Hypergraph Attention Networks for Multimodal Learning

Presented by Taehee Kim

DAVIAN
Data and Visual Analytics Lab

Motivation

- Most of the previous researches on learning multimodal inputs commonly make input features of each modality as vector forms after applying pre-trained pre-processing methods
- And integrate the multiple input features into a common vector space
- And apply problem-specific modules (usually fully connected neural networks)
- The feature vectors from different modalities are considered as abstracted information on the equivalent level, even though those are obtained from totally different pre-processing steps

Introduction

- One of the fundamental problems that arise in multimodel learning task is the disparity of information levels between different modalities
- There is a severe lack of consideration regarding the adequate form of the input representations of the multimodal data to learn
- This paper discovers that alignment of the information levels between the modalities is important
- The symbolic graphs are very powerful ways to represent the information of the low-level signals in alignment
- SOTA on GQA(Question Answering on Image Scene Graphs) dataset

Hypergraph Attention Networks

- The main purpose of the suggested method is to align information levels and to integrate the inputs within the same information level

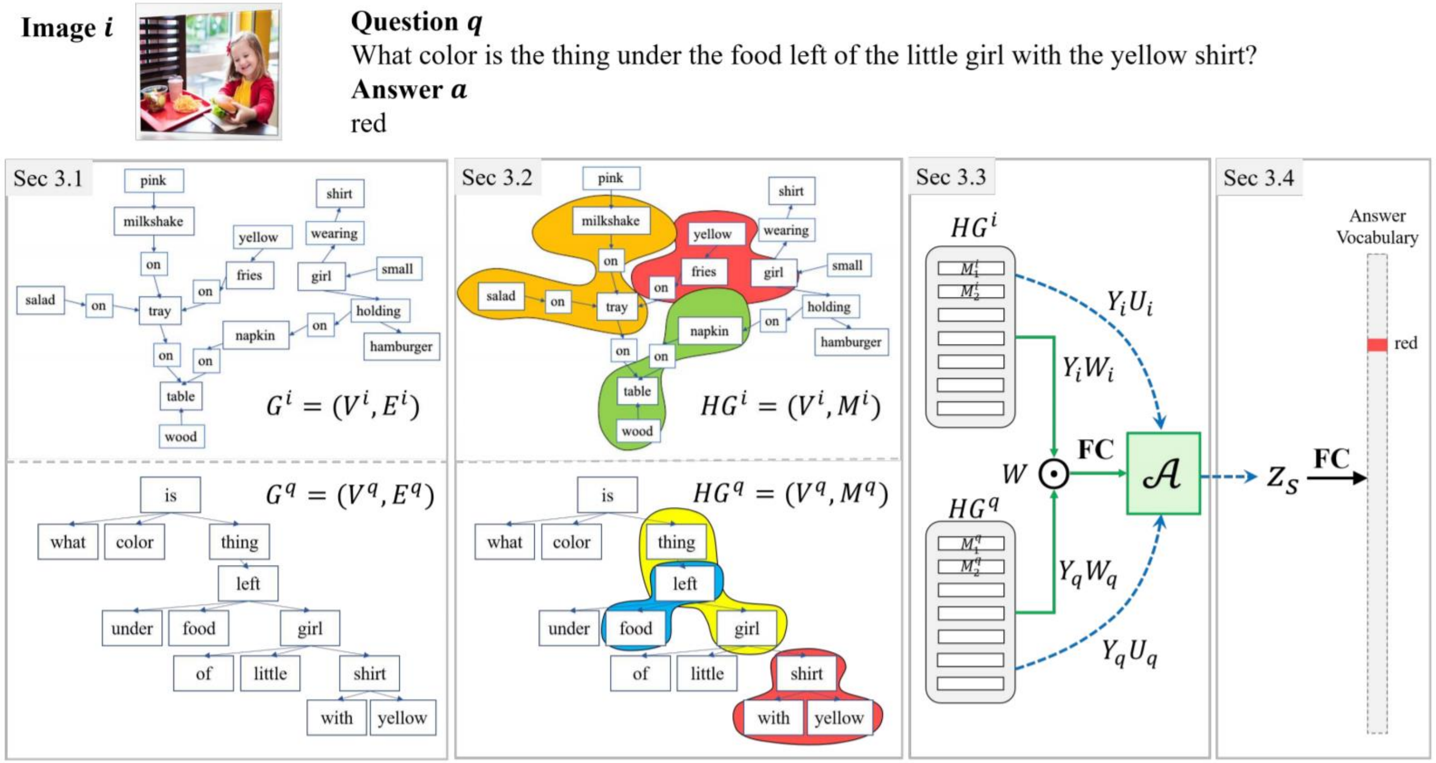


