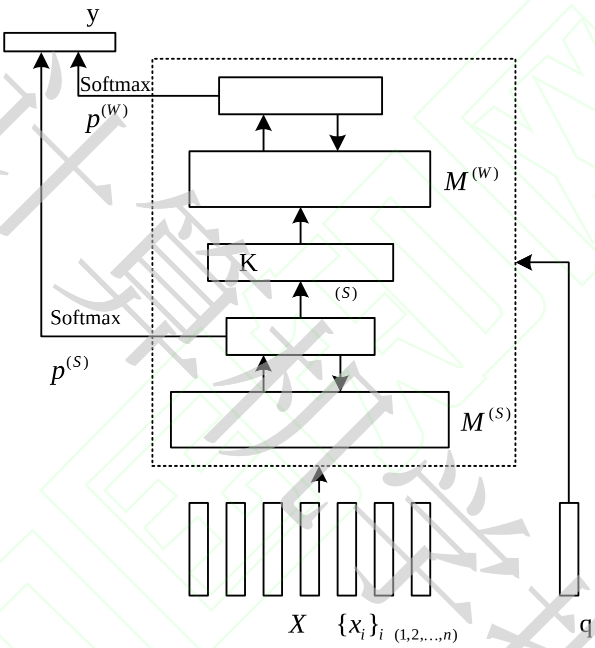
h_t softmax k
k

p_t

$$p_t \quad \mathrm{soft\,max}(U_p[h_t, x_t, \;_t(V)]) \quad b_p \qquad 112$$

3.3.10 HMN

27



27 HMN

n **K**

$X \quad \{x_i\}_{i \quad (1,2,\ldots,n)}$ q X (S) k

$M^{(S)}$

$M^{(W)}$

soft-search k GRU

$\{\;_t\}_{t \quad (1,2,\ldots,|t|)}$ t

y GRU GRU

$\vec{h}_t \quad G\vec{R}U(C^{R}\;_t) \quad \overleftarrow{h}_t \quad G\overleftarrow{R}U(C^{R}\;_t)$

C^R

$p^{(S)}$ GRU

$\vec{h}_t$ GRU $\overleftarrow{h}_t$

$M^{(W)}$ $\{m_t\}_{t \quad (1,2,\ldots,|t|)}$ m_t $\vec{h}_t$ $\overleftarrow{h}_t$

m_t X t

$\bar{w}_t$

$\hat{}$ $\{\hat{}_t\}_{t \quad (1,2,\ldots,|\hat{t}|)}$

$\hat{X}$

$u_R^{(S)}$ $\hat{w}$

$\{\hat{m}_t\}_{t \quad (1,2,\ldots,\hat{t}}$,

$^{(W)}$ $\{ {}_t^{(W)}\}_{t \quad (1,2,\ldots,|\hat{t}|)}$

$${}_t^{(W)} \quad \mathrm{softmax}(v^T \tanh(W u_R^{(S)} \quad U\hat{m}_t)) \qquad 113$$

v $^{d \ 1}$, W $^{d \ d}$, U ? ? $^{d? \ d}$

Oriol Vinyals [53]

$\hat{w}$

$$p^{(W)}(\ ) \quad trans(p^{(W)}(\hat{})) \quad trans(\ ^{(W)}) \qquad 114$$

$trans(\ )$

$p^{(W)}(\hat{w})$ Î ¡ $^{|\hat{t}|}$

$p^{(W)}(w)$ Î ¡ $^{|V|}$

0

$M^{(S)}$

$p^{(S)}$ $M^{(W)}$ $p^{(W)}$

$$p(w) = p^{(S)}(w) + p^{(W)}(w) \qquad 115$$

y

[56]

HMN

HMN

3.4

3.4.1

urlanello

Active Long Term Memory Networks A-LTM [58]

Distillation Loss [59]

McClelland [60]

