Two New Datasets and Tasks on Visual Reasoning

Peng Wang

School of Computing and Information Technology

University of Wollongong

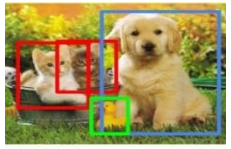
Fast thinking vs. Slow thinking

Fast thinking

- Object recognition

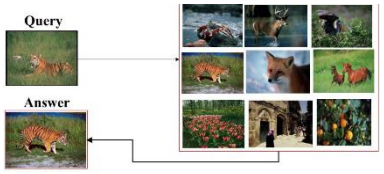
CAT

- Object detection



CAT, DOG, DUCK

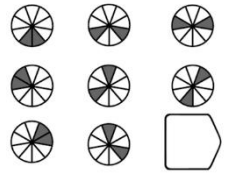
- Image retrieval



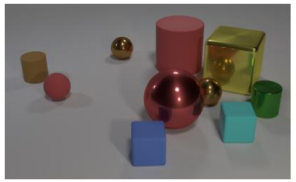
- Speech recognition

Slow thinking

- Raven's Progressive Matrices

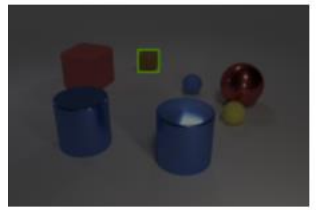


- VQA (CLEVER)



Q: Are there an equal number of large things and metal spheres?

- Referring Expression (CLEVER-Ref)

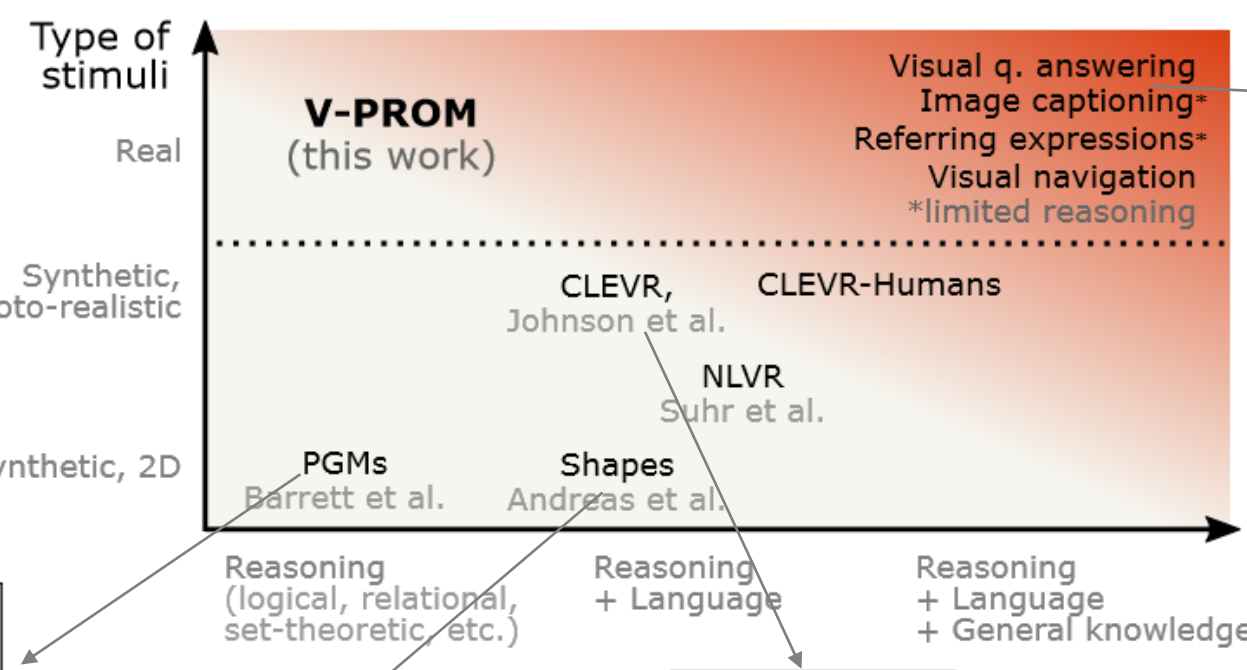


E: Any other things that are the same shape as th fourth one of the rubber thing(s) from right

- VQA (GQA)

Q: Are the napkin and the cup the same color?

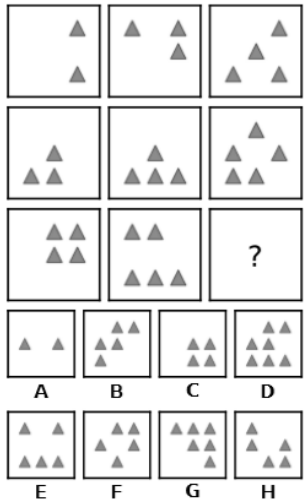
Reasoning tasks: type of stimuli vs. skills required



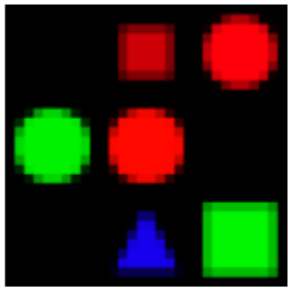
Increasing real-world applicability (Unsolved tasks)

cvpr19

Q: Are the napkin and the cup the same color?

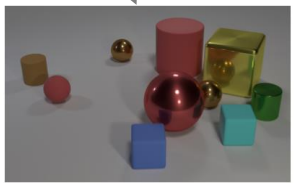


ICML18



cvpr16

is there a red shape above a circle?

