Learning Distributed Representations of Symbolic Structure Using Binding and Unbinding Operations

Shuai Tang, Paul Smolensky, Virginia R. de Sa

Shuai Tang
Cognitive Science
UC San Diego

Paul Smolensky
Cognitive Science
Johns Hopkins University
Microsoft Research AI

Virginia R. de Sa
Cognitive Science
UC San Diego

Outline

- Motivations
- Our Proposed Recurrent Unit
- Experiments
- Conclusions

Outline

- Motivations

Distributed Representations

- Inducing structure in data
- Considerable power in statistical inference
- Encoding word knowledge
- Efficient usage of representation space

Symbolic Computing Systems

- Symbol ---- Substructure
  - Representations maintain the structure of data explicitly
  - Each substructure can be retrieved with no loss
- Inducing implicit structure from data
  - unique symbol ---- potential substructure

Distributed Representations + Symbolic Computing Systems

- Inducing structure in data
- Considerable power in statistical inference
- Encoding word knowledge
- Efficient usage of representation space

- Symbol ---- Substructure
  - Representations maintain the structure of data explicitly
  - Each substructure can be retrieved with no loss
- Inducing implicit structure from data
  - unique symbol ---- potential substructure

Learning Structured Distributed Representations

Tensor Product Representations (TPRs)

$$\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{r}_i \otimes \boldsymbol{f}_i = \sum_{i=1}^{N} \boldsymbol{r}_i \boldsymbol{f}_i^\top = \boldsymbol{R}\boldsymbol{F}^\top$$

- Binding Operation $\boldsymbol{r}_i \otimes \boldsymbol{f}_i$
- Unbinding Operation $\boldsymbol{f}_i = \boldsymbol{u}_i^\top \boldsymbol{S}$

$$\boldsymbol{u}_i^\top \boldsymbol{r}_j = \delta_{ij}$$

Tensor Product Representations (TPRs)

Symbols

$$\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{r}_i \otimes \boldsymbol{f}_i = \sum_{i=1}^{N} \boldsymbol{r}_i \boldsymbol{f}_i^\top = \boldsymbol{R}\boldsymbol{F}^\top$$

- Binding Operation $\boldsymbol{r}_i \otimes \boldsymbol{f}_i$
- Unbinding Operation $\boldsymbol{f}_i = \boldsymbol{u}_i^\top \boldsymbol{S}$

$$\boldsymbol{u}_i^\top \boldsymbol{r}_j = \delta_{ij}$$

Positions in a string

Part-of-speech tags

Context

Tensor Product Representations (TPRs)

$$\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{r}_i \otimes \boldsymbol{f}_i = \sum_{i=1}^{N} \boldsymbol{r}_i \boldsymbol{f}_i^\top = \boldsymbol{R}\boldsymbol{F}^\top$$

- Binding Operation $\boldsymbol{r}_i \otimes \boldsymbol{f}_i$
- Unbinding Operation $\boldsymbol{f}_i = \boldsymbol{u}_i^\top \boldsymbol{S}$

$$\boldsymbol{u}_i^\top \boldsymbol{r}_j = \delta_{ij}$$

Vector Representations

$$\boldsymbol{b} = \boldsymbol{R}\boldsymbol{f} \quad \text{and} \quad \boldsymbol{f} = \boldsymbol{U}^\top \boldsymbol{b}$$

Tensor Product Representations (TPRs)

$$\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{r}_i \otimes \boldsymbol{f}_i = \sum_{i=1}^{N} \boldsymbol{r}_i \boldsymbol{f}_i^\top = \boldsymbol{R}\boldsymbol{F}^\top$$

- Binding Operation $\boldsymbol{r}_i \otimes \boldsymbol{f}_i$
- Unbinding Operation $\boldsymbol{f}_i = \boldsymbol{u}_i^\top \boldsymbol{S}$

$$\boldsymbol{u}_i^\top \boldsymbol{r}_j = \delta_{ij}$$

simplify

Vector Representations

simplify

$$\boldsymbol{b} = \boldsymbol{R}\boldsymbol{f} \quad \text{and} \quad \boldsymbol{f} = \boldsymbol{U}^\top \boldsymbol{b}$$