Hippocampus Neocortex

Catastrophic Inference CI

A-LTM

N

H

N

H

N

A-LTM

N

N

N 0

N H H

H

w_0^0 w_0^* w_1^0 w_1^* $w_2^0 \sim N(0,\ )$ 116

$f(w_0^*, w_1^*; x_1)$

N y_1

3.4.2

Colin Raffel

[61] 28

Bahdanau [62]

h_t

a_{tj}

h_j c_t

$$c_t = \sum_{j=1}^{T} a_{tj} h_j \quad 117$$

T a_{tj} h_j

t

$$e_{tj} = a(s_{t-1}, h_j), a_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T} \exp(e_{tk})} \quad 118$$

a

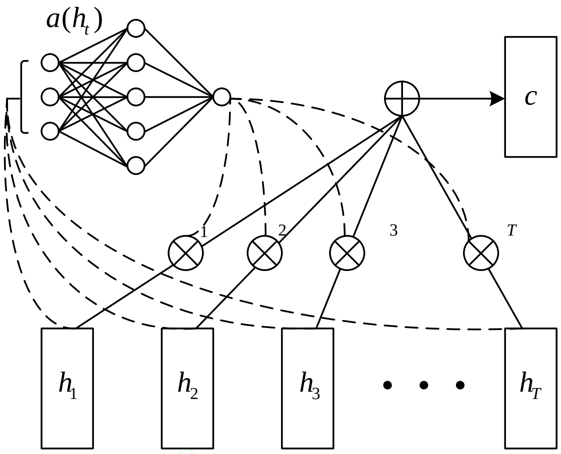
s_{t-1} h_j

s_t

s_t s_{t-1}

c_t t-1

$s_t = f(s_{t-1}, c_t, y_{t-1})$



28

h_t

$a(h_t)$ a a

h_t c_t

c_t

$$e_t = a(h_t), a_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_t)}, c = \sum_{t=1}^{T} a_t h_t \quad 119$$

a_t h_t

a_t h_t

c

T

T

h_t

$$c = \frac{1}{T} \sum_{t=1}^{T} h_t$$

x_t

$h_t = \mathrm{LReLU}(W_{xh}x_t + b_{xh})$ 120

W_{xh} $^{D\times 2}$, b_{xh} ^{D} $\mathrm{LReLU}(x)$

$$\mathrm{LReLU}(x) = \max(x, 0.01x)$$

$$a(h_t) = \tanh(W_{hc}h_t + b_{hc})$$

104 c

s

$s = \mathrm{LReLU}(W_{cs}c + b_{cs}), W_{cs}$ $^{D\times D}$, b_{cs} ^{D}

121

y

$y = \mathrm{LReLU}(W_{sy}s + b_{sy}), W_{sy}$ $^{1\times D}$, b_{sy}

122

y

adam

[63]