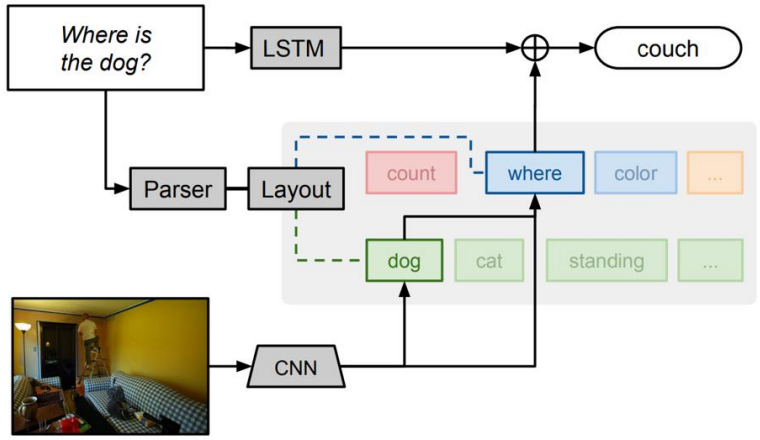


cvpr17

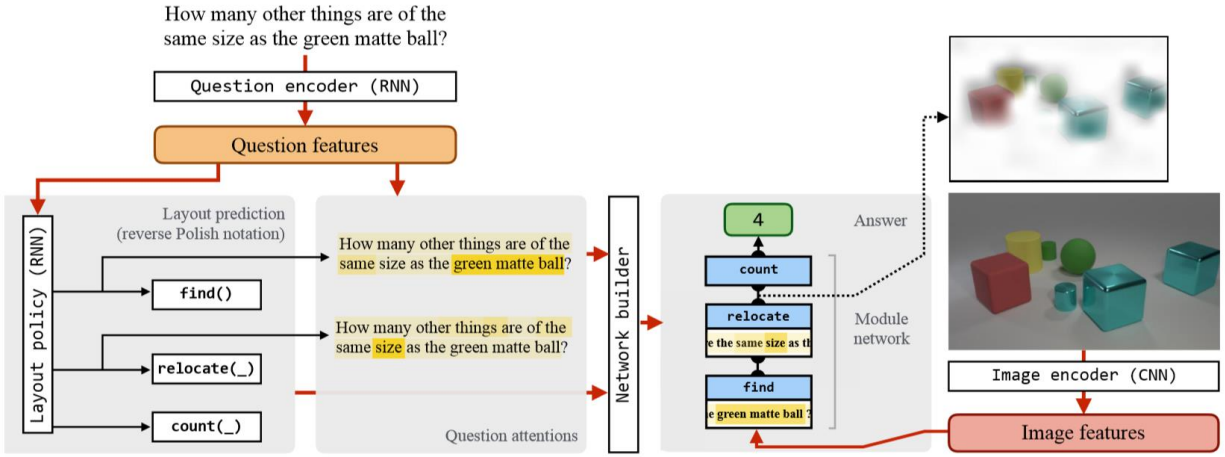
Are there an equal number of large things and metal spheres?

Typical solutions: function program

- Module networks [1]
  - Custom architecture for each question
  - Use existing linguistic tool to convert question into module sequence



- End-to-end module networks [2]
  - Implement question→program sequence using seq-to-seq learning
  - Require program function labelling

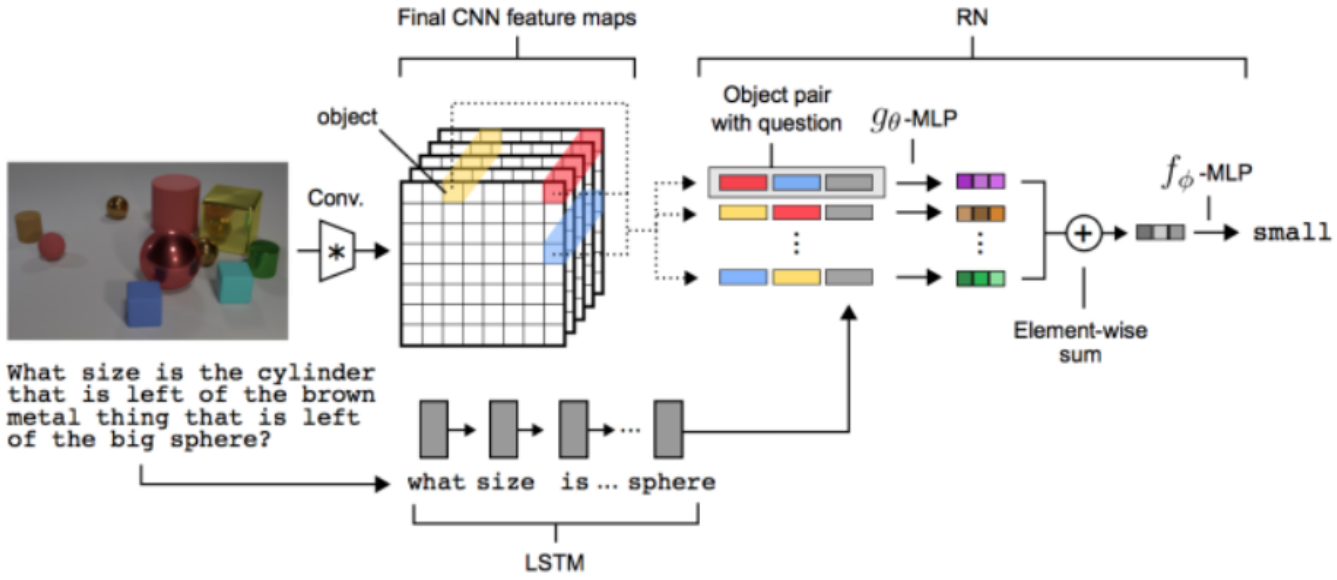


[1]. J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In CVPR, 2016.

