Yahong HAN, Aming WU, Linchao ZHU, Yi YANG, 2021. Visual commonsense reasoning with directional visual connections. *Frontiers of Information Technology & Electronic Engineering*, 22(5):625-637.
https://doi.org/10.1631/FITEE.2000722

Visual commonsense reasoning with directional visual connections

Key words: Visual commonsense reasoning; Directional connective network; Visual neuron connectivity; Contextualized connectivity; Directional connectivity

Corresponding author: Yahong HAN
E-mail: yahong@tju.edu.cn
ORCID: https://orcid.org/0000-0003-2768-1398

Visual commonsense reasoning (VCR)

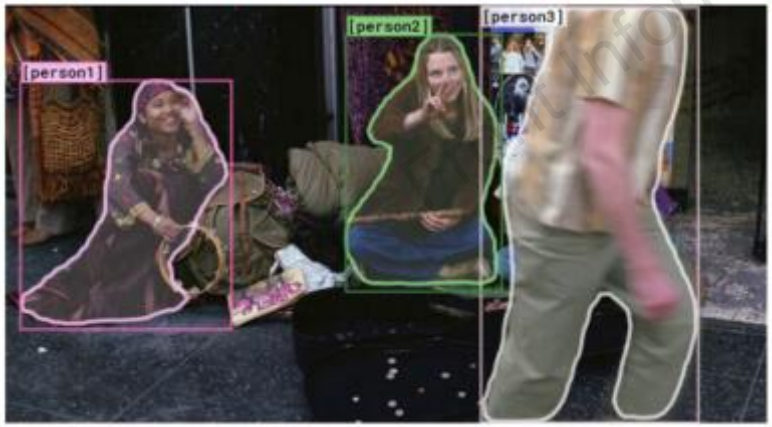
- Given an image, a list of regions, and a question, a model must **answer the question** and **provide a rationale** explaining why its answer is right.

Why is [person4] pointing at [person1]?

a) **He is telling [person3] that [person1] ordered the pancakes.**
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

I chose a) because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) **[person3] is delivering food to the table, and she might not know whose order is whose.**



How did [person2] get the money that's in front of her?

a) [person2] is selling things on the street.
b) **[person2] earned this money playing music.**
c) She may work jobs for the mafia.
d) She won money playing poker.

I chose b) because...

a) She is playing guitar for money.
b) [person2] is a professional musician in an orchestra.
c) **[person2] and [person1] are both holding instruments, and were probably busking for that money.**
d) [person1] is putting money in [person2]'s tip jar, while she plays music.

Motivation

- Recent studies on brain networks have suggested that brain function or cognition can be described as **the global and dynamic integration of local neuronal connectivity**. Such a global and dynamic integration is context-sensitive with respect to a specific cognition task.