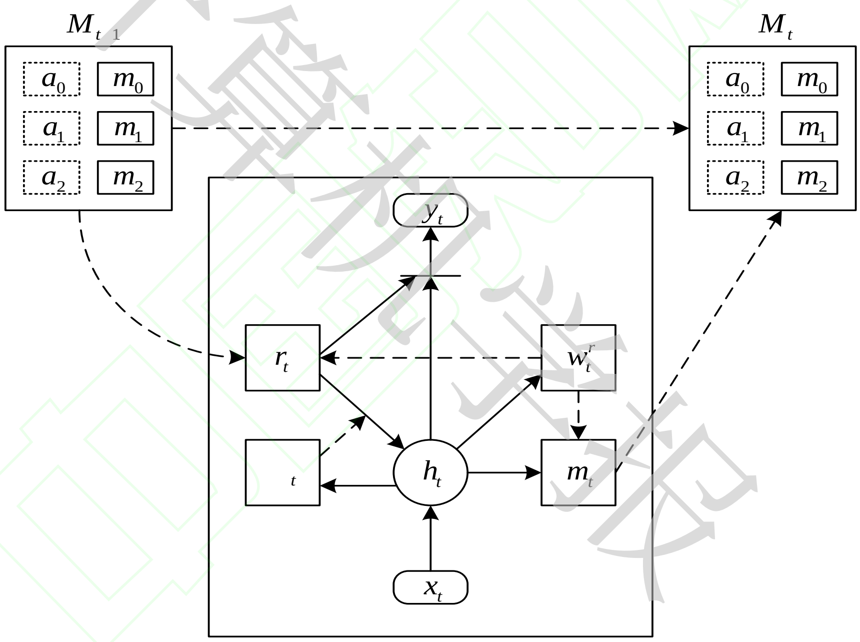
3.4.3

Gulcehre

Temporal Automatic Relation Discovery In Sequences TARDIS

[64]

TARDIS



29 TARDIS

TARDIS 29

M_t RNN

w_t^r M_t

r_t $(M_t)^T w_t^r$ TARDIS

$M_t[i]$ $W_m h_t$

h_t

r_t x_t

h_{t-1} h_t (x_t, h_{t-1}, r_t)

TARDIS

TARDIS M_t

$A_t \in \mathbb{R}^{k\times a}$ $C_t \in \mathbb{R}^{k\times c}$

$M_t = [A_t; C_t] \in \mathbb{R}^{k\times (c+a)}$

$\bar{w}_t^r$

$$\alpha_t[i] = a^T \tanh(W_h h_t + W_x x_t + W_m M_t[i] + W_u u_t)$$
$$\bar{w}_t^r = \mathrm{softmax}(\alpha_t) \quad (123)$$

$\{a, W_h^g, W_x^g, W_m^g, W_u^g\}$

w_t^r $\bar{w}_t^r$ $\arg\max \bar{w}_t^r$

u_t

$$u_t = \mathrm{norm}\left(\sum_{i=1}^{t-1} w_i^r\right) \quad (124)$$

u_t

TARDIS TARDIS LSTM

Gumbel-sigmoid

LSTM c_t

$$\tilde{c}_t = \tanh(\alpha_t W_h^g h_{t-1} + W_x^g x_t + \beta_t W_r^g r_t)$$
$$c_t = f_t c_{t-1} + i_t \tilde{c}_t \quad (125)$$
$$h_t = o_t \tanh(c_t)$$

f_t, i_t, o_t LSTM

α_t, β_t

LSTM

c_t

3.4.4

Sarthak Jain[65]

Freebase

n n-gram n

L

1 $F = \{f_1, \dots, f_{|F|}\}$

f (s, R, o) R s o

2 $q \in \mathbb{R}^d$

$f \in F$

$g(f)$ $h(f)$

$g(f)$

q (s, R) $g(f) = q^T(s + R)$

softmax $h(f)$

$h(f)$

$F = \{f_1, \ldots, f_{|F|}\}$ q

1

2 $f = (s, R, o)$

q

q q $h(f)(s \quad R)$ 126
$f \quad F$

3 f

o

f F

q

$h(f)$

$$L_{QA} = \sum_{(q,A)} \sum_{n=1}^{L} \frac{n}{L} \left\| |F_n| \sum_{a \in A} a - |A| \sum_{f \in F_n} h(f) ? o \right\|^2$$ 127

D -

q A

n n L

o

a $h(f)$

3.4.5

Rae

Sparse Access Memory SAM [66]

N SAM

$(\log N)$

(N) (1)