Tensor Product Representations (TPRs)

$$\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{r}_i \otimes \boldsymbol{f}_i = \sum_{i=1}^{N} \boldsymbol{r}_i \boldsymbol{f}_i^{\top} = \boldsymbol{R}\boldsymbol{F}^{\top}$$

- Binding Operation $\boldsymbol{r}_i \otimes \boldsymbol{f}_i$
- Unbinding Operation $\boldsymbol{f}_i = \boldsymbol{u}_i^{\top} \boldsymbol{S}$

$$\boldsymbol{u}_i^{\top} \boldsymbol{r}_j = \delta_{ij}$$

Vector Representations

$$\boldsymbol{b} = \boldsymbol{R}\boldsymbol{f} \quad \text{and} \quad \boldsymbol{f} = \boldsymbol{U}^{\top}\boldsymbol{b}$$

binding complex

binding complex

Tensor Product Representations (TPRs)

$$\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{r}_i \otimes \boldsymbol{f}_i = \sum_{i=1}^{N} \boldsymbol{r}_i \boldsymbol{f}_i^\top = \boldsymbol{R}\boldsymbol{F}^\top$$

- Binding Operation $\boldsymbol{r}_i \otimes \boldsymbol{f}_i$
- Unbinding Operation $\boldsymbol{f}_i = \boldsymbol{u}_i^\top \boldsymbol{S}$

$$\boldsymbol{u}_i^\top \boldsymbol{r}_j = \delta_{ij}$$

Vector Representations

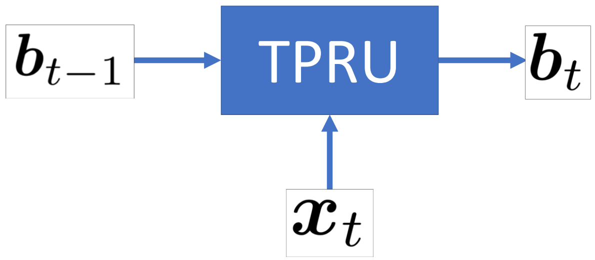
$$\boldsymbol{b} = \boldsymbol{R}\boldsymbol{f} \quad \text{and} \quad \boldsymbol{f} = \boldsymbol{U}^\top \boldsymbol{b}$$

learnt

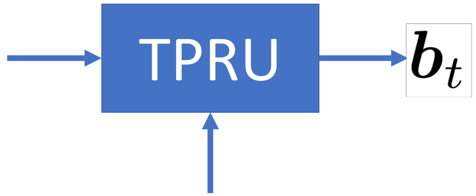
Outline

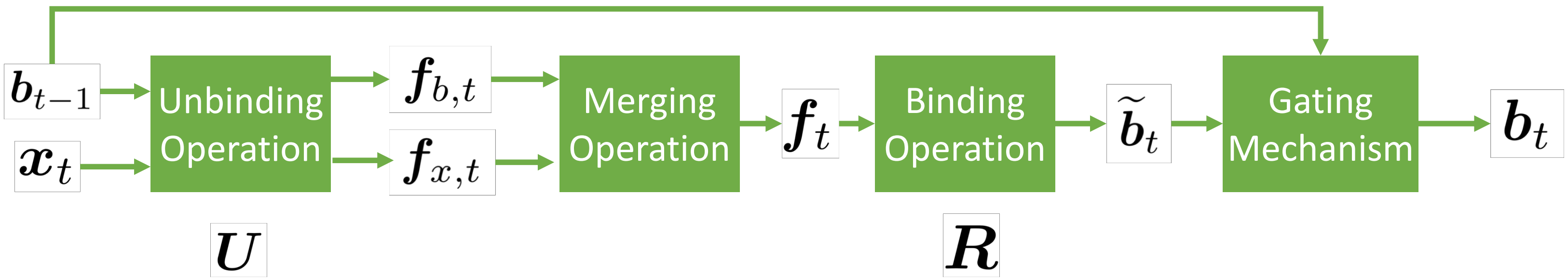
- Our Proposed Recurrent Unit

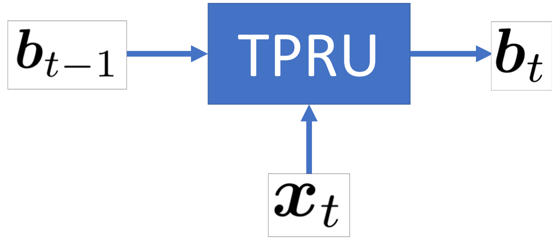
TPRU – Recurrent Unit



TPRU – Recurrent Unit







TPRU – Recurrent Unit

- Unbinding operation

$$\boldsymbol{f}_{b,t} = \boldsymbol{U}^\top \boldsymbol{b}_{t-1} \in \mathbb{R}^{N\times 1}, \qquad \boldsymbol{f}_{x,t} = \boldsymbol{U}^\top \boldsymbol{W}\boldsymbol{x}_t \in \mathbb{R}^{N\times 1}$$