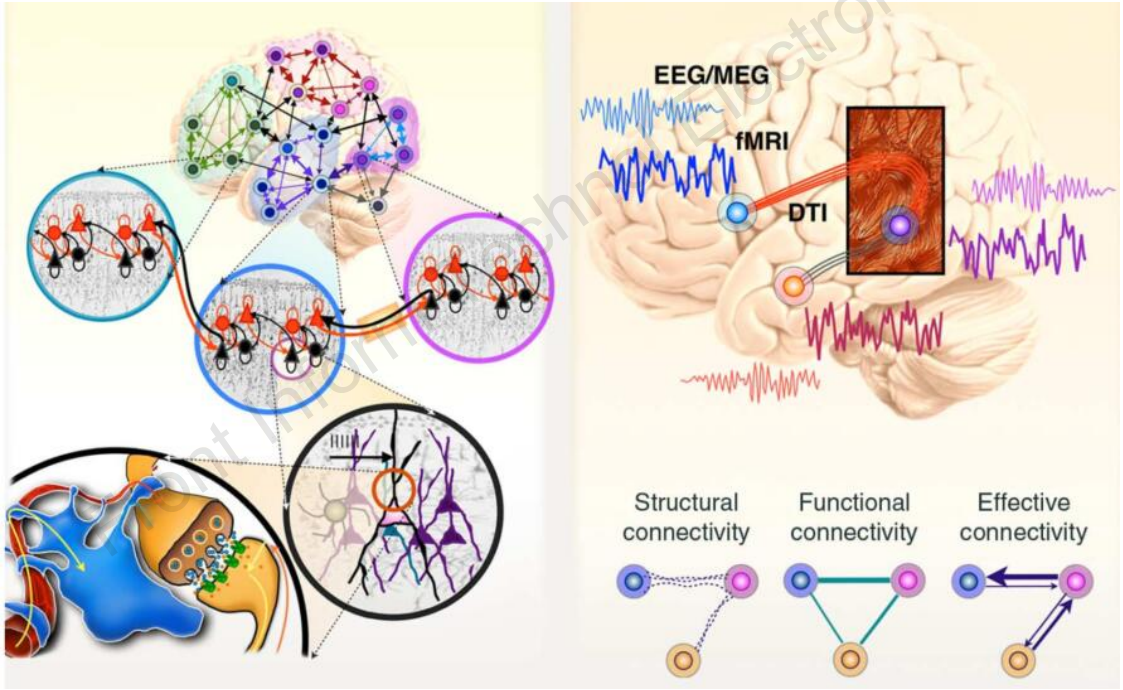


Fig. 1 The connectivity mechanism of brain neurons

(H. Park and K. Friston. Structural and functional brain networks: from connections to cognition. *Science*, 342(6158):1238411, 2013.)

Motivation (Cont'd)

- We propose a **directional connective network (DCN)** for visual commonsense reasoning. This network consists mainly of three modules, i.e., **visual neuron connectivity**, **contextualized connectivity**, and **directional connectivity for reasoning**.

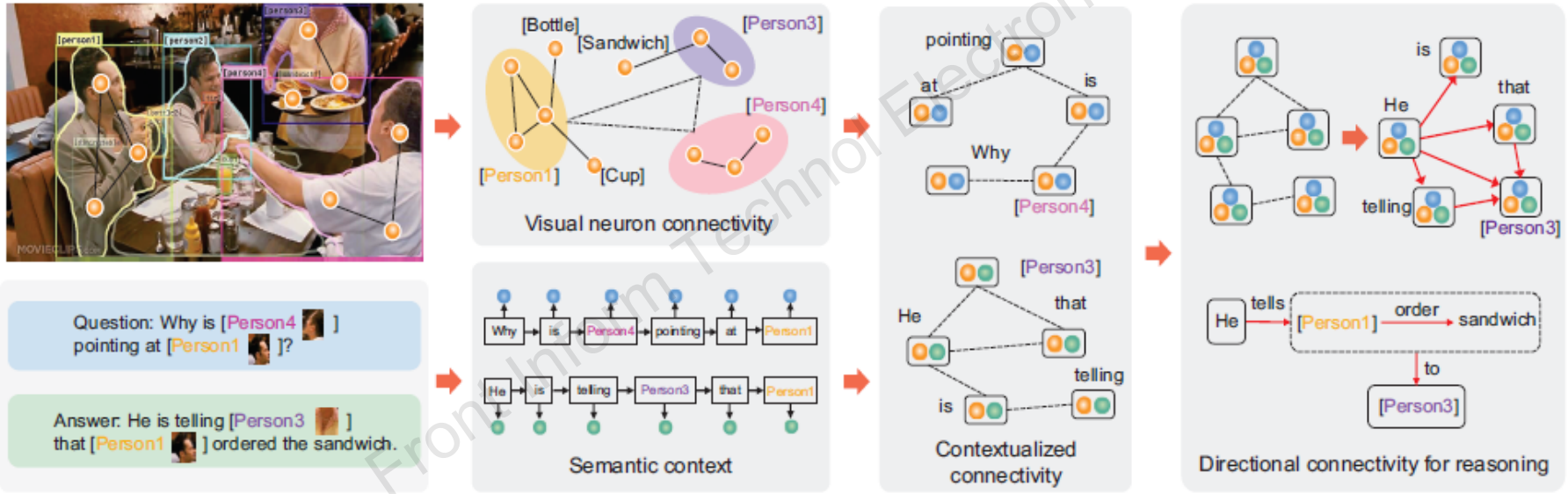


Fig. 2 Overview of our DCN method. The yellow, blue, and green circles indicate visual elements, questions, and answer representations, respectively. Our method includes mainly visual neuron connectivity, contextualized connectivity, and directional connectivity for reasoning

Directional connective network (DCN)