SAM

1

$$\tilde{r}_t \quad \sum_{i \quad 1}^{K} \tilde{w}_t^R(s_i) M_t(s_i)$$ 128

$\tilde{w}_t^R$ $s_1, \ldots, s_K$

w_t^R K

0 K

q K

Approximate nearest neighbors ANN

$\tilde{w}_t^R$

2

$$w_t^W \quad {}_t({}_t w_{t\ 1}^R \quad (1 \quad {}_t) I_t^U)$$ 129

t t

$w_{t\ 1}^R$

I_t^U 0

K

I_t^U

U 1 0

$$U_T^{(1)}(i) = \sum_{t=0}^{T} l^{T-t}(w_t^W(i) + w_t^R(i)) \quad (130)$$

l

$$U_T^{(2)}(i) = T - \max\{t : w_t^W(i) + w_t^R(i) > d\} \quad (131)$$

d

3 LSTM

LSTM x_t

r_{t-1}

$p_t = (q_t, a_t, a_t, g_t)$

r_t LSTM

y_t

3.4.6

NTM

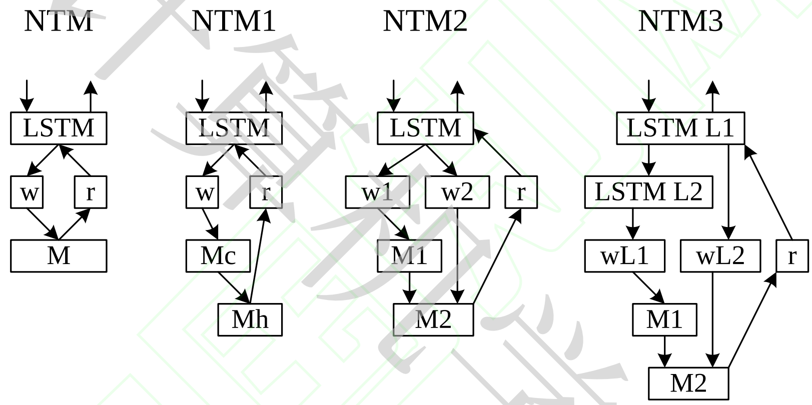
Zhang

[67]

Zhang

NTM NTM1 NTM1 NTM1

30



30 NTM

1 NTM1 NTM

M_h M_h

M_c M_h

M_c

t NTM1

$$M_c(t) = h(M_c(t-1), w(t-1), c(t))$$
$$M_h(t) = aM_h(t-1) + bM_c(t) \quad (132)$$
$$r(t) = w_r(t)M_h(t)$$

M_c $t-1$ $w(t-1)$

t

M_h t

$r(t)$ $c(t)$ h

a b $M_h(t)$

2 NTM2 M_1 M_2

M_2 w_2

M_1 w_1

M_2 M_1 w_2 NTM2

$$M_1(t), w_1(t) = h(M_1(t-1), w_1(t-1), c(t))$$
$$\tilde{M}_2(t), w_2(t) = h(M_2(t-1), w_2(t-1), c(t))$$
$$M_2(t) = a\tilde{M}_2(t) + bM_1(t)$$
$$r(t) = w_r(t)M_2(t)$$

133

M_1 M_2

M_2

3 NTM3

Luyang Li
[71]

Jianpeng Cheng LSTM
[39]

Machine Comprehension
Pan
[73]

Full-orientation matching

Hamid Palangi LSTM
[74]

Web

Sun

J
[80]

Mingxuan Wang
RNN [75]

Ganhotra
[76]

Huang RNN -
-
[77]

Tang D
[78]

LSTM

Chen P
[79]

Tran N
LSTM[81] LSTM

4.2

Xu Jia
[82]
LSTM
LSTM

- Arnav Kumar Jain
[54]
- -

Chunseong Park C
[83] Context Sequence Memory Network CSMN

Instagram

Oriol Vinyals
[40]

Adam Santoro
[32]

Wang J
[88]

Donahue J
[87] Long-Term Recurrent Convolutional Networks
LTRCN LSTM

Caiming Xiong

[43] Ma C
[84]

Kim K M
[85]

AI

Joel Moniz
[86] LSTM

Scene labeling
Abdulnabi A H
[89] CNN

Kaiser Ł
life-long [90]
one-shot

Parmar N

[91]

ImageNet

Fernando T
(Memory Augmented Conditional Generative Adversarial Networks, MC-GAN) [92]

