Exploiting the Social-Like Prior in Transformer for Visual Reasoning

Yudong Han[1], Yupeng Hu[1], Xuemeng Song[2], Haoyu Tang[1], Mingzhu Xu[1], Liqiang Nie[3]

[1]School of Software, Shandong University
[2]School of Computer Science and Technology, Shandong University
[3]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
{hanyudong.sdu, sxmustc, nieliqiang}@gmail.com, {huyupeng, tanghao258, xumingzhu}@sdu.edu.cn

Abstract

Benefiting from instrumental global dependency modeling of *self-attention* (SA), transformer-based approaches have become the pivotal choices for numerous downstream visual reasoning tasks, such as visual question answering (VQA) and referring expression comprehension (REC). However, some studies have recently suggested that SA tends to suffer from rank collapse thereby inevitably leads to representation degradation as the transformer layer goes deeper. Inspired by *social network theory*, we attempt to make an analogy between social behavior and regional information interaction in SA, and harness two crucial notions of *structural hole* and *degree centrality* in social network to explore the possible optimization towards SA learning, which naturally deduces two plug-and-play social-like modules. Based on *structural hole*, the former module allows to make information interaction in SA more structured, which effectively avoids redundant information aggregation and global feature homogenization for better rank remedy, followed by latter module to comprehensively characterize and refine the representation discrimination via considering *degree centrality* of regions and *transitivity* of relations. Without bells and whistles, our model outperforms a bunch of baselines by a noticeable margin when considering our social-like prior on five benchmarks in VQA and REC tasks, and a series of explanatory results are showcased to sufficiently reveal the social-like behaviors in SA.

Introduction

The success of transformer-based methods (Vaswani et al. 2017) in the natural language domain paves the way for the prosperity of vision reasoning tasks (Goyal et al. 2017), and numerous well-devised transformer variants have achieved promising performance on various benchmarks (Johnson et al. 2017). Due to the powerful global modeling ability of *self-attention* (SA) mechanism in transformer, these methods not only facilitate the intra-modality context learning, but also excel in inter-modality alignment and complementation. However, as discussed by several extant works (Dong, Cordonnier, and Loukas 2021; Dong et al. 2021), the traditional SA mechanism easily leads to quicker rank collapse and representation degradation without Feed-Forward Network (FFN) (Vaswani et al. 2017) and residual connection (He et al. 2016). Therefore, how to further optimize the effective learning of SA and generate more expressive representations for vision-and-language tasks remains an imperative issue.

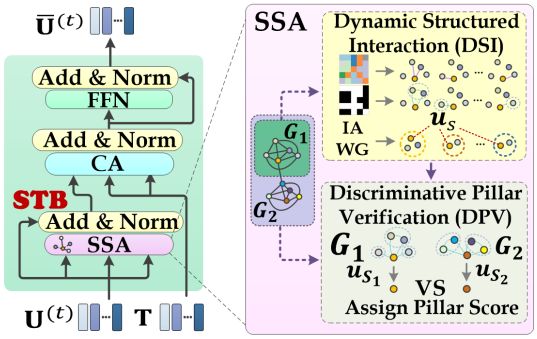


Figure 1: Illustration of Social-like Transformer Block (STB), where vanilla *self-attention* is replaced with SSA. Add&Norm denotes addition and layer normalization, respectively, CA represents the cross-modality attention, and FFN is the feed-forward network. In SSA, "IAWG" denotes *information aggregation within groups*, $\mathbf{G}_i$ and $\mathbf{u}_s$ represent the i-th subgraph and *structural hole*, respectively.

Intriguingly, we accidentally perceive that the feature aggregation amongst different image regions in SA seemingly shares similar philosophy with the information communication in social network. Moving forward in a purposeful way, we attempt to equip the *social network theory* (SNT) (Tabassum et al. 2018) into SA learning to gain some merits. For the sake of clarity, we first shed light on the counterparts between SA and social network. Regarding visual SA modeling in transformer, each visual unit, which can be grid feature (Jiang et al. 2020) or salient object (Anderson et al. 2018), follows to perform feature aggregation from other visual units based on the similarity score. We perceive that this process parallels social behavior, in other words, each visual

Corresponding author: Yupeng Hu.
This work was supported in part by the National Natural Science Foundation of China, No.:62276155, No.:62376137, No.:62206156, and No.:62206157; in part by the NSF of Shandong Province, No.:ZR2021MF040 and No.:ZR2022QF047;

region can be likened to a social member, and each member tends to make friends who has *kindred spirits*.