[2]. Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross B Girshick. Inferring and executing programs for visual reasoning. In ICCV, 2017.

Typical solutions: relation network

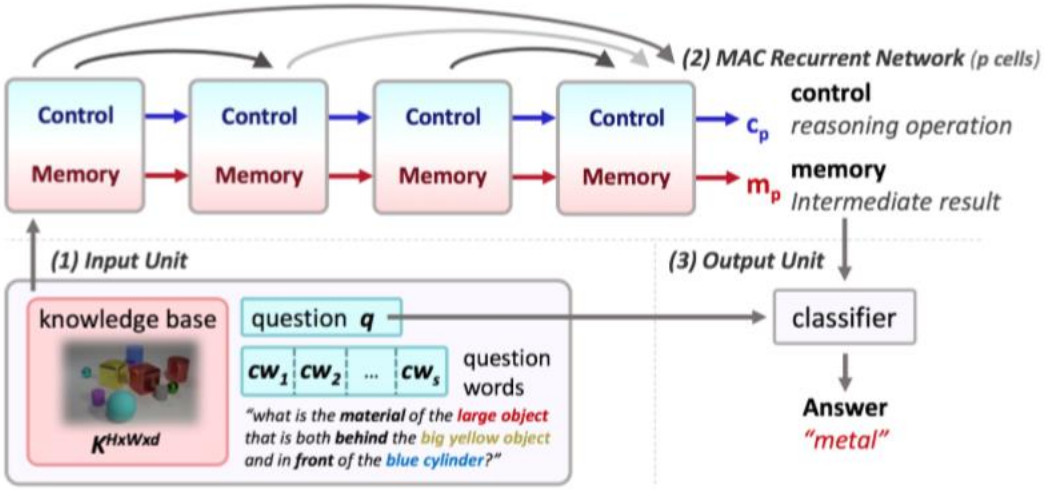
- Relation network [3]
  - Use paired convolutional features for relational reasoning
  - No additional supervision but better performance
  - Generalize to more complex visual stimuli and semantic relationships?



[3]. A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In NIPS, 2017.

Typical solutions: iterative attention-based reasoning

- Memory, Attention, and Composition (MAC) [4]
  - A series of attention-based reasoning steps, each performed by a MAC cell.
  - Fully differentiable
  - No additional supervision
  - Better performance



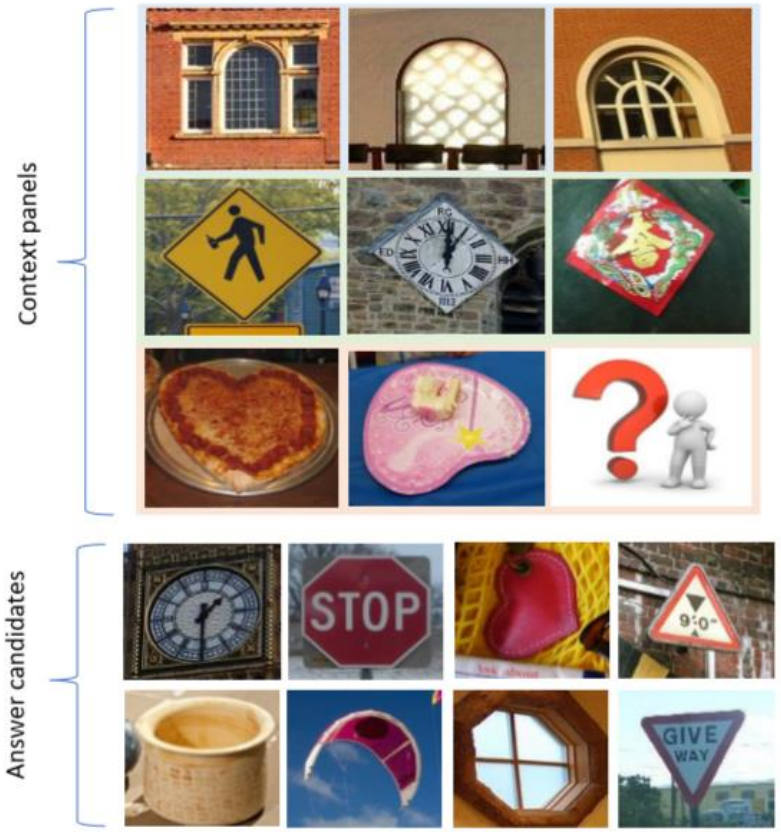
[4]. Hudson, D.A., Manning, C.D. Compositional attention networks for machine reasoning. In ICLR, 2018.