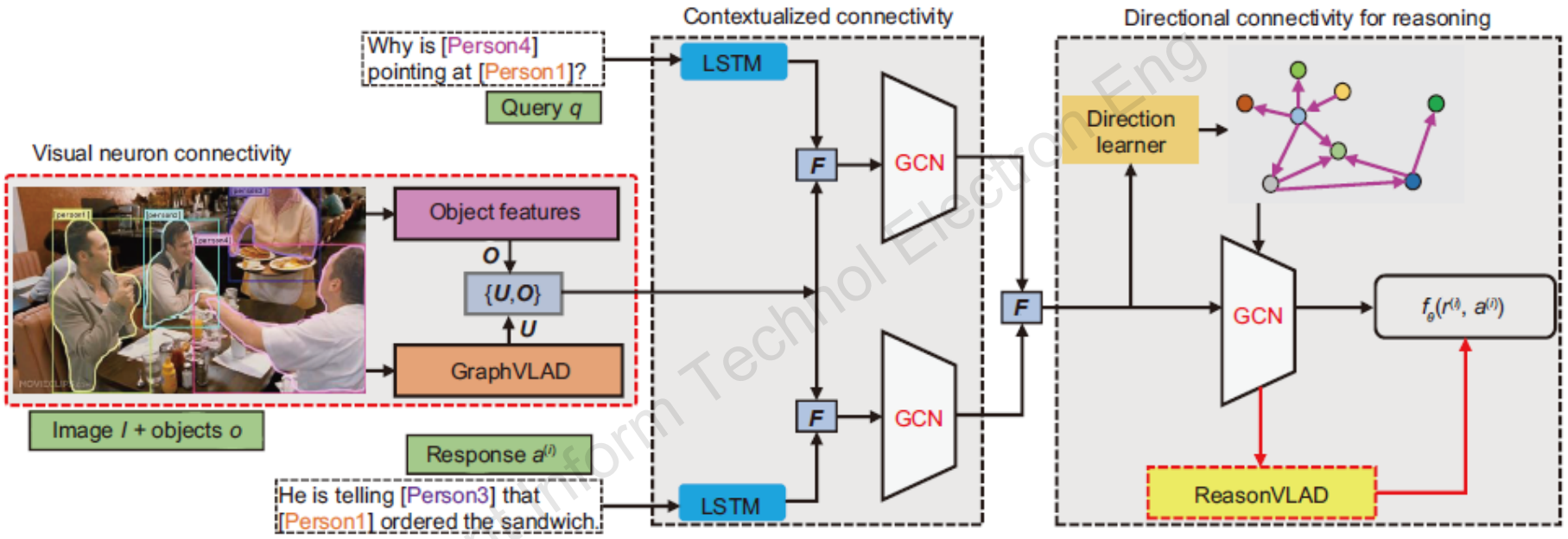


Fig. 3 Framework of the DCN method, including mainly visual neuron connectivity, contextualized connectivity, and directional connectivity for reasoning. {*U*, *O*}: the set including the output *U* of GraphVLAD and object features *O*; f_θ: the prediction function for responses (answers or rationales); ***F***: a fusion operation.

Directional connective network (DCN) (Cont'd)