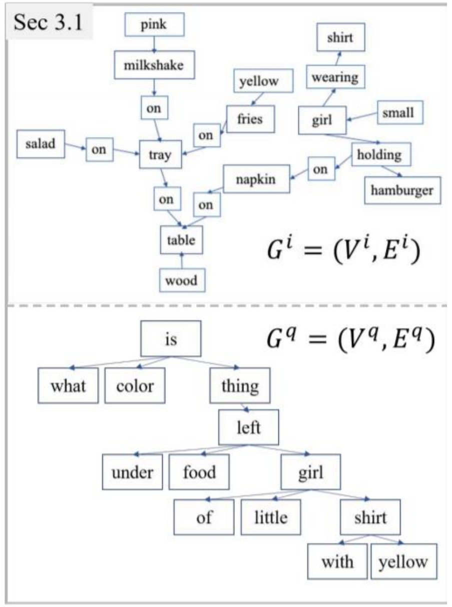
Figure 1. The overall architecture of the suggested model. For a given pair of image and question, two symbolic graphs are constructed. After constructing the symbolic graphs G^i and G^q, two hypergraphs HG^i and HG^q with random-walk based hyperedge are constructed. By comparing the semantics of each hyperedges, a co-attention map $\mathcal{A}$ is constructed. The two hypergraphs are combined by the co-attention map $\mathcal{A}$, and the final representation z_s is used to predict an answer for the given question.

I. Constructing Symbolic Graphs

- The symbolic representations of the two modalities are defined with graph forms

Image *i*

Question *q*
What color is the thing under the food left of the little girl with the yellow shirt?
Answer *a*
red



G^i: *Scene graph from image*

V^i: *Set of nodes (objects, attributes, relations)*

E^i: *Set of edges (object ∼ attribute, object ∼ relation)*

G^q: *Dependency tree*

V^q: *Set of tokens*

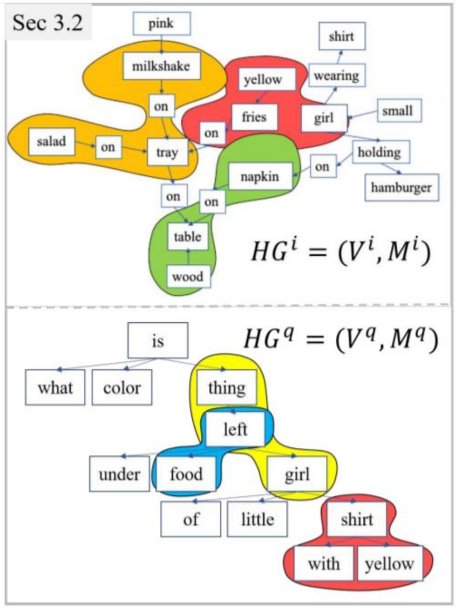
E^q: *Dependency between the tokens*

2. Constructing the Hypergraphs

- They consider each hyperedge as a sub-graph
- A hyperedge is a sequence of nodes sampled by random-walk algorithm along with directed edges

Image *i*

Question *q*
What color is the thing under the food left of the little girl with the yellow shirt?
Answer *a*
red