V-PROM: A Benchmark for Visual Reasoning Using Visual Progressive Matrices

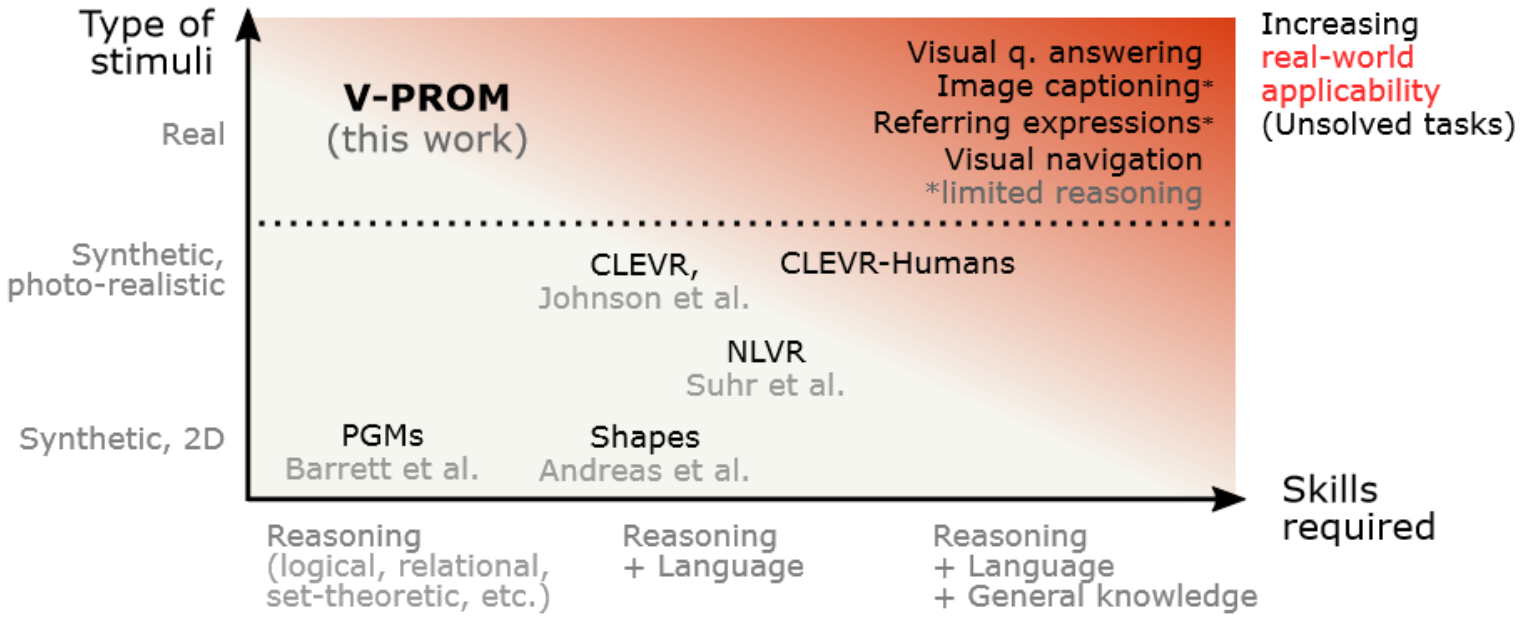
Damien Teney*, **Peng Wang***, Jiewei Cao, Lingqiao Liu, Chunhua Shen, Anton ven den Hengel

Task definition

- Each test instance is a matrix of 3 x 3 images, the task is to identify the correct candidate for the 9th image from a set of candidates.
- The task requires identifying a plausible explanation for the provided triplets of images, i.e. a relation that could have generated them.
- The task focuses on fundamental visual properties and relationships such as logical and counting operations over multiple images.



Datasets and tasks for visual reasoning: Recap



Guess the answer?

Generating descriptions of task instances

- Each instance is a visual reasoning matrix (VRM). $I_{i,j}$ denotes the j^{th} image of the i^{th} row.
- Each image describes one visual element $a_{i,j} = \phi(I_{i,j})$ which can be an attribute, an object, or object count. $v(I_{i,j}) \in \{A, O, C\}$ denotes the type of visual element the image corresponds to, i.e. attributes, objects, or object counts.
- Each VRM represents on type of visual elements and one specific type of relationship $r \in \{\text{And}, \text{Or}, \text{Union}, \text{Progression}\}$.

$$\text{And} : \phi(I_{i,3}) = \phi(I_{i,j}),\ \forall j \in 1, 2.$$

$$\text{Or} : \phi(I_{i,3}) = \phi(I_{i,1}) \text{ or } \phi(I_{i,3}) = \phi(I_{i,2}).$$

$$\text{Union} : \{\phi(I_{1,j}) \forall j\} = \{\phi(I_{2,j}) \forall j\} = \{\phi(I_{3,j}) \forall j\}.$$

$$\text{Progression} : v(I_{i,c}) = C,\ \forall i, j;\ \text{and}\ \phi(I_{i,t+1}) - \phi(I_{i,t}) = \phi(I_{j,t+1}) - \phi(I_{j,t})\ \forall i, j, t \in 1, 2$$

Mining images from Visual Genome

- Desired principle:
  - Richness: diversity of visual elements, and of the images representing each visual element
  - Purity: constrain the complexity of the image
  - Visual relatedness: properties that have a clear visual depiction
  - Independence: exclude objects that frequently co-occur with other objects, e.g. *sky, road, water*.
- Collect data using VG's region-level annotations of categories, attributes, and natural language descriptions.

	Object attributes	Human attributes	Object categories	Object counts
Nb. visual elements	84	38	346	10
Nb. images	36,750	12,249	82,905	11,730
Nb. task instances	45,000	45,000	45,000	100,000[2]