- Visual neuron connectivity

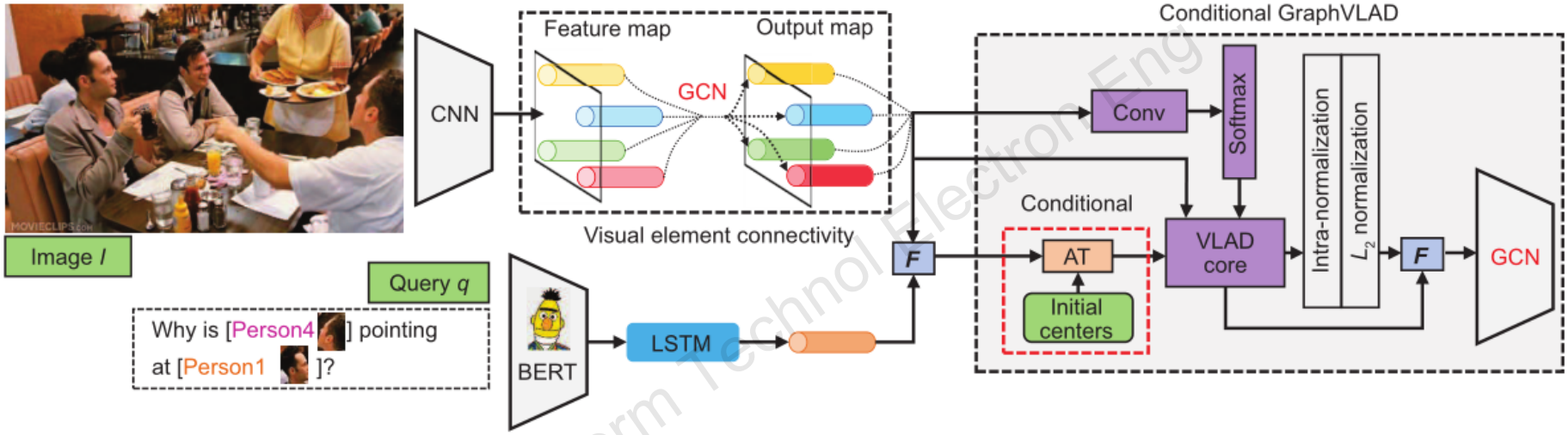


Fig. 4 Details of visual neuron connectivity. After obtaining the representation of the image and its corresponding query, we first devise a **visual element connectivity** to extract the relations of the image regions. Then, based on the fusion between the region relations and the query's representation, we devise a **conditional GraphVLAD** module to achieve a better joint representation

Directional connective network (DCN) (Cont'd)

- Visual element connectivity

  In general, there exists certain relation between objects of an image. For example, in Fig. 5, relations exist not only between elements in the same object region, but also among various objects. Obviously, capturing these relations is helpful for a thorough understanding of the entire scene.

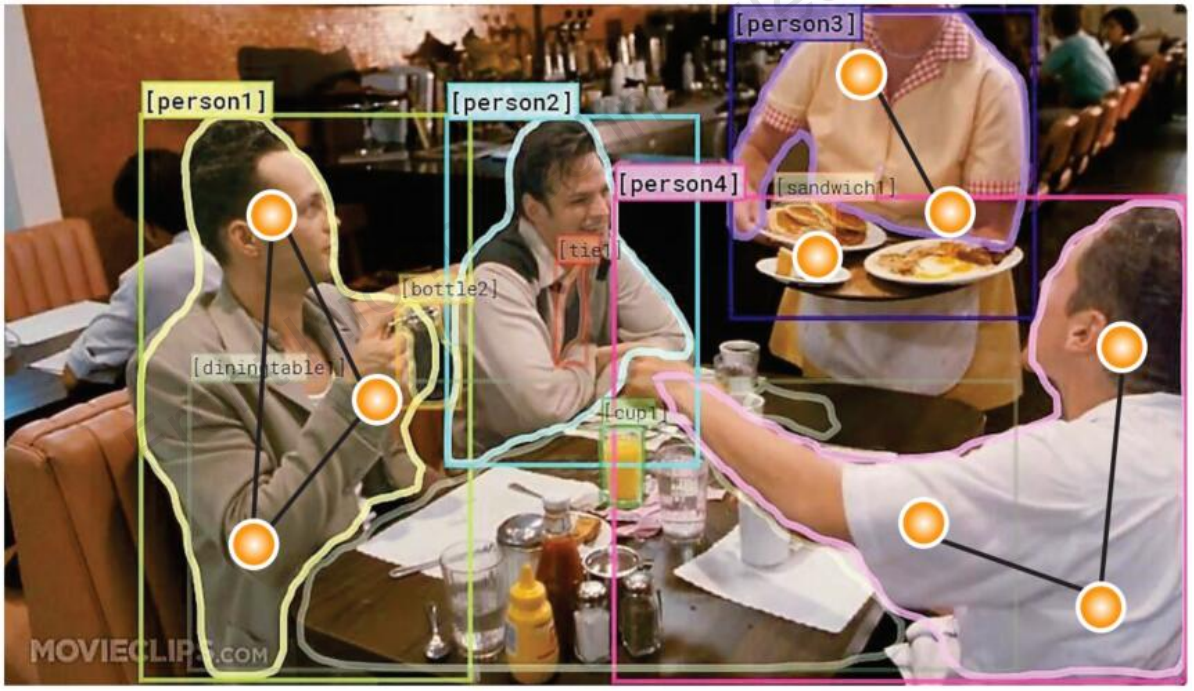


Fig. 5 Relations of visual elements

Directional connective network (DCN) (Cont'd)

- Visual element connectivity

Here, we employ **graph convolutional neural network (GCN)** to capture these relations. Specifically, we seek to construct an undirected graph $G_g = (V, \xi, \boldsymbol{A})$, where ξ is the set of graph edges to learn and $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ is the corresponding adjacency matrix.

$$\begin{cases} \boldsymbol{M} = \boldsymbol{A}\tilde{\boldsymbol{X}}, \\ \tilde{\boldsymbol{M}} = \tanh(\boldsymbol{w}_f^c * \boldsymbol{M} + \boldsymbol{b}_f^c) \odot \sigma(\boldsymbol{w}_g^c * \boldsymbol{M} + \boldsymbol{b}_g^c) \end{cases}$$

$\tilde{\boldsymbol{X}}$ is the image feature extracted by ResNet network. $\boldsymbol{A} = \mathrm{softmax}_{\mathrm{r}}\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{\mathrm{T}}\right) + \boldsymbol{I}_d$, where $\boldsymbol{I}_d$ indicates the identity matrix. Each row of the matrix $\boldsymbol{M}$ represents a feature vector of a node, which is a weighted sum of the neighboring node features of the current node.