HG^i: *Hypergraph from image*

V^i: *Set of nodes (objects, attributes, relations)*

M^i: *Set of edges (object ~ attribute, object ~ relation)*

HG^q: *Hypergraph from dependency tree*

V^q: *Set of tokens*

M^q: *Dependency between the tokens*

Random-walk Algorithm

The initial probability that a node v_i will be selected is defined with, where $deg^+(v_i)$ represent the number of out-going edges from node v_i

$$P^0_{v_i} = \frac{deg^+(v_i) + \epsilon}{\sum_{j=1}^{N} deg^+(v_j)}$$

the transition probability is defined with,

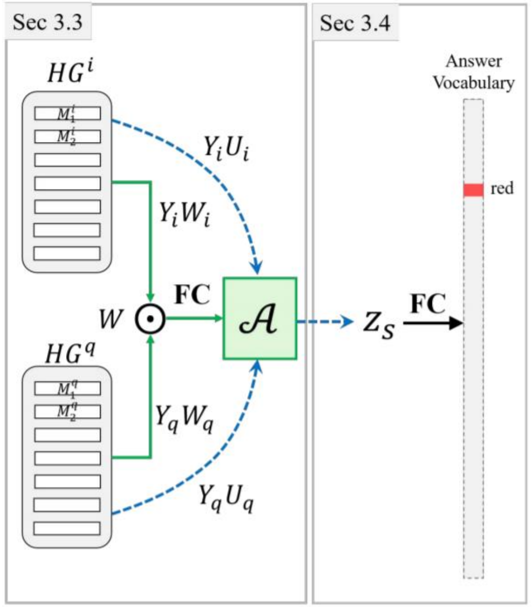
$$P_{v,u} = \begin{cases} \frac{1-\epsilon}{deg^+(v)}, & \text{if } (v,u) \in E \\ \epsilon, & \text{if } (v,u) \notin E \end{cases}$$

3. Building Co-attention Maps between Hyper-graphs

- The semantic of each hyperedge M can be defined by combining the word representations(e.g. GloVe) within the same hyperedge
- The co-attention map is based on comparing the semantics with the symbolic representations
- This method can consider the inherent structures by constructing the hypergraphs

Image *i*

Question *q*
What color is the thing under the food left of the little girl with the yellow shirt?
Answer *a*
red