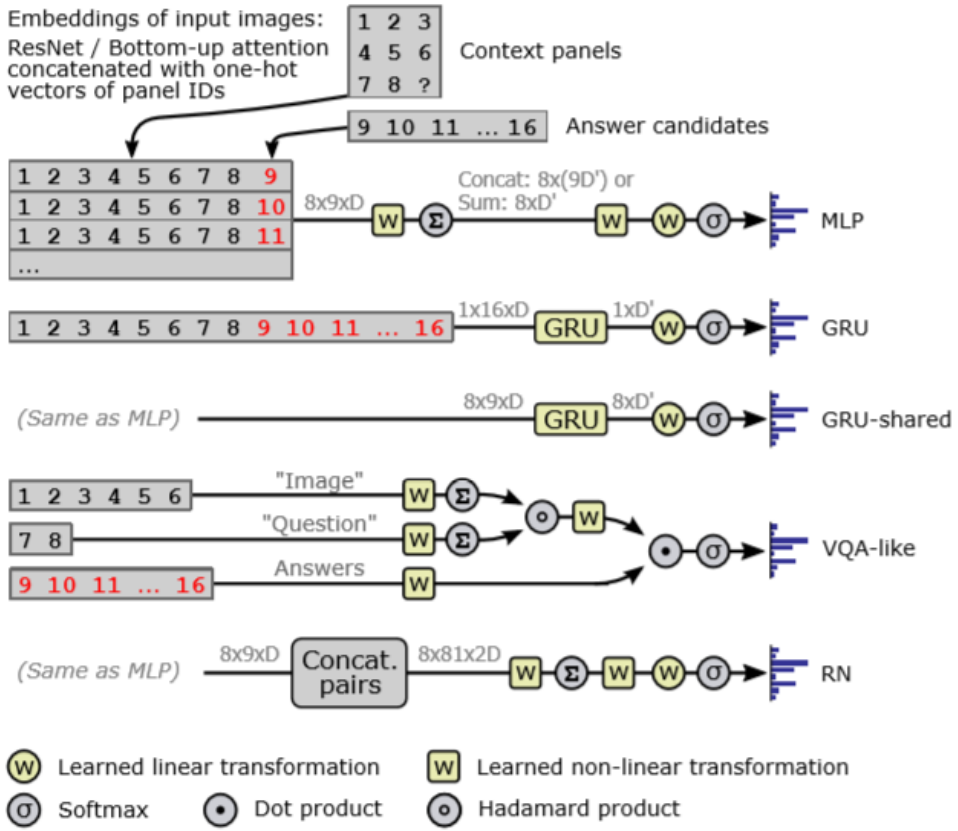
Table 1. Statistics of the V-PROM dataset.

Data splits to measure generalization

- Neural: The training and test sets are both sampled from the whole set of relationships and visual elements.
- Interpolation/extrapolation: These two splits evaluate generalization for counting. Counts (1,3,5,7,9)/(1,2,3,4,5) are used for training and counts (2,4,6,8,10)/(6,7,8,9,10) are used for testing.
- Held-out attributes/objects: A set of attributes/objects are held-out for testing only.
- Held-out pairs of relationships/attributes: A subset of relationship/attributes are held-out for testing only.
- Held-out pairs of relationships/objects: For each type of relationship, 1/3 of objects are held-out.

Evaluated models

- Each image is passed through a pretrained ResNet101 or Bottom-Up Attention Network to extract visual features.
- The feature maps are average-pooled and L2 normalized.
- The vector of each image is concatenated with a one-hot representation of index 1-16.