Directional connective network (DCN) (Cont'd)

- Conditional GraphVLAD

  Since $\widetilde{\boldsymbol{M}}$ captures only relations between visual elements and does not have the capability to fully understand the image, we propose a module of conditional GraphVLAD to enhance the representation of an image.

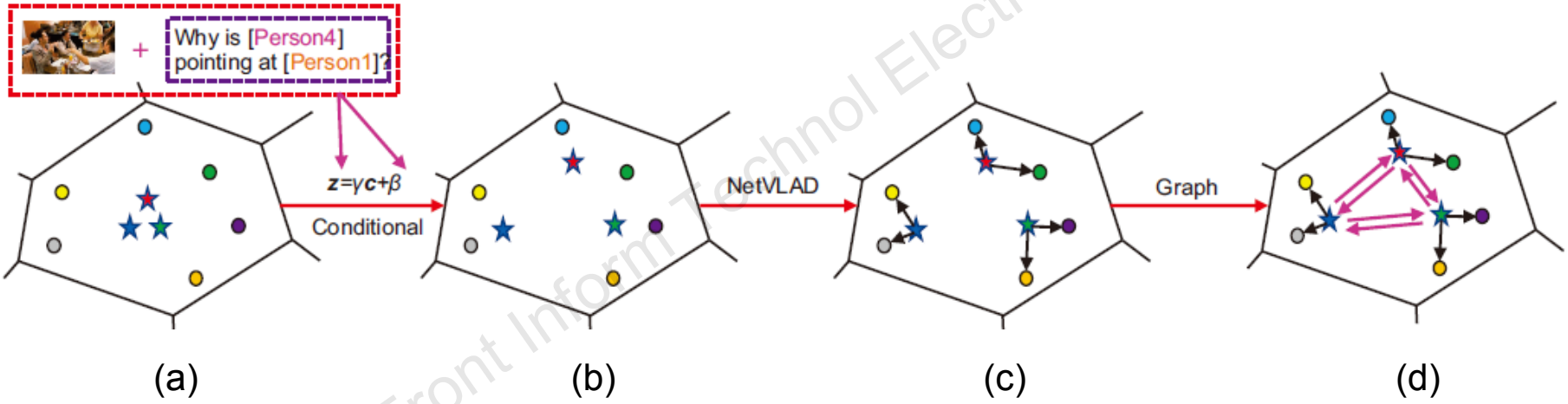


Fig. 6 Process of conditional GraphVLAD: (a) initial state of NetVLAD; (b) conditional centers after an affine transformation. Here, we use the fusion of image and question to compute the parameters γ and β. (c) and (d) show the results of NetVLAD and GraphVLAD, respectively

Directional connective network (DCN) (Cont'd)

- Conditional GraphVLAD

  We consider making an **affine transformation** for the initial centers and using these transformed centers to represent an image.

  Concretely, we first define the initial centers $\boldsymbol{C} = \{\boldsymbol{c}_i \in \mathbb{R}^n, i = 1, 2, \ldots, K\}$. Next, based on the current input query-image pairs, we make the affine transformation for the initial centers.

$$\begin{cases} \gamma = f(\langle \tilde{\boldsymbol{M}}, \tilde{\boldsymbol{Y}} \rangle) \\ \beta = h(\langle \tilde{\boldsymbol{M}}, \tilde{\boldsymbol{Y}} \rangle) \\ \boldsymbol{z}_i = \gamma \boldsymbol{c}_i + \beta \end{cases}$$

  $< a, b >$ represents the concatenation of a and b. $\boldsymbol{z}$ indicates the transformed center.