$Y_q \in \mathbb{R}^{N^q \times 300}$

$Y_i \in \mathbb{R}^{N^i \times 300}$

$W_q, W_i \in \mathbb{R}^{300 \times h}$

$W \in \mathbb{R}^{N^q \times N^i}$

$U_q, U_i \in \mathbb{R}^{300 \times h}$

$$\mathcal{A} = \mathrm{softmax}(W \circ (Y_qW_q)(Y_iW_i)^\top)$$

$$\mathbf{z}_s = (Y_qU_q)^\top \mathcal{A}(Y_iU_i)$$

Results

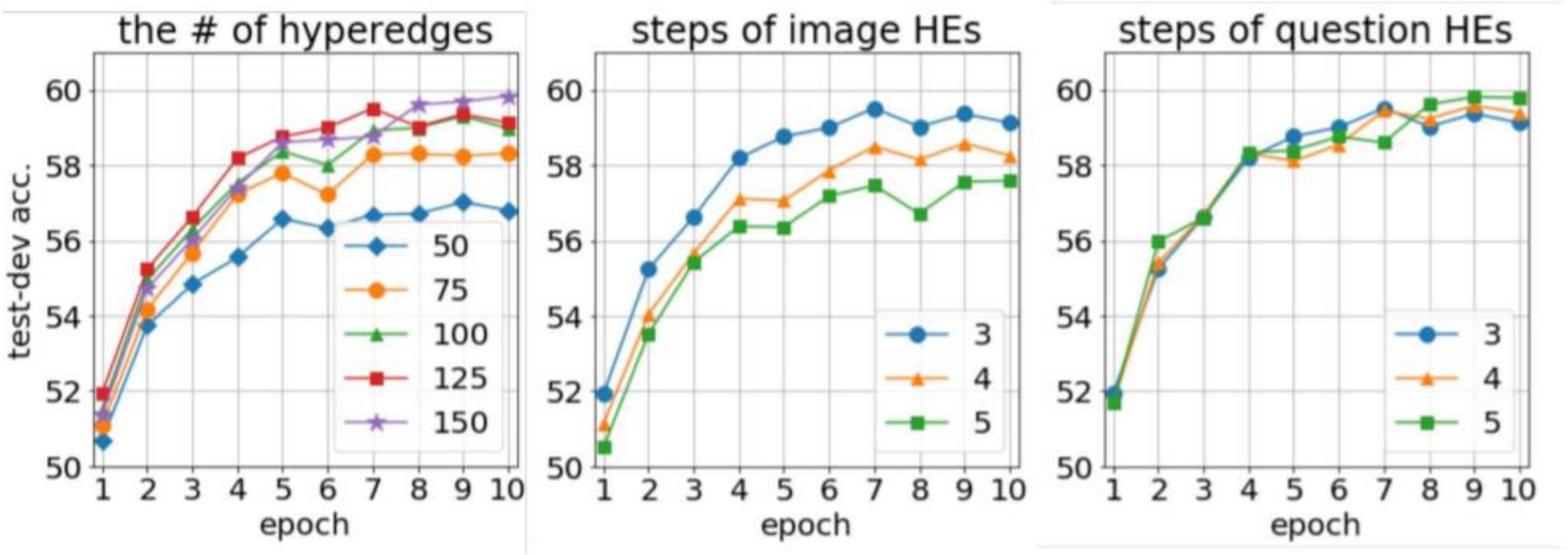


Figure 2. Test-dev accuracies with various hyper-parameter combinations. **Left:** Test-dev accuracy with various number of image hyperedges when the number of hyperedges for questions are fixed to 50. **Middle:** Accuracy with various number of steps (k) of image hyperedges with three-step question hyperedges. **Right:** Accuracy with various number of steps (k) of question hyperedges with three step image hyperedges.

