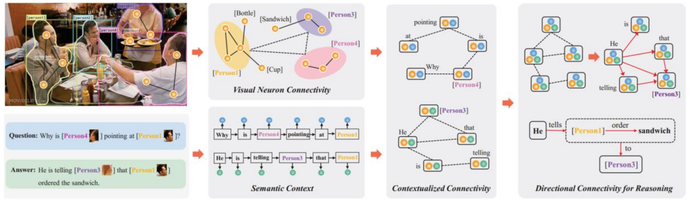
Connective Cognition Network for Directional Visual Commonsense Reasoning

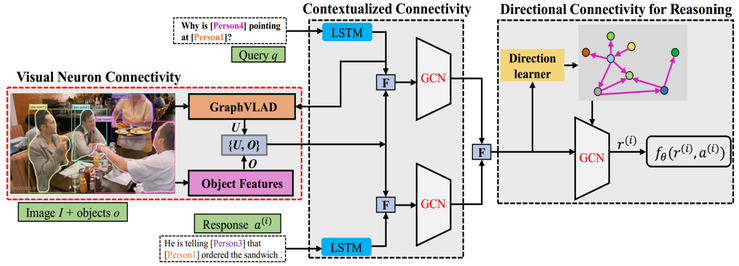
Aming Wu[1], Linchao Zhu[2], Yahong Han[1], Yi Yang[2]
[1]Tianjin University, [2]ReLER, University of Technology Sydney
{tjwam,yahong}@tju.edu.cn, {Linchao.Zhu, yi.yang}@uts.edu.au

Introduction

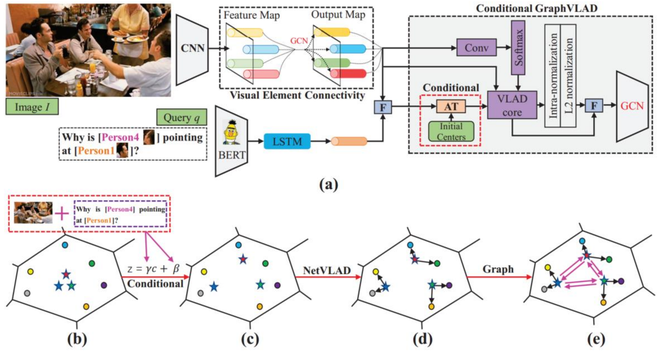
- **Visual Commonsense Reasoning (VCR)** answering challenging visual questions providing a rationale explaining why its answer is true.
- **Connective Cognition Network (CCN)** Inspired by neuroscience advances from brain connections to cognition, we propose CCN for VCR.



Connective Cognition Network



Conditional GraphVLAD



$$M = \mathbf{A}\widetilde{X}, \qquad \widetilde{M} = \tanh\left(w_f^c * M + b_f^c\right) \odot \sigma\left(w_g^c * M + b_g^c\right), \quad (1)$$

$$\gamma = f(|\widetilde{M}, \widetilde{Y}|), \quad \beta = h(|\widetilde{M}, \widetilde{Y}|), \quad z_i = \gamma c_i + \beta, \quad (2)$$

$$D_j = \sum_{i=1}^{N} \frac{e^{w_j^T \widetilde{M}_i + b_j}}{\sum_{j'} e^{w_{j'}^T \widetilde{M}_i + b_{j'}}} (\widetilde{M}_i - z_j) \quad (3)$$

Contextualized and Directional Connectivity

$$F_{qu} = softmax(\tilde{Q}U^T), \quad F_{qo} = softmax(\tilde{Q}O^T), \quad Q_U = F_{qu}U, \quad Q_O = F_{qo}O, \quad (4)$$

$$D_{qa} = \emptyset(E_{qa}), \quad G_t = D_{qa}D_{qa}^T, \quad D_t = sign(G_t), \quad V_e = softmax(abs(G_t)), \quad (5)$$

$$\mathbf{H} = D_t \odot V_e + I_d, \quad M_t = \mathbf{H}E_{qa}, \quad R_t = \tanh\left(w_f^r * M_t + b_f^r\right) \odot \sigma(w_g^r * M_t + b_g^r), \quad (6)$$

Experiments on VCR dataset

- ***Q*⟶*A*:** given a question, select the correct answer.
- ***QA*⟶*R*:** given a question and correct answer, select the correct rationale.
- ***Q*⟶*AR*:** given a question, select the correct answer, then the correct rationale.