Directional connective network (DCN) (Cont'd)

- Conditional GraphVLAD

  Computation of GraphVLAD:

$$D_j = \sum_{i=1}^{N} \frac{e^{\boldsymbol{w}_j^{\mathrm{T}} \tilde{\boldsymbol{M}}_i + \boldsymbol{b}_j}}{\sum_{j'} e^{\boldsymbol{w}_{j'}^{\mathrm{T}} \tilde{\boldsymbol{M}}_i + \boldsymbol{b}_{j'}}} (\tilde{\boldsymbol{M}}_i - \boldsymbol{z}_j),$$

  where $\{\boldsymbol{w}_j\}$ and $\{\boldsymbol{b}_j\}$ are sets of trainable parameters for each center $\boldsymbol{z}_j$.

  As NetVLAD is computed based on visual elements, we consider that there should exist certain relations between NetVLAD outputs *D*. Here, we employ GCN to capture these relations. Finally, the set $\boldsymbol{S} = \{\boldsymbol{U}, \boldsymbol{O}\}$ is taken as the global representation of an image, where $\boldsymbol{U}$ indicates the output of GraphVLAD and $\boldsymbol{O}$ indicates the object features.

Directional connective network (DCN) (Cont'd)

- Contextualized connectivity

  The goal of contextualized connectivity is to not only capture the relevance between linguistic features and the global representation $\boldsymbol{S}$, but also extract deep semantic existing in sentences.

$$\begin{cases} \boldsymbol{F}_{qu} = \mathrm{softmax_r}(\tilde{\boldsymbol{Q}}\boldsymbol{U}^{\mathrm{T}}), \\ \boldsymbol{F}_{qo} = \mathrm{softmax_r}(\tilde{\boldsymbol{Q}}\boldsymbol{O}^{\mathrm{T}}), \\ \boldsymbol{Q}_U = \boldsymbol{F}_{qu}\boldsymbol{U}, \\ \boldsymbol{Q}_O = \boldsymbol{F}_{qo}\boldsymbol{O}. \end{cases}$$

  $\tilde{\boldsymbol{Q}}$ indicates the query representation extracted by an LSTM unit. Then, we take the concatenation of $\boldsymbol{Q}_U$, $\boldsymbol{Q}_O$, and $\tilde{\boldsymbol{Q}}$ as $\boldsymbol{Q}_F$. Here, we obtain only sequential features, rather than **the structural information** which is helpful for a better understanding of the sentence semantic. We consider using GCN to extract the structural information. Finally, we obtain the query representation $\boldsymbol{Q}_g$ and response representation $\boldsymbol{A}_g$.

Directional connective network (DCN) (Cont'd)

- Directional connectivity with ReasonVLAD

Directional information is an important clue for cognitive reasoning. Using directional information could improve the accuracy of reasoning. We propose a **semantic direction based GCN** for reasoning.

$$\begin{cases} \boldsymbol{D}_{qa} = \phi(\boldsymbol{E}_{qa}) \\ \boldsymbol{G}_t = \boldsymbol{D}_{qa}\boldsymbol{D}_{qa}^{\mathrm{T}} \\ \boldsymbol{D}_t = \mathrm{sign}(\boldsymbol{G}_t) \\ \boldsymbol{V}_e = \mathrm{softmax}_{\mathrm{r}}(\mathrm{abs}(\boldsymbol{G}_t)) \end{cases}$$

Next, based on the output $\boldsymbol{D}_t$ of the sign function, we compute the adjacency matrix.

$$\begin{cases} \boldsymbol{H} = \boldsymbol{D}_t \odot \boldsymbol{V}_e + \boldsymbol{I}_d, \\ \boldsymbol{M}_t = \boldsymbol{H}\,\boldsymbol{E}_{qa}, \\ \boldsymbol{R}_t = \tanh(\boldsymbol{w}_f^r * \boldsymbol{M}_t + \boldsymbol{b}_f^r) \odot \sigma(\boldsymbol{w}_g^r * \boldsymbol{M}_t + \boldsymbol{b}_g^r), \end{cases}$$

where ***H*** indicates the adjacency matrix.

Directional connective network (DCN) (Cont'd)

- Directional connectivity with ReasonVLAD

  After obtaining the output of the directional connectivity module, to enhance the information association of different modals and improve the reasoning ability, we design a **ReasonVLAD** module.