4.3

Kim Y B

[93] Speaker-Sensitive Dual Memory Networks SSDMN
multi-turn slot tagging

Huang C Z A
[94]

text-to-speech TTS Naihan Li
Tacotron2 TTS [95]

4.4

Parisotto E
[96]

2D
Baskar M K
[97]

Bornschein J
[98]

[99]

Trang Pham DeepCare[100]

Prakash A
[101] Condensed Memory
Neural Networks CMNN

Schwaller P
[102] Molecular
Transformer MT

Jonathan Woodbridge LSTM
Domain Generate Algorithm DGA
[103] DGA
DGA

Lee S W
[104]

Fernando T [105]

Zadeh A
[106]

Khan A
[107]

Pham T
[108]

Knowledge Tracing

Zhang J -
[109] Dynamic Key-Value Memory Networks
DKV-MemNN

Sprechmann P

4.5

9

2D

9

	MN Skip-Thought MN DMN KV-MemNN MemN2N MN LSTM LSTM WTN ANN Evolved Transformer RNN MemN2N - RAN LSTM
	LSTM KV-MemNN CSMN LTRCN DMN MC-GAN
	SSDMN Music Transformer TTS
2D	DeepCare CMNN Molecular Transformer LSTM DGA DKV-MemNN

RNN LSTM NTM MN

5

5.1

RNN
LSTM
NTM
MN
RNN

-

-

5.2

1

2

3 MN

MN

LSTM MN

MN

[111]

4

Tishby [112]

[113]

5

MN

LSTM

6

[23]

7

B.ЗMNÈOÎX

9

16

10

CNN

GAN [90]

11

12

13

[100-102] [103] [107]

14

15

[1] Sepp Hochreiter, Jürgen Schmidhuber. Long Short-Term Memory. Neural Computation 9(8): 1735-1780 (1997)

[2] Jason Weston, Sumit Chopra, Antoine Bordes. Memory Networks. CoRR abs/1410.3916 (2014).

[3] Alex Graves, Greg Wayne, Ivo Danihelka. Neural turing machines. arXiv preprint arXiv:1410.5401, 2014.

[4] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus. End-to-end memory networks// Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, Roman Garnett. Advances in neural information processing systems. Montreal, Quebec, Canada. NIPS 2015: 2440-2448.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[6] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, Jürgen Schmidhuber. LSTM: A search space odyssey// IEEE Trans. Neural Netw. Learning Syst. 28(10): 2222-2232 (2017)

[7] Fader A, Zettlemoyer L, Etzioni O. Paraphrase-driven learning for open question answering[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013: 1608-1618.

[8] Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models[C]//Joint European conference

on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2014: 165-180.

[9] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//Eleventh annual conference of the international speech communication association. 2010.

[10] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[11] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.

[12] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[13] Kalchbrenner N, Espeholt L, Simonyan K, et al. Neural machine translation in linear time[J]. arXiv preprint arXiv:1610.10099, 2016.

[14] Zhou J, Cao Y, Wang X, et al. Deep recurrent models with fast-forward connections for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 371-383.

[15] Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv:1609.08144, 2016.

[16] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 1243-1252.

[17] Shazeer N, Mirhoseini A, Maziarz K, et al. Outrageously large neural networks: The sparsely-gated mixture-of

Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, Roman Garnett, Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain, NIPS 2016: 3630-3638

[41] Fei Liu, Julien Perez. Gated end-to-end memory networks// Mirella Lapata, Phil Blunsom, Alexander Koller, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain,.EACL 2017: 1-10

[42] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, Richard Socher. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing// Maria-Florina Balcan, Kilian Q. Weinberger, Proceedings of the 33nd International Conference on Machine Learning, New York City, NY, USA, ICML 2016: 1378-1387

[43] Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, Jason Weston. Key-Value Memory Networks for Directly Reading Documents// Jian Su, Xavier Carreras, Kevin Duh, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA, EMNLP 2016: 1400-1409

[44] Sarath Chandar, Sungjin Ahn, Hugo Larochelle, Pascal Vincent, Gerald Tesauro, Yoshua Bengio. Hierarchical memory networks. arXiv preprint arXiv:1605.07427, 2016.