Results

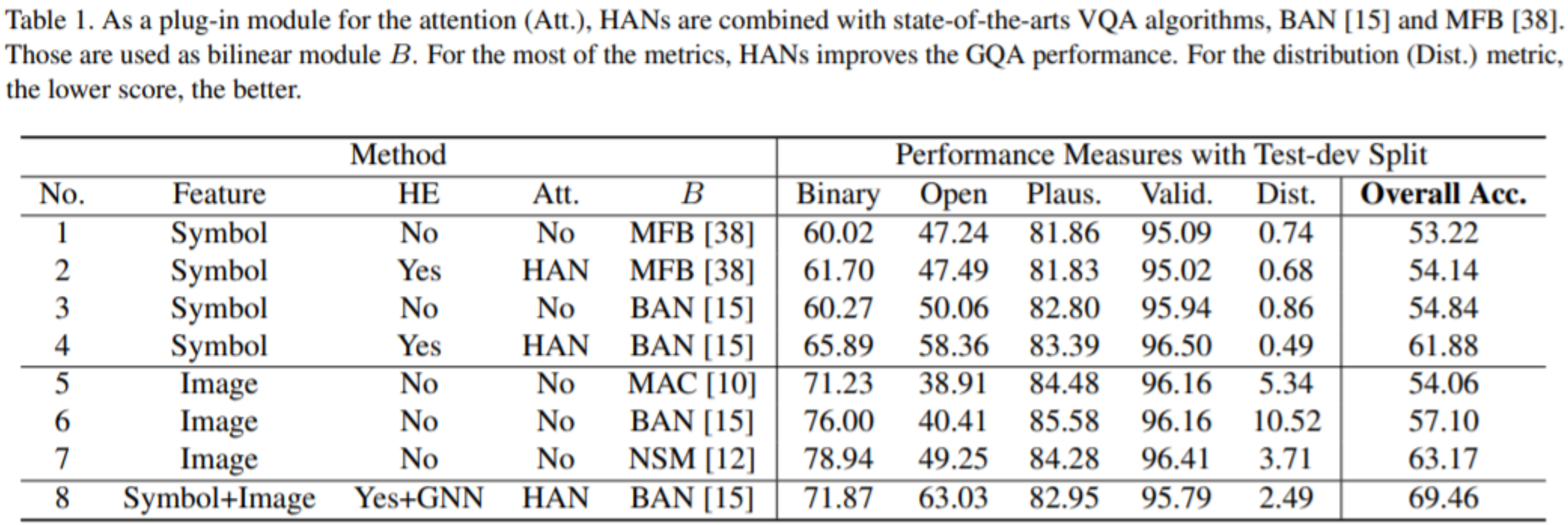
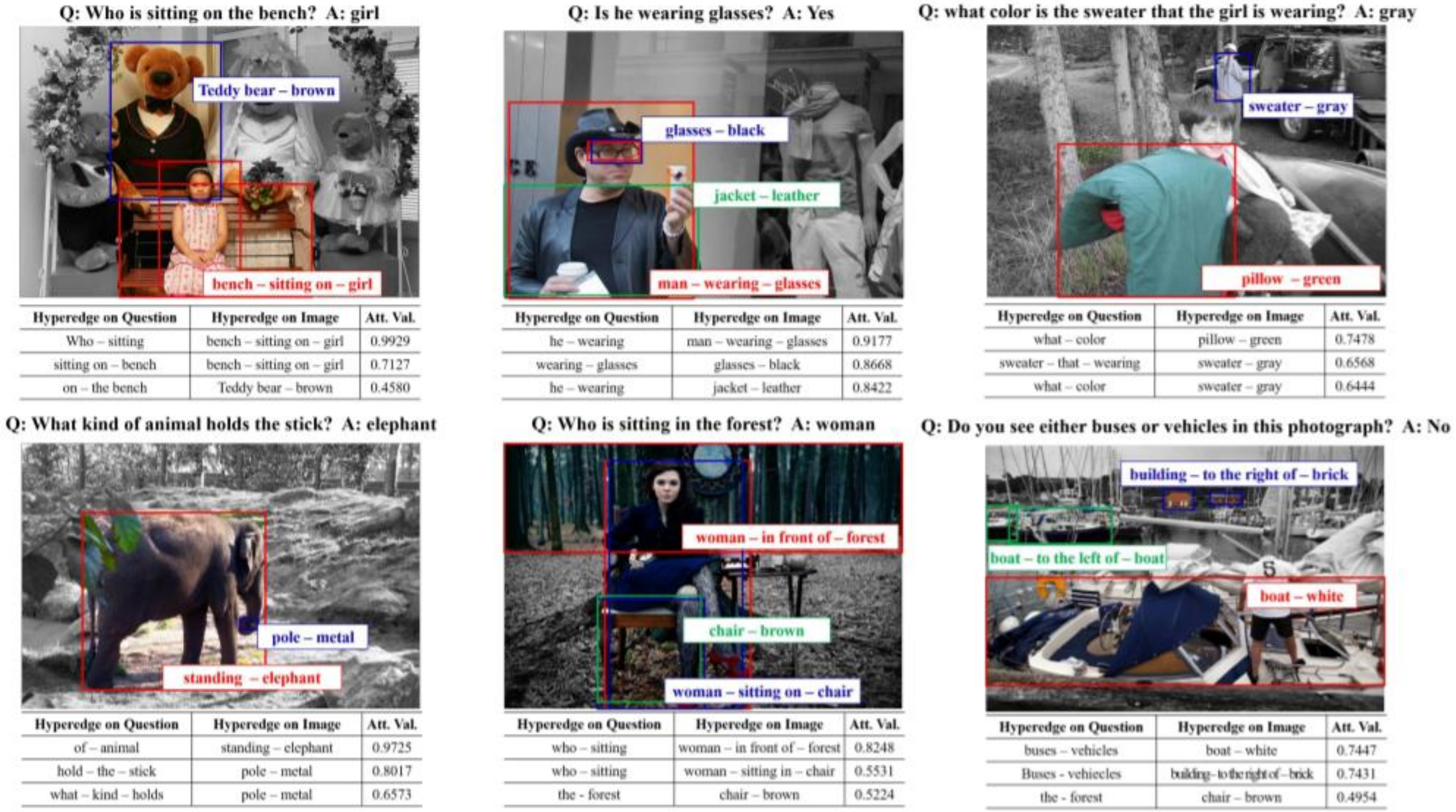


Table 1. As a plug-in module for the attention (Att.), HANs are combined with state-of-the-arts VQA algorithms, BAN [15] and MFB [38]. Those are used as bilinear module B. For the most of the metrics, HANs improves the GQA performance. For the distribution (Dist.) metric, the lower score, the better.