Table 1: The performance of our CCN model on the VCR dataset.

Model	$Q \to A$ Val	$Q \to A$ Test	$QA \to R$ Val	$QA \to R$ Test	$Q \to AR$ Val	$Q \to AR$ Test
Revisited VQA [16]	39.4	40.5	34.0	33.7	13.5	13.8
BottomUpTopDown [1]	42.8	44.1	25.1	25.1	10.7	11.0
MLB [18]	45.5	46.2	36.1	36.8	17.0	17.2
MUTAN [5]	44.4	45.5	32.0	32.2	14.6	14.6
R2C (baseline) [38]	63.8	65.1	67.2	67.3	43.1	44.0
CCN	**67.4**	**68.5**	**70.6**	**70.5**	**47.7**	**48.4**

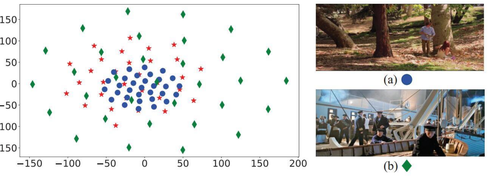
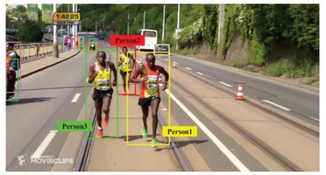


Table 2: Ablation analysis of GraphVLAD.

Method	$Q \to A$	$QA \to R$	$Q \to AR$
No-C + No-G	65.8	68.3	45.6
No-C	66.5	69.6	46.6
No-G	66.9	69.4	46.5
C + G	**67.4**	**70.6**	**47.7**

Table 3: Ablation of Directional Reasoning.

Method	$Q \to A$	$QA \to R$	$Q \to AR$
No-R	65.9	67.9	45.3
LSTM-R	64.8	67.1	43.9
GCN	66.5	69.4	46.4
D-GCN	**67.4**	**70.6**	**47.7**



What will [Person1, Person3] do if [Person2] catches up to them?

a) [Person1, Person3] will start to pick up their paces and run faster if [Person2] catches up.	97.5%
b) [Person1, Person3] will fly away.	1.2%
c) [Person1, Person3] will scream for [Person2].	1.0%
d) [Person1, Person3] hug, and follow [Person2] to their destination.	0.3%

The rationale is …

a) If [Person2] closes the distance between himself and [Person1, Person3], then [Person1, Person3] will be concerned that [Person2] is going to push ahead of them, so they will run faster because they want to keep in the lead.	99.3%
b) [Person2] looks like he is really picking up his legs to try to set himself apart from the other runners.	0.2%
c) [Person1, Person3] flank [Person2] as he walks between them.	0.4%
d) If [Person2] gets short, [Person1, Person3] will have a chance of catching him on foot.	0.1%

Why are [Person1, Person2] and [Person3] have brunch?

a) [Person1, Person2] and [Person3] are just enjoying the view while out for a walk.	10.9%
b) [Person1, Person2] and [Person3] are employees of [Person1].	0.4%
c) They enjoy reading books and consider reading to be a worthwhile hobby.	27.3%
d) They are on vacation at a resort.	61.4%

The rationale is …

a) There is sand everywhere and the ocean is in the backdrop. Many of the men have their shirts off and there are women wearing bikinis.	5.2%
b) They have just married and it is custom to enjoy a vacation after a wedding.	23.7%
c) They are dressed in clothing someone might wear on vacation. There is a lake behind them and a large building can be seen in the background.	41.9%
d) It is light food with a lot of fruits. It seems like in the afternoon which is brunch time.	29.2%

Conclusion

- We propose a cognition connectivity network for directional visual commonsense reasoning.
- A conditional GraphVLAD module is proposed to represent an image.
- Experimental results demonstrate our method is effective.