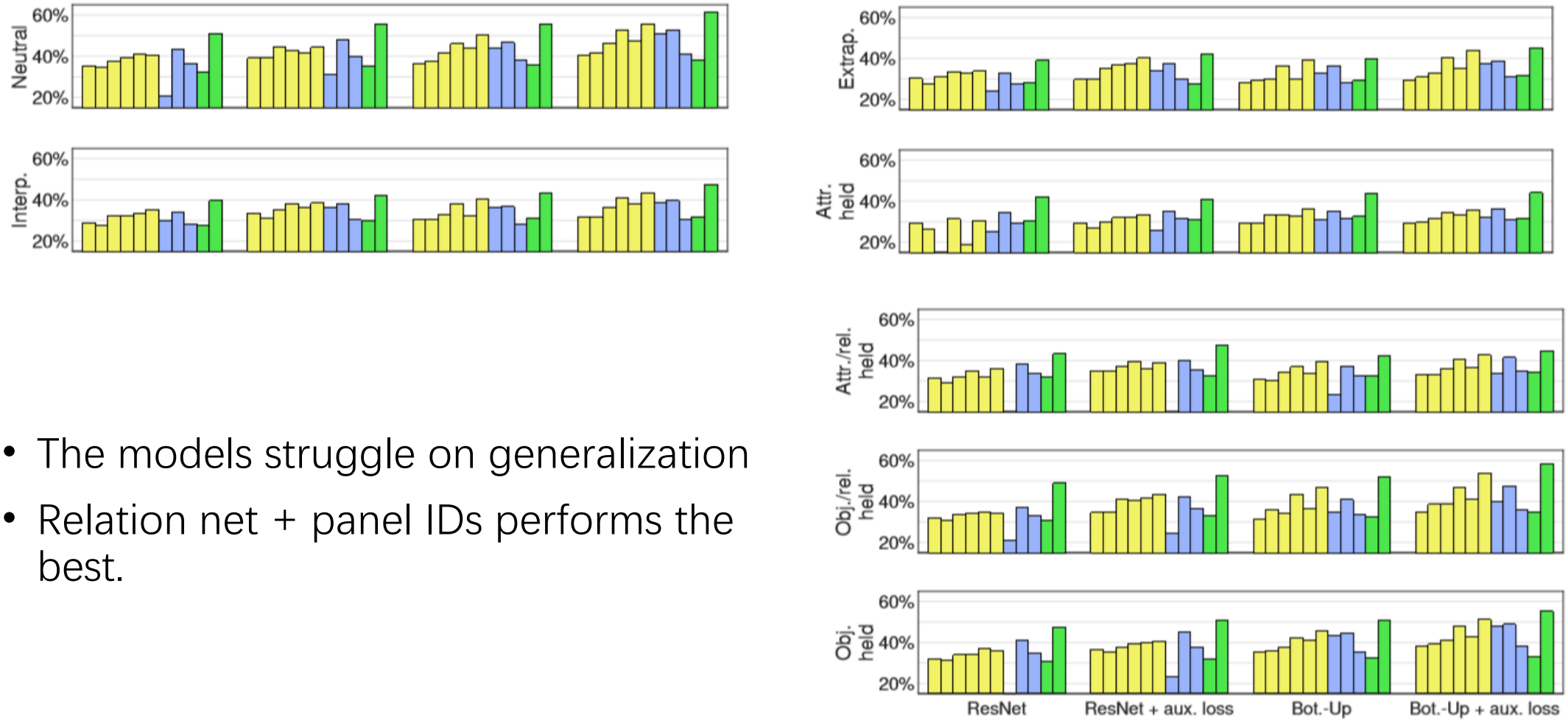


Performance comparison

	ResNet	ResNet +aux.loss	B.-up	B.-up +aux.loss
Human evaluation		77.8		
RN with shuffled inputs	12.5	12.5	12.5	12.5
MLP-sum-6 layers	40.7	44.5	50.4	55.7
GRU-shared	43.4	48.2	46.7	52.7
VQA-like	36.7	39.7	37.9	41.0
Relational network (RN)	**51.2**	**55.8**	**55.4**	**61.3**

- Bottom-Up features have better performance;
- Relational network performs the best;
- Auxiliary loss helps;
- Humans tend to use high-level semantics to infer the answer, which harm the performance.

Performance comparison on different splits



- The models struggle on generalization
- Relation net + panel IDs performs the best.

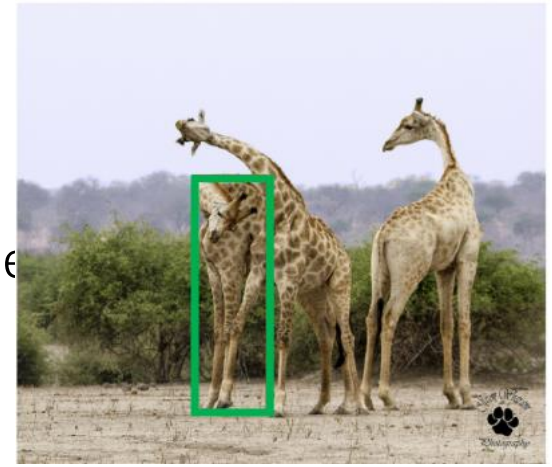
Cops-Ref: A new Dataset and Task on Compositional Referring Expression Comprehension

Zhenfang Chen**, Peng Wang**, Lin Ma, Kwan-Yee K. Wong, Qi Wu

Introduction: Task Description

- Referring expression comprehension
  - Referring expression comprehension (REF) aims at identifying a particular object in a scene by a natural language expression.

First giraffe on le



[Yu *et al.*, ECCV 16]

- **Applications**
  - Visual Question Answering;
  - Text Based Image retrieval;
  - Description Generation;
  - …

Introduction: Limitations of current datasets

Current datasets:

RefCOCO, RefCOCO+, RefCOCOg and CLEVR-Ref+

- Limitations
  - Their expressions are short, typically describing only some simple distinctive properties of the object.
  - Their images contain limited distracting information.
  - Mainly Evaluate the ability of objection recognition, attribution recognition and simple relation detection.
  - Fail to provide an ideal test bed for evaluating the reasoning ability of the REF models.

Introduction: Our Task and dataset

- Compositional Referring Expression Comprehension
  - The task requires a model to identify a target object described by a compositional referring expression from a set of images including not only the target image but also some other images with varying distracting factors as well.

- Query expression: The cat on the left that is sleeping and resting on the white towel.

(a) The image with the target "cat"

(b) Distractors of different categories

(c) Distractors with "cat"

(d) Distractors with "sleeping cat"

(e) Distractors with "cat" and "towel"

Cops-Ref Dataset

- To better evaluate the reasoning ability of the REF models, the Cops-Ref dataset has two main features:
  - Flowery and compositional expressions, requiring complex reasoning ability to understand;
  - It includes controlled distractors with similar visual properties to the referent.

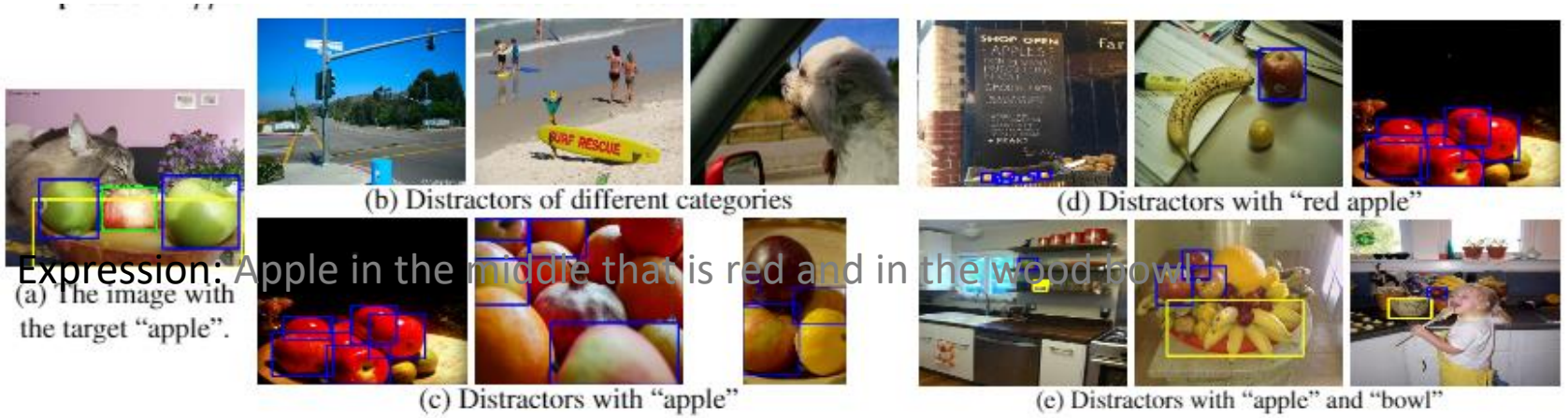
- The construction of the dataset mainly includes,
  - Expression engine
  - Discovery of distracting images