[45] Alex Auvolat, Pascal Vincent. Clustering is efficient for approximate maximum inner product search. arXiv preprint arXiv:1507.05910, 2015.

[46] Zhong S. Efficient online spherical k-means clustering// Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on. IEEE, 2005, 5: 3180-3185.

[47] Marcin Andrychowicz, Karol Kurach. Learning efficient algorithms with hierarchical attentive memory. arXiv preprint arXiv:1602.03218, 2016.

[48] Ronald J. Williams. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. Machine Learning 8: 229-256 (1992)

[49] Da-Rong Liu, Shun-Po Chuang, Hung-yi Lee. Attention-based Memory Selection Recurrent Network for Language Modeling. arXiv preprint arXiv:1611.08656, 2016.

[50] Antoine Bordes, Nicolas Usunier, [illegible]

[65] Sarthak Jain. Question Answering over Knowledge Base using Factual Memory Networks// Proceedings of the Student Research Workshop, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

[86] Joel Moniz, Christopher J. Pal. Convolutional residual memory networks[J]. arXiv preprint arXiv:1606.05262, 2016.

[87] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description// IEEE Transactions on Pattern Analysis and Machine Intelligence. 39(4): 677-691 (2017)

[88] Junbo Wang, Wei Wang, Yan Huang, Liang Wang, Tieniu Tan. M3: Multimodal Memory Modelling for Video Captioning// 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, CVPR 2018: 7512-7520

[89] Abrar H. Abdulnabi, Bing Shuai, Stefan Winkler, Gang Wang. Episodic CAMN: Contextual Attention-Based Memory Networks with Iterative Feedback for Scene Labeling// 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, CVPR 2017: 6278-6287

[90] Lukasz Kaiser, Ofir Nachum, Aurko Roy, Samy Bengio. Learning to remember rare events. arXiv preprint arXiv:1703.03129, 2017.

[91] Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer[J]. arXiv preprint arXiv:1802.05751, 2018.

[92] Tharindu Fernando, Simon Denman, Sridha Sridharan, Clinton Fookes. Task Specific Visual Saliency Prediction with Memory Augmented Conditional Generative Adversarial Networks// 2018 IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, WACV 2018: 1539-1548

[93] Young-Bum Kim, Sungjin Lee, Ruhi Sarikaya. Speaker-sensitive dual memory networks for multi-turn slot tagging// 2017 IEEE Automatic Speech Recognition and Understanding Workshop, Okinawa, Japan, ASRU 2017: 541-546

[94] Huang C Z A, Vaswani A, Uszkoreit J, et al. Music transformer: Generating music with long-term structure[J]. 2018.

[95] Li N, Liu S, Liu Y, et al. Close to human quality TTS with transformer[J]. arXiv preprint arXiv:1809.08895, 2018.

[96] Emilio Parisotto, Ruslan Salakhutdinov. Neural map: Structured memory for deep reinforcement learning. arXiv preprint arXiv:1702.08360, 2017.

[97] Murali Karthick Baskar, Martin Karafiát, Lukás Burget, Karel Veselý, Frantisek Grézl, Jan Cernocký. Residual memory networks: Feed-forward approach to learn long-term temporal dependencies// 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, ICASSP 2017: 4810-4814

[98] Jörg Bornschein, Andriy Mnih, Daniel Zoran, Danilo Jimenez Rezende. Variational Memory Addressing in Generative Models// Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, Roman Garnett, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, NIPS 2017: 3923-3932

[99] Diederik P. Kingma, Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[100] Trang Pham, Truyen Tran, Dinh Q. Phung, Svetha Venkatesh. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine// James Bailey, Latifur Khan, Takashi Washio, Gillian Dobbie, Joshua Zhexue Huang, Ruili Wang, Advances in Knowledge Discovery and Data Mining - 20th Pacific-Asia Conference, Auckland, New Zealand, PAKDD (2) 2016: 30-41