$$(\widetilde{\boldsymbol{f}}_{b,t})_n = \mathrm{ReLU}\left((\boldsymbol{f}_{b,t})_n + b_b\right), \qquad (\widetilde{\boldsymbol{f}}_{x,t})_n = \mathrm{ReLU}\left((\boldsymbol{f}_{x,t})_n + b_x\right)$$

$$(\boldsymbol{f}_t)_n = \frac{\left((\widetilde{\boldsymbol{f}}_{b,t})_n + (\widetilde{\boldsymbol{f}}_{x,t})_n\right)^2}{\sum_{m=1}^{N}\left((\widetilde{\boldsymbol{f}}_{b,t})_m + (\widetilde{\boldsymbol{f}}_{x,t})_m\right)^2}$$

- Binding operation

$$\widetilde{\boldsymbol{b}}_t = \boldsymbol{R}\boldsymbol{f}_t$$

- Input Gate

$$\boldsymbol{b}_t = \boldsymbol{g}_t \circ \tanh(\widetilde{\boldsymbol{b}}_t) + (1 - \boldsymbol{g}_t) \circ \boldsymbol{b}_{t-1}$$
$$\boldsymbol{g}_t = \sigma(\boldsymbol{W}_b \boldsymbol{b}_{t-1} + \boldsymbol{W}_x \boldsymbol{x}_t)$$

TPRU – Unbinding Vectors

$\boldsymbol{U} = \boldsymbol{W}_u \boldsymbol{V}$

$\boldsymbol{R} = \boldsymbol{W}_r \boldsymbol{V}$

- Unbinding operation

$$\boldsymbol{f}_{b,t} = \boldsymbol{U}^\top \boldsymbol{b}_{t-1} \in \mathbb{R}^{N\times 1}, \qquad \boldsymbol{f}_{x,t} = \boldsymbol{U}^\top \boldsymbol{W} \boldsymbol{x}_t \in \mathbb{R}^{N\times 1}$$

- Binding operation

$$\widetilde{\boldsymbol{b}}_t = \boldsymbol{R}\boldsymbol{f}_t$$

- Input Gate

TPRU – Binding Vectors

$\boldsymbol{U} = \boldsymbol{W}_u \boldsymbol{V}$

$\boldsymbol{R} = \boldsymbol{W}_r \boldsymbol{V}$

- Unbinding operation

$$\boldsymbol{f}_{b,t} = \boldsymbol{U}^\top \boldsymbol{b}_{t-1} \in \mathbb{R}^{N\times 1}, \qquad \boldsymbol{f}_{x,t} = \boldsymbol{U}^\top \boldsymbol{W} \boldsymbol{x}_t \in \mathbb{R}^{N\times 1}$$

- Binding operation

$$\widetilde{\boldsymbol{b}}_t = \boldsymbol{R} \boldsymbol{f}_t$$

- Input Gate

TPRU – Parameters

$$\boldsymbol{U} = \boxed{\boldsymbol{W}_u}\boldsymbol{V} \qquad \boldsymbol{R} = \boxed{\boldsymbol{W}_r}\boldsymbol{V}$$