Cops-Ref Dataset: Expression engine

- **Expression engine** aims to generate grammatically correct, unambiguous and flowery expressions with various compositionality for each of the described regions. We propose to generate expressions from scene graphs based on some **expression logic forms**.

Cops-Ref Dataset: Distractor discovery

- **Introducing distracting images** provides more complex visual reasoning context, reduces dataset bias.



Cops-Ref Dataset

- Dataset statistics
  - 148k expressions on 75k images making our dataset the current largest real-world image dataset for referring expressions.
  - The average length of the expressions is 14.4 and the size of the vocabulary is 1,596.

	Object Cat.	Att. Num.	Rel. Num.	Exp. length	Cand Num.	Cat Cand Num.
refCOCO	80[1]	-	-	3.5	10.6	4.9
refCOCOg	80	-	-	8.5	8.2	2.6
CLEVR-Ref+	3	12	5	22.4	-	-
Cops-Ref	508	601	299	14.4	262.5	20.3

- Most frequent categories, attributes and relations.

(a) The most frequent object names (b) The most frequent attributes (c) The most frequent relations

Figure 2: The most frequent object names, attributes and relations of Cops-Ref. The size of words indicates frequency.

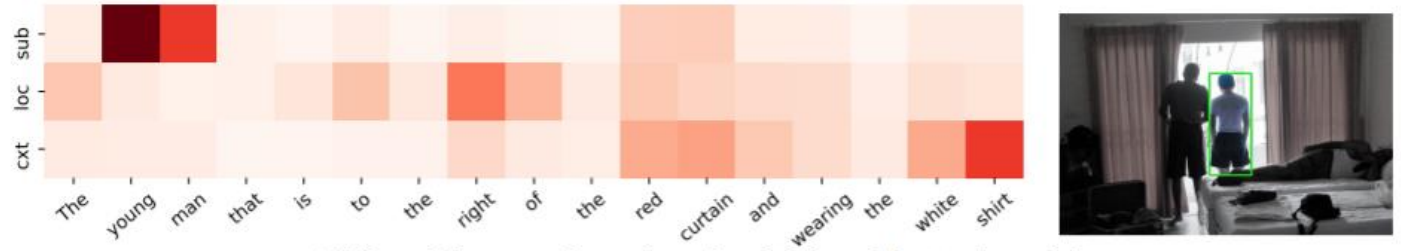
Methods: Modular hard mining strategy

- MattNet estimates matching score between expression q and the j-th r_j ,
  - $\mathrm{s}(r_j|q) = \sum_{md} w^{md} \mathrm{s}(r_j|q^{md})$,
    Where $md \in \{\mathrm{sub}, \mathrm{loc}, \mathrm{cxt}\}$.
- Ranking Loss:
  - $L_{rank} = \sum_m ([\Delta - \mathrm{s}(r_m|q_m) + \mathrm{s}(r_m|q_n)]_+ + [\Delta - \mathrm{s}(r_m|q_m) + \mathrm{s}(r_o|q_m)]_+)$,

  Where r_o and q_n are other random unaligned regions and expressions in the same image.
- Mining possibility:
  - $s_{m,n}^{md} = f(q_m^{md}, q_n^{md})$,
  - $p_{m,n}^{md} = \frac{\exp(s_{m,n}^{md})}{\sum_{n=1,n\neq m}^{n=N_C} \exp(s_{m,n}^{md})}$,
- Mining Loss:
  - $L_{rank} = \sum_m \sum_{md} (\left[\Delta - \mathrm{s}(r_m|q_m) + \mathrm{s}\left(r_m|q_n^{md}\right)\right]_+ + [\Delta - \mathrm{s}(r_m|q_m) +$

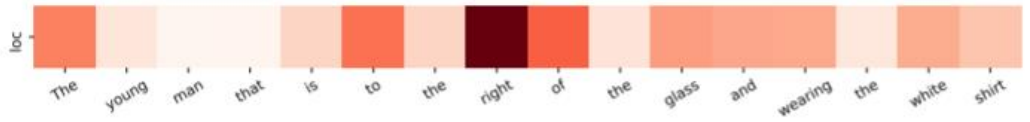
Methods: Modular hard mining strategy

A typical mining example of modular hard mining strategy

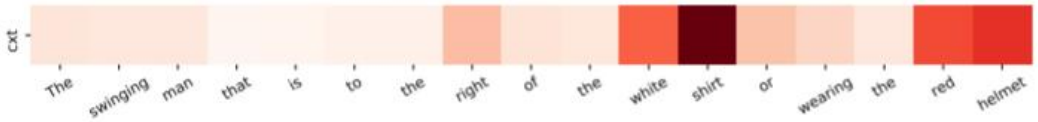


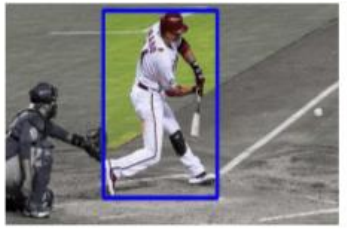
(a) The positive expression-region pair and their module attentive weight.