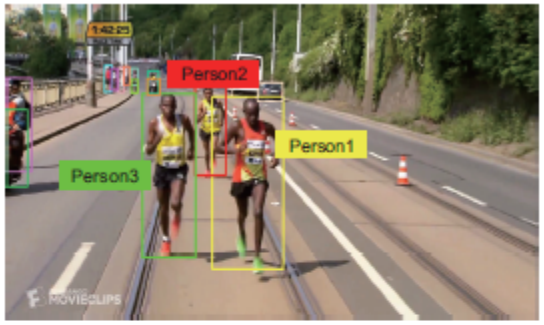
  Concretely, we first define the initial centers $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i \in \mathbb{R}^n,\ i = 1, 2, \ldots, K\}$. Next, the processes of ReasonVLAD are shown as follows:

$$\boldsymbol{\Phi}_j = \sum_{i=1}^{J} \frac{e^{\boldsymbol{w}_j^{\mathrm{T}} \tilde{\boldsymbol{S}}_i + \boldsymbol{b}_j}}{\sum_{j'} e^{\boldsymbol{w}_{j'}^{\mathrm{T}} \tilde{\boldsymbol{S}}_i + \boldsymbol{b}_{j'}}} (\tilde{\boldsymbol{S}}_i - \boldsymbol{\theta}_j)$$

  The advantage of ReasonVLAD is mainly that, with the help of the learned centers, this module can sufficiently capture the fused information, which reduces the loss of related information and then improves the reasoning ability.

Experimental results

We evaluate our method on the VCR dataset. This dataset contains 2.9×10^5 pairs of questions, answers, and rationales, over 1.1×10^5 unique movie scenes.

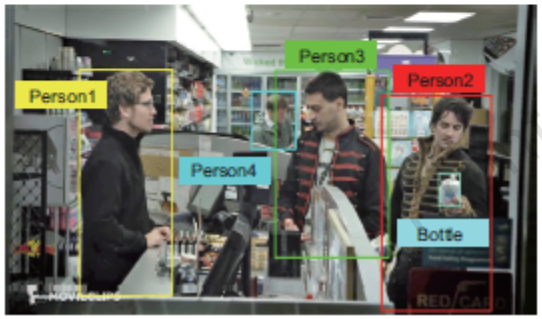


What will [Person1, Person3] do if [Person2] catches up with them?

a) [Person1, Person3] will start to pick up their paces and run faster if [Person2] catches up. 98%
b) [Person1, Person3] will fly away.
c) [Person1, Person3] will scream for [Person2].
d) [Person1, Person3] hug, and follow [Person2] to their destination.

The rationale is ...

a) If [Person2] closes the distance, then [Person1, Person3] will be concerned that [Person2] is going to push ahead of them, so they will run faster. 99%
b) [Person2] looks like he is really picking up his legs to try to set himself apart from the other runners.
c) [Person1] flanks [Person2] as he walks between them.
d) If [Person2] gets short, [Person1, Person3] will have a chance of catching him on foot.



What does [Person3] do next?

a) [Person3] pays for the items he wants. 58%
b) He walks out the building.
c) Put his hands up slowly to show he is not reaching for a weapon.
d) After finishing speaking to [Person2], he moves to the chair behind him, and sinks into the chair while running his hand over his face in disbelief.

The rationale is ...

a) [Person1] is a cashier who you give money to in exchange for products. 69%
b) [Person1] is breaking into a mine where such items as valuable metals or buried treasure would be located.
c) [Person2] wants to purchase new boots but does not know the proper size.
d) [Person1] is wearing a uniform and standing behind a register.

Fig. 7 Two qualitative examples for DCN. Correct choices are highlighted in **blue**

Experimental results (Cont'd)

- Performance of our method

Table 1 Performance of our DCN model on the VCR dataset

Model	Accuracy (%)		
	$Q \rightarrow A$	$QA \rightarrow R$	$Q \rightarrow AR$
VisualBERT	70.8	73.2	52.2
ViLBERT	72.4	74.5	54.0
Unicoder-VL	72.6	74.5	54.5
VL-BERT	73.8	74.4	55.2
Revisited VQA	39.4	34.0	13.5
BottomUpTopDown	42.8	25.1	10.7
MLB	45.5	36.1	17.0
MUTAN	44.4	32.0	14.6
R2C (baseline)	63.8	67.2	43.1
DCN	**67.6**	**70.7**	**47.9**

Experimental results (Cont'd)