- Unbinding operation

$$\boldsymbol{f}_{b,t} = \boldsymbol{U}^\top \boldsymbol{b}_{t-1} \in \mathbb{R}^{N\times 1}, \qquad \boldsymbol{f}_{x,t} = \boldsymbol{U}^\top \boxed{\boldsymbol{W}} \boldsymbol{x}_t \in \mathbb{R}^{N\times 1}$$

$$(\widetilde{\boldsymbol{f}}_{b,t})_n = \text{ReLU}\left((\boldsymbol{f}_{b,t})_n + b_b\right), \qquad (\widetilde{\boldsymbol{f}}_{x,t})_n = \text{ReLU}\left((\boldsymbol{f}_{x,t})_n + b_x\right)$$

$$(\boldsymbol{f}_t)_n = \frac{\left((\widetilde{\boldsymbol{f}}_{b,t})_n + (\widetilde{\boldsymbol{f}}_{x,t})_n\right)^2}{\sum_{m=1}^{N}\left((\widetilde{\boldsymbol{f}}_{b,t})_m + (\widetilde{\boldsymbol{f}}_{x,t})_m\right)^2}$$

- Binding operation

$$\widetilde{\boldsymbol{b}}_t = \boldsymbol{R}\boldsymbol{f}_t$$

- Input Gate

$$\boldsymbol{b}_t = \boldsymbol{g}_t \circ \tanh(\widetilde{\boldsymbol{b}}_t) + (1 - \boldsymbol{g}_t) \circ \boldsymbol{b}_{t-1}$$

$$\boldsymbol{g}_t = \sigma(\boxed{\boldsymbol{W}_b}\boldsymbol{b}_{t-1} + \boxed{\boldsymbol{W}_x}\boldsymbol{x}_t)$$

Outline

- Experiments

Experiments

- Tasks
  - Logical Entailment in Propositional Logic (Evans et al., 2018)
  - Multi-genre Natural Language Inference (Williams et al., 2018)
  - General Purpose Sentence Representations (Conneau & Kiela, 2018)

- Plain & BiDAF architecture
  - BiDAF – Bi-Directional Attention Flow (Seo et al., 2017)

Logical Entailment in Propositional Logic

- Training set
- Validation set
- Test set
  - easy, big, hard, massive, exam

Connectives matter

Table 4: A truth table for $A = p \wedge q$ and $B = q$.

p	q	A	B	
T	T	T (1)	T (1)	$(1 = 1)$
T	F	F (0)	F (0)	$(0 = 0)$
F	T	F (0)	T (1)	$(0 < 1)$
F	F	F (0)	F (0)	$(0 = 0)$

Logical Entailment in Propositional Logic

- Training set
- Validation set
- Test set
    - easy, big, hard, massive, exam

A: ((g>((x|s)|((q&i)&o)))&(s&((i|v)|x)))

B: (~(((r|s)|q))>(~((q&(q|(s|r))))>(v|r)))

Connectives matter

Table 4: A truth table for $A = p \wedge q$ and $B = q$.

p	q	A	B	
T	T	T (1)	T (1)	$(1 = 1)$
T	F	F (0)	F (0)	$(0 = 0)$
F	T	F (0)	T (1)	$(0 < 1)$
F	F	F (0)	F (0)	$(0 = 0)$

Logical Entailment in Propositional Logic

model		valid	test easy	test hard	test big	test massive	test exam	# params
Mean $2^{\#Vars}$		75.7	81.0	184.4	3310.8	848,570.0	5.8	
Plain (BiDAF) Architecture - *dim 64*								
LSTM		71.7 (88.5)	71.8 (88.7)	64.1 (**74.5**)	64.2 (**73.8**)	53.7 (**66.8**)	68.3 (**80.0**)	65.5k (230.0k)
GRU		75.1 (87.9)	**77.1** (88.3)	63.7 (72.5)	63.8 (71.3)	54.4 (66.1)	73.7 (78.0)	49.1k (172.4k)
Ours	8	66.8 (86.2)	67.2 (87.1)	59.3 (69.1)	60.9 (68.2)	51.9 (62.5)	67.0 (74.3)	40.1k (131.3k)
	32	73.7 (88.4)	73.7 (88.4)	62.7 (71.1)	62.8 (70.1)	53.0 (64.9)	**76.7** (77.0)	
	128	75.9 (88.5)	76.0 (88.6)	**64.9** (71.5)	64.0 (69.8)	53.8 (64.1)	75.7 (**80.0**)	
	512	**76.8** (**88.6**)	76.8 (**89.2**)	64.4 (72.6)	**64.6** (71.2)	**54.6** (64.4)	75.3 (**80.0**)	
Plain (BiDAF) Architecture - *dim 128*								
LSTM †		64.5 (**88.6**)	64.2 (**89.3**)	59.7 (**74.7**)	62.1 (73.5)	50.9 (**67.4**)	65.0 (78.3)	196.6k (917.5k)
GRU ‡		**80.8** (86.2)	**80.3** (85.7)	65.9 (69.1)	**66.0** (69.1)	55.0 (63.1)	77.3 (72.7)	147.5k (688.1k)
Ours	8	63.7 (87.1)	63.4 (87.3)	57.5 (69.4)	59.6 (68.1)	51.3 (62.7)	65.0 (76.0)	131.1k (524.3k)
	32	71.5 (88.2)	71.7 (88.5)	62.6 (71.6)	62.4 (70.3)	52.0 (64.4)	78.3 (78.3)	
	128	72.8 (88.4)	73.1 (89.0)	63.8 (72.4)	62.8 (**71.5**)	52.6 (66.3)	71.3 (**80.0**)	
	512	79.6 (**88.6**)	79.6 (89.2)	**66.1** (72.7)	65.9 (70.8)	**55.2** (64.9)	**80.3** (79.7)	