[101] Aaditya Prakash, Siyuan Zhao, Sadid A. Hasan, Vivek V. Datla, Kathy Lee, Ashequl Qadir, Joey Liu, Oladimeji Farri. Condensed Memory Networks for Clinical Diagnostic Inferencing// Satinder P. Singh, Shaul Markovitch, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, AAAI 2017: 3274-3280

[102] Schwaller P, Laino T, Gaudin T, et al. Molecular transformer for chemical reaction prediction and uncertainty estimation[J]. arXiv preprint arXiv:1811.02633, 2018

[103] Jonathan Woodbridge, Hyrum S. Anderson, Anjum Ahuja, Daniel Grant. Predicting domain generation algorithms with long short-term memory networks. arXiv preprint arXiv:1611.00791, 2016.

[104] Sang-Woo Lee, Chung-yeon Lee, Dong-Hyun Kwak, Jung-Woo Ha, Jeonghee Kim, Byoung-Tak Zhang. Dual-memory neural networks for modeling cognitive activities of humans via wearable sensors. Neural Networks 92: 17-28 (2017)

[105] Tharindu Fernando, Simon Denman, Aaron McFadyen, Sridha Sridharan, Clinton Fookes. Tree Memory Networks for modelling long-term temporal dependencies. Neurocomputing 304: 64-81 (2018)

[106] Zadeh A, Liang P P, Mazumder N, et al Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, Louis-Philippe Morency. Memory Fusion Network for Multi-view Sequential Learning// Sheila A. McIlraith, Kilian Q. Weinberger, Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, AAAI 2018: 5634-5641

[107] Muhammad Asjad Khan, Hung Le, Kien Do, Truyen Tran, Aditya Ghose, Hoa Dam, Renuka Sindhgatta. Memory-Augmented Neural Networks for Predictive Process Analytics. CoRR abs/1802.00938 (2018)

[108] Trang Pham, Truyen Tran, Svetha Venkatesh. Relational dynamic memory networks. CoRR abs/1808.04247 (2018)

[109] Jiani Zhang, Xingjian Shi, Irwin King, Dit-Yan Yeung. Dynamic Key-Value Memory Networks for Knowledge Tracing// Rick Barrett, Rick Cummings, Eugene Agichtein, Evgeniy Gabrilovich, Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, WWW 2017: 765-774
[110] Pablo Sprechmann, Siddhant M. Jayakumar, Jack W. Rae, Alexander Pritzel, Adrià Puigdomènech Badia, Benigno Uria, Oriol Vinyals, Demis Hassabis, Razvan Pascanu, Charles Blundell. Memory-based parameter adaptation. arXiv preprint arXiv:1802.10542, 2018.
[111] , , , . [J]. Journal of Software, 2017, 11: 2905-2924.
[112] Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle[C]//2015 IEEE Information Theory Workshop (ITW). IEEE, 2015: 1-5.
[113] , . : [J]. , 2018, 1(11): 86.

Background

Deep memory network is a general term for neural network models with memory function, which is mainly to solve the prediction problem of sequence-dependent dependence, and can be predicted by memorizing the effective information learned before. Memory network usually have independent memory modules or other structures capable of memory function. The former stores important information in an independently readable and writable memory and reads it when needed;while the latter method usually modify the internal structure of the cell to retain the information that needs to be remembered.

Deep memory network have achieved unprecedented performance in a wide variety of different application areas. For example, image classification, face recognition, human-level concept learning, playing Atari games and AlphaGo.

Deep memory network combines the benefits of memory network and deep learning. On one hand, memory network has a wider scope of applicability since it can enhance the memory of the model. On the other hand, deep learning can extract a good representation at different levels of abstraction, which disentangles better the factors of variations underlying the data.