- Visualization analysis

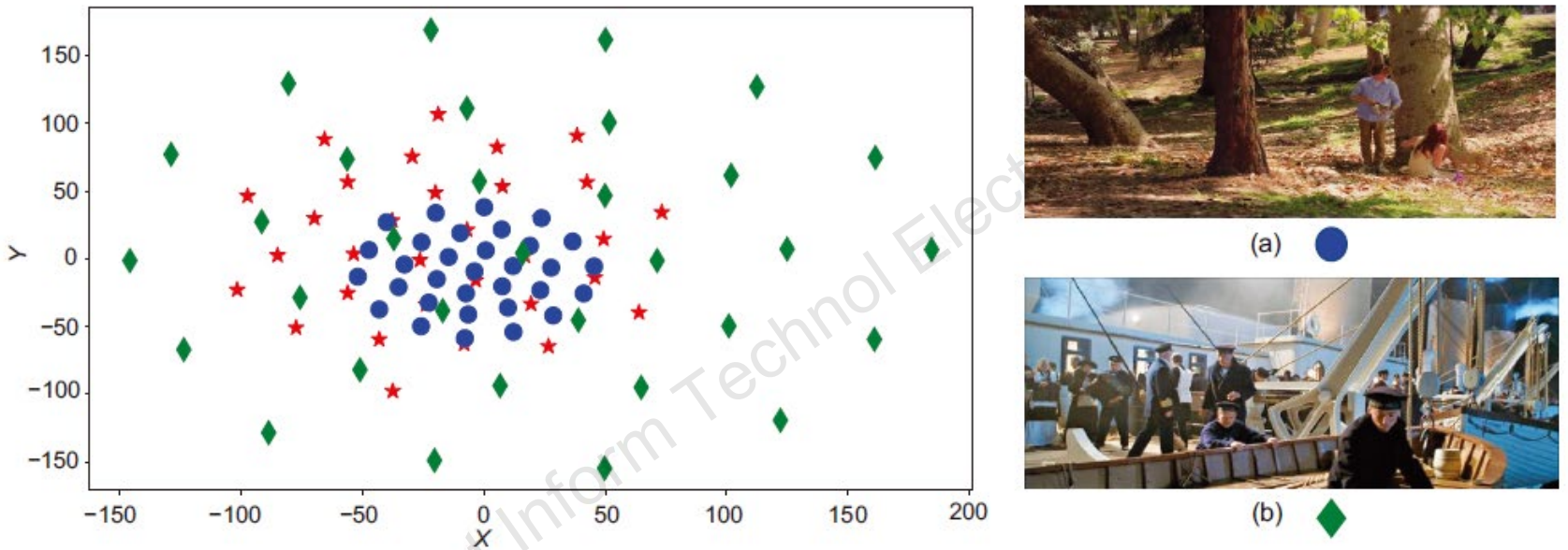


Fig. 8 t-SNE plot of conditional centers. Here, the red stars, blue circles, and green rhombuses indicate the initial centers and two different conditional centers, respectively. (a) and (b) are the computational conditions of blue and green centers, respectively

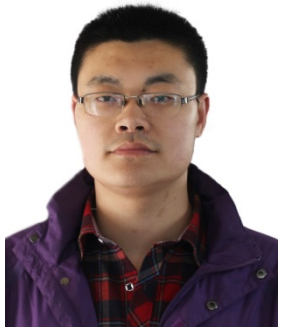
Conclusions

- In this paper, we have proposed a directional connective network for visual commonsense reasoning, which includes mainly three graph-based modules, i.e., visual neuron connectivity, contextualized connectivity, and directional connectivity with ReasonVLAD.

- Visual neuron connectivity promotes a thorough understanding of visual content. Contextualized connectivity captures the relevance between linguistic features and global representations. Directional connectivity enhances the reasoning ability based on learned direction information.

- Experimental results and visualization analysis demonstrated the effectiveness of the proposed method.

Yahong HAN received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2012. He is currently a full professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. From Nov. 2014 to Nov. 2015, he visited Prof. Bin YU's group at UC Berkeley as a visiting scholar. His current research interests include multimedia analysis, computer vision, and machine learning.



Aming WU received the Ph.D. degree from Tianjin University, Tianjin, China, in 2021. He joins Xidian University as a pre-tenured associate professor at the School of Electronic Engineering in Jan. 2021. His current research interests include computer vision, multimedia analysis, and machine learning.

Linchao ZHU received the Ph.D. degree in computer science from University of Technology Sydney, Australia, in 2019. He received the B.E. degree from Zhejiang University, China, in 2015. He is currently a lecturer in the Australian Artificial Intelligence Institute, University of Technology Sydney, Australia. His research interest is video analysis and understanding.

Yi YANG received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a professor with the University of Technology Sydney, Australia. He was a postdoctoral researcher in the School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis, and video content understanding.