Multi-genre Natural Language Inference

- 5 genres available in training set
- 10 genres presented in dev and test set

Both structure and word meaning matter

Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** N N N N	All of the children love working in their gardens.
At 8:34, the Boston Center controller received a third transmission from American 11	9/11 **entailment** E E E E	The Boston Center controller got a third transmission from American 11.
In contrast, suppliers that have continued to innovate and expand their use of the four practices, as well as other activities described in previous chapters, keep outperforming the industry as a whole.	OUP **contradiction** C C C C	The suppliers that continued to innovate in their use of the four practices consistently underperformed in the industry.

Multi-genre Natural Language Inference

model		MNLI dev matched	MNLI dev mismatched	# params
Plain (BiDAF) Architecture - *dim 512*				
LSTM		72.0 (76.0)	73.2 (75.5)	10.5m (29.4m)
GRU		72.1 (74.2)	72.8 (74.8)	7.9m (22.0m)
Ours	16	72.4 (73.9)	73.5 (75.0)	5.8m (15.7m)
	64	73.0 (74.8)	73.5 (75.5)	
	256	73.1 (75.9)	**73.9 (76.8)**	
	1024	**73.2 (76.2)**	73.8 (76.6)	
Plain (BiDAF) Architecture - *dim 1024*				
LSTM		72.5 (75.5)	73.9 (76.6)	25.2m (83.9m)
GRU		72.6 (74.8)	73.6 (75.9)	18.9m (62.9m)
Ours	16	72.9 (73.9)	73.7 (74.8)	14.7m (46.1m)
	64	73.4 (75.2)	74.4 (76.0)	
	256	73.7 (75.5)	74.6 (76.7)	
	1024	**74.2 (76.7)**	**74.7 (77.3)**	