(b) The hard mining expression-region pair of the *sub* module.



(c) The hard mining expression-region pair of the *loc* module.





(d) The hard mining expression-region pair of the *cxt* module.

Experiments: set up

- Evaluation Setting
  - **Full** denotes the case when all the distractors are added while **WithoutDist** denotes no distractor is added. **DiffCat**, **Cat** and **Cat&attr**, respectively, represent the cases when certain type of distractors are added.

- Methods
  - GroundeR: a simple CNN-LSTM model for referring expression;
  - MattNet: one of the most popular REF models;
  - CM-Att-Erase: model with the best performance;
  - MattNet-Mine: MattNet with the proposed hard mining training strategy.

Experiments: performance comparison

Method	Full	DiffCat	Cat	Cat&attr	Cat&cat	WithoutDist
Chance	0.4	1.7	1.8	1.9	1.7	6.6
GroundeR [30]	19.1	60.2	38.5	35.7	38.9	75.7
Deaf-GroundeR	2.2	7.7	7.9	8.0	8.0	27.1
Shuffle-GroundeR	13.1	41.8	28.6	27.2	27.6	58.5
Obj-Attr-GroundeR	15.2	53.1	32.6	29.6	32.7	68.8
MattNet-refCOCO	8.7	22.7	17.0	16.7	18.9	42.4
MattNet [38]	26.3	69.1	45.2	42.5	45.8	77.9
CM-Att-Erase [23]	28.0	**71.3**	47.1	43.4	48.4	**80.4**
SCAN [18]+MattNet	18.8	-	-	-	-	-
MattNet-Mine	**33.8**	70.5	**54.4**	**46.8**	**52.0**	78.4

- Existing REF models achieve unsatisfactory performance when distractors are added;
- Existing REF models mainly rely on object and attribution recognition to ground the expression;
- The proposed MattNet-Net can constantly improve the performance especially when the distractors are added.

Experiments: ablation

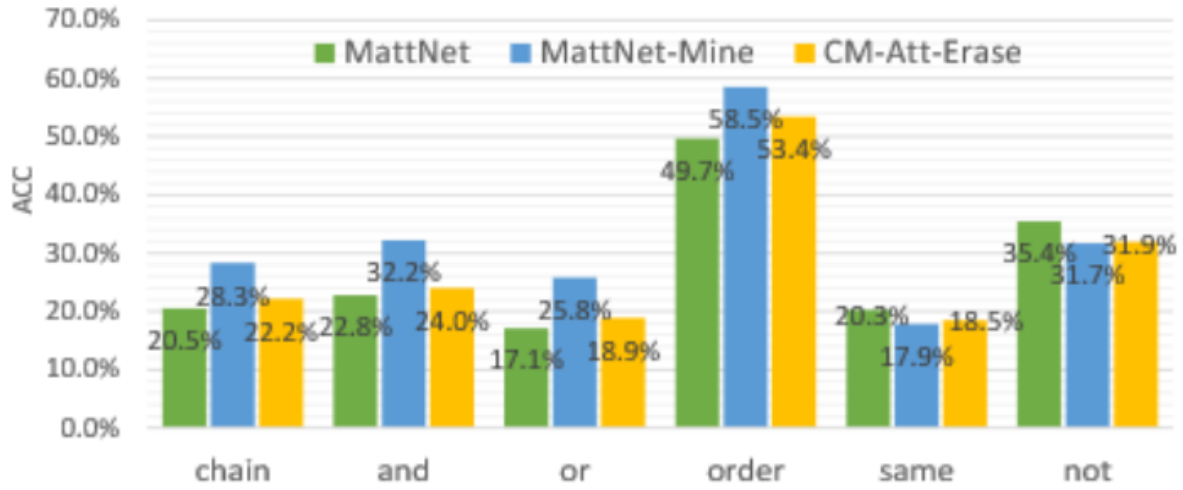


Figure 3: Accuracy of expressions of different logic form.

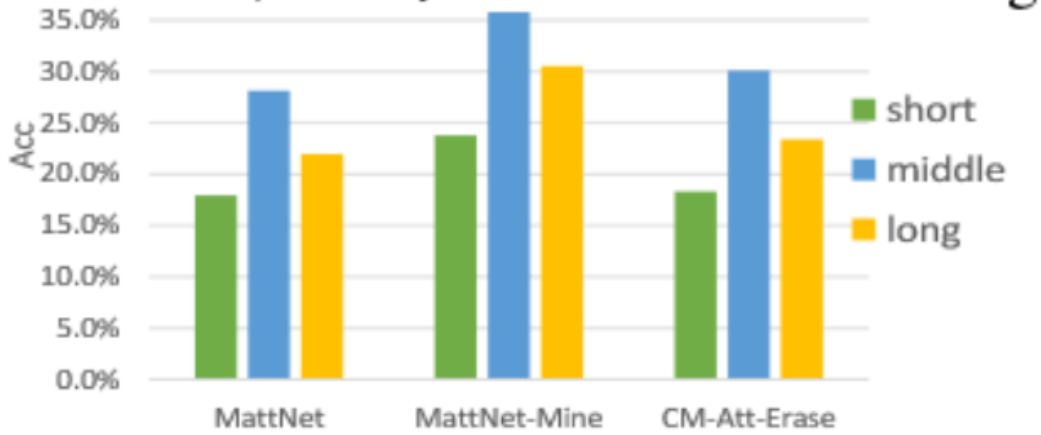


Figure 4: Accuracy of expressions of different lengths.

Thank You