Sparse Attentive Backtracking: Temporal credit assignment through reminding

Nan Rosemary Ke[1,2], Anirudh Goyal[1], Olexa Bilaniuk [1],
Jonathan Binas[1] Chris Pal[2,4], Mike Mozer [3], Yoshua Bengio[1,5]

[1]Mila, Université de Montréal
[2]Mila, Polytechnique Montreal
[3]University of Colorado, Boulder
[4]Element AI
[5]CIFAR Senior Fellow

Credit assignment

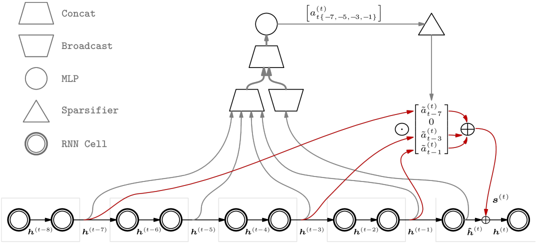
- Credit assignment: The correct division and attribution of blame to one's past actions in leading to a final outcome.
- Credit assignment in recurrent neural networks uses backpropgation through time (BPTT).
  - Detailed memory of all past events
  - Assign soft credit to almost all past events
  - Diffusion of credit?

Credit assignment through time and memory

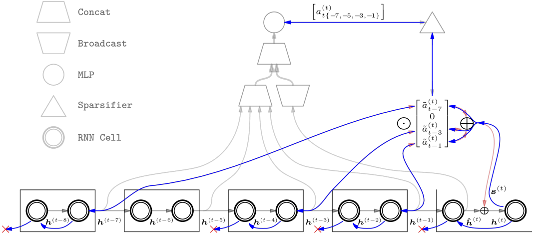
- Humans selectively recall memories that are relevant to the current behavior.
- Automatic reminding:
  - Triggered by contextual features.
  - Can serve a useful computation role in ongoing cognition.
  - Can be used for credit assignment to past events?
- Assign credit through only a few states, instead of all states:
  - Sparse, local credit assignment.
  - How to pick the states to assign credit to?

Sparse Attentive Backtracking

- Forward pass



- Backward pass



Some results

	k_{trunc}	k_{top}	Copying (T=100) acc.	Copying (T=100) CE_{10}	Copying (T=100) CE	Copying (T=200) acc.	Copying (T=200) CE_{10}	Copying (T=200) CE	Copying (T=300) acc.	Copying (T=300) CE_{10}	Copying (T=300) CE
LSTM	*full BPTT*		99.8	0.030	0.002	56.0	1.07	0.046	35.9	0.197	0.047
LSTM	*full self-attn.*		100.0	0.0008	0.0000	100.0	0.001	0.000	100.0	0.002	7.5e-5
LSTM	1	-	20.6	1.984	0.165				14.0	2.077	0.065
LSTM	5	-	31.0	1.737	0.145	17.1	2.03	0.092			
LSTM	10	-	29.6	1.772	0.148	20.2	1.98	0.090			
LSTM	20	-	30.5	1.714	0.143	35.8	1.61	0.073	25.7	1.848	0.197
LSTM	150	-	-	-	-	35.0	1.596	0.073	24.4	1.857	0.058
SAB	1	1	57.9	1.041	0.087	39.9	1.516	0.069	43.1	0.231	0.045
SAB	1	5	**100.0**	**0.001**	**0.000**				89.1	0.383	0.012
SAB	5	5	**100.0**	**0.000**	**0.000**	**100.0**	**0.000**	**0.000**	**99.9**	**0.007**	**0.001**
SAB	10	10	**100.0**	**0.000**	**0.001**	**100.0**	**0.000**	**0.000**			

Table 2: Test accuracy and cross-entropy (CE) loss performance on the copying task with sequence lengths of T=100, 200, and 300. Accuracies are given in percent for the last 10 characters. CE_{10} corresponds to the CE loss on the last 10 characters. These results are with mental updates; Compare with Table 4 for without.

Image class.	k_{trunc}	k_{top}	k_{att}	pMNIST acc.	CIFAR10 acc.
LSTM	*full BPTT*			90.3	58.3
LSTM	300	-	-		51.3
SAB	20	5	20	89.8	
SAB	20	10	20	90.9	
SAB	50	10	50	**94.2**	
SAB	16	10	16		**64.5**
Transformer (Vasvani'17)				**97.9**	62.2

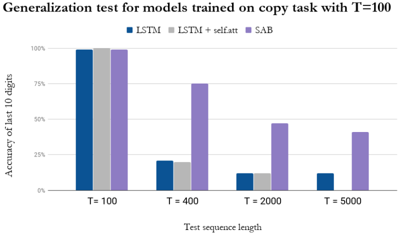
Table 4: Test accuracy for the permutated MNIST and CIFAR10 classification tasks.

Generalization and attention map

- Generalization on longer sequences

Transfer Learning Results

Copy len. (T)	LSTM	LSTM +self-a.	SAB
100	99%	100%	99%
200	34%	52%	**95%**
300	25%	28%	**83%**
400	21%	20%	**75%**
2000	12%	12%	**47%**
5000	12%	OOM	**41%**



- Learned attention over different timesteps during training

Copy Task with T = 200