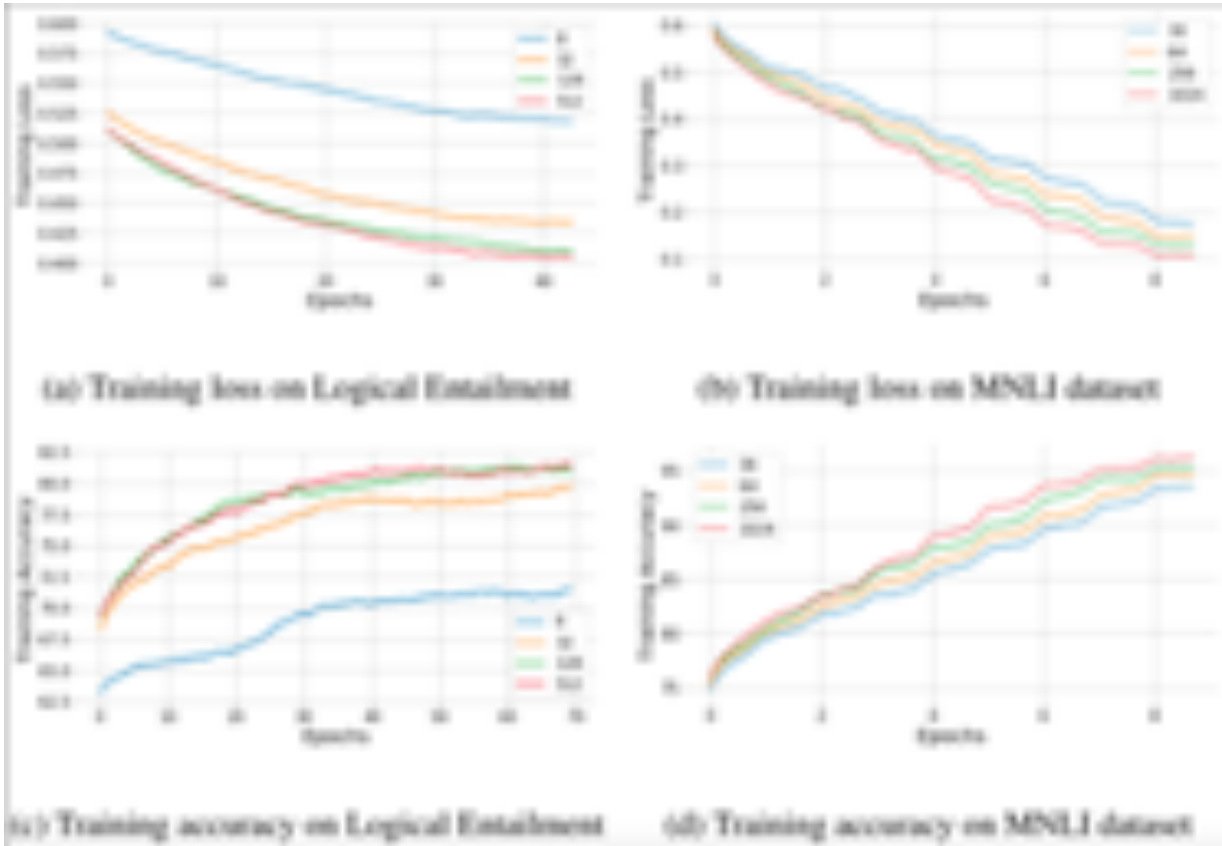
General Purpose Sentence Representations

Model		Downstream Tasks in SentEval						
		Binary	SST-5	TREC	SICK-E	STS (Su.)	STS (Un.)	MRPC
Measure		Accuracy				Pearson's $\rho \times 100$		Acc./F1
Plain Architecture - dim 512								
LSTM		87.0	47.5	89.7	84.4	81.8	62.5	**77.8** / 83.8
GRU		87.0	47.5	**91.1**	84.8	80.3	62.5	76.9 / 83.4
Ours	16	86.8	47.0	89.5	84.8	80.0	60.7	76.3 / 82.8
	64	87.1	46.9	89.9	85.1	80.8	62.1	76.8 / 83.3
	256	87.2	47.2	90.1	85.2	81.3	62.6	77.4 / **84.1**
	1024	**87.4**	**48.1**	90.5	**85.4**	**82.4**	**62.8**	77.1 / 83.9
Plain Architecture - dim 1024								
LSTM		87.6	47.3	**92.7**	85.0	**81.7**	63.3	77.0 / 83.6
GRU		87.5	**48.9**	92.6	85.8	81.2	62.8	**77.6** / 84.0
Ours	16	87.4	47.5	91.3	85.6	79.6	60.9	76.2 / 83.2
	64	87.8	47.8	92.0	85.6	80.7	62.3	77.5 / 83.8
	256	87.8	47.9	92.5	**86.0**	80.6	63.3	**77.6** / 83.9
	1024	**87.9**	48.5	91.9	85.9	81.5	**63.9**	77.5 / **84.4**

Incorporating more role vectors…

$$\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{r}_i \otimes \boldsymbol{f}_i$$

- Faster convergence rate
- Better performance



(a) Training loss on Logical Entailment

(b) Training loss on MNLI dataset

(c) Training accuracy on Logical Entailment

(d) Training accuracy on MNLI dataset

Outline

- Conclusions

Conclusions

- A TPRU (Recurrent Unit) is proposed to leverage both
  - Distributed Representations
  - Neural-Symbolic Computing
- Compared to LSTM and GRU
  - symbolic execution
  - reduced total number of parameters
  - comparable or better performance
- Incorporating more role vectors leads to
  - faster convergence rate and better results

Thank you!

@Shuai93Tang shuaitang93@ucsd.edu