Method					Performance Measures with Test-dev Split					
No.	Feature	HE	Att.	B	Binary	Open	Plaus.	Valid.	Dist.	**Overall Acc.**
1	Symbol	No	No	MFB [38]	60.02	47.24	81.86	95.09	0.74	53.22
2	Symbol	Yes	HAN	MFB [38]	61.70	47.49	81.83	95.02	0.68	54.14
3	Symbol	No	No	BAN [15]	60.27	50.06	82.80	95.94	0.86	54.84
4	Symbol	Yes	HAN	BAN [15]	65.89	58.36	83.39	96.50	0.49	61.88
5	Image	No	No	MAC [10]	71.23	38.91	84.48	96.16	5.34	54.06
6	Image	No	No	BAN [15]	76.00	40.41	85.58	96.16	10.52	57.10
7	Image	No	No	NSM [12]	78.94	49.25	84.28	96.41	3.71	63.17
8	Symbol+Image	Yes+GNN	HAN	BAN [15]	71.87	63.03	82.95	95.79	2.49	69.46

Results



Q: Who is sitting on the bench? A: girl

Hyperedge on Question	Hyperedge on Image	Att. Val.
Who – sitting	bench – sitting on – girl	0.9929
sitting on – bench	bench – sitting on – girl	0.7127
on – the bench	Teddy bear – brown	0.4580

Q: Is he wearing glasses? A: Yes

Hyperedge on Question	Hyperedge on Image	Att. Val.
he – wearing	man – wearing – glasses	0.9177
wearing – glasses	glasses – black	0.8668
he – wearing	jacket – leather	0.8422

Q: what color is the sweater that the girl is wearing? A: gray

Hyperedge on Question	Hyperedge on Image	Att. Val.
what – color	pillow – green	0.7478
sweater – that – wearing	sweater – gray	0.6568
what – color	sweater – gray	0.6444

Q: What kind of animal holds the stick? A: elephant

Hyperedge on Question	Hyperedge on Image	Att. Val.
of – animal	standing – elephant	0.9725
hold – the – stick	pole – metal	0.8017
what – kind – holds	pole – metal	0.6573

Q: Who is sitting in the forest? A: woman

Hyperedge on Question	Hyperedge on Image	Att. Val.
who – sitting	woman – in front of – forest	0.8248
who – sitting	woman – sitting in – chair	0.5531
the - forest	chair – brown	0.5224

Q: Do you see either buses or vehicles in this photograph? A: No

building – to the right of – brick
boat – to the left of – boat
boat – white

Hyperedge on Question	Hyperedge on Image	Att. Val.
buses – vehicles	boat – white	0.7447
Buses - vehiecles	building–to the right of–brick	0.7431
the - forest	chair – brown	0.4954

Figure 3. The visualization for co-attention maps $\mathcal{A}$ of HANs with six examples. Among the all pairs of images and question hyperedges, three hyperedge pairs with top-3 attention-value are presented. The question is shown on the top of the image and the hyperedge pairs are on the bottom. Corresponding regions attended by HANs are represented on the image.

Appendix

- Merging Visual Features

$Y'_q \in \mathbb{R}^{N^q \times d'}$

$Y'_i \in \mathbb{R}^{N^v \times d'}$

$U'_q \in \mathbb{R}^{d' \times h}$

$U'_i \in \mathbb{R}^{d' \times h}$

$$\mathbf{z}_v = (\hat{Y}_q \hat{U}_q)^\top \mathcal{A}^* (\hat{V}_i \hat{U}_i)$$

- Graph Neural Network

$X \in \mathbb{R}^{S \times d}$

$A \in \{0,1\}^{S \times S}$ adjacency matrix

$D_{in}, D_{out} \in \mathbb{R}^{N \times N}$ indegree, outdegree matrix of A

$X_{new} \in \mathbb{R}^{S \times d}$

$$Z_{\text{in}} = \sigma(D_{in}^{-1} A X W_{\text{in}} + X W_{\text{in}})$$

$$Z_{\text{out}} = \sigma(D_{out}^{-1} A^\top X W_{\text{out}} + X W_{\text{out}})$$

$$X_{new} = \sigma((Z_{\text{in}} \circ Z_{\text{out}}) W_{\text{mrg}})$$