In this work, we make contributions towards optimizing SA learning from two kinds of social perspectives: (1) Inspired by the concept of *structural holes* (Tabassum et al. 2018) mentioned in SNT, we argue that not all pair-wise regions tend to establish connections in SA, and most region except *structural hole* is inclined to interacts with a limited number of regions whereas *structural hole* can be served as a link for communication with several isolated regional groups. This mechanism makes information interaction across different regions more structured, and effectively avoids redundant information aggregation and global feature homogenization, which elicits our **Dynamic Structured Interaction** (DSI). (2) After performing DSI, the excavated *structural holes* are responsible for receiving more sources of information as information hub, because of their greater degree and larger possibility to contact the representative regions within different groups. In light of this, we take the notion of *degree centrality* and *transitivity* in SNT into consideration to further delve into the spontaneous formation condition of *pillar*: one region has larger possibility to be pillar 1) when it obtains more focus from other regions or 2) it tends to establish closer connection to other pillars, which naturally leads to our **Discriminative Pillar Verification** (DPV) for further attention re-assignment. These two components are sequentially integrated into vanilla SA, and systematically work as a whole (i.e., SSA) for structured rank optimization and discriminative representation refinement.

We further equipped several transformer-based models with our SSA for two highly competitive tasks, VQA and REC, and we conducted extensive experiments on five popular public benchmarks, namely, VQA 2.0 (Goyal et al. 2017), CLVER (Johnson et al. 2017), RefCOCO (Yu et al. 2016), RefCOCO+ (Yu et al. 2016), and RefCOCOg (Mao et al. 2016a). The quantitative comparison shows that our method surpasses a bunch of baselines by a noticeable margin, and even achieves the new state-of-the-art (SOTA), and related visualization and mechanism exploration results sufficiently indicate the social behavior that SSA brings.

Our contributions are summarized as follows:

- We draw the inspiration from the *social network theory* to optimize the self-attention learning for two competitive visual reasoning tasks.
- We introduce SSA, where two novel modules are elaborately deployed. The former bespeaks the great potential to retard the rank collapse via structured interaction, followed by the latter one to further refine the discriminative representation based on region centrality.
- We conducted extensive experiments on five public available datasets to demonstrate the gain of two social-like modules and convincingly showcase their mechanism.

Related Work

Visual Question Answering

Visual question answering (VQA) targets at correctly answering the related questions based on given images, which is generally considered as a classification task with fixed number of categories. The recent years have witnesses the prosperity of VQA, numerous well-designed multimodal methods (Shi, Zhang, and Li 2019) and various benchmarks (Goyal et al. 2017; Johnson et al. 2017) have been proposed to facilitate this task. Earlier works utilize various multimodal embedding techniques (Hedi et al. 2019) to optimize the joint representation of image and question. Different from these lines of studies, lots of interpretable studies (Shi, Zhang, and Li 2019; Hu et al. 2018) open these black-box models to better explore the compositional reasoning. With the prevailing advent of attention mechanism in computer vision, a series of attention-based VQA models have been introduced to this task. Bottom-up and top-down attention mechanism (Anderson et al. 2018) is exploited to locate the crucial visual parts and learn more discriminative visual features. Concurrent with these works, some researchers (Wu et al. 2018; Li et al. 2019) focus on strengthening context-aware representation via exploring the relations between different visual regions. However, the increasing difficulty of sufficiently understanding questions in VQA often exceeds the ability of these single-step approaches. In this case, multiple-step attentions (Peng et al. 2019; Cadène et al. 2019) or stacked attentions (Gao et al. 2019) are further dedicated to gradually refine the visual/text information. The state-of-the-art methods (Gao et al. 2019; Zhou et al. 2021) exploit mainstreaming transformer-based architecture to capture the delicate relationships within and across modalities in a stacked manner.

Referring Expression Comprehension

Referring Expression Comprehension (REC) aims to localize the referred region in an image according to the given textual description. Existing methods typically base the object detection framework to improve the performance of REC task, which can be roughly categorized as two-stage framework and one-stage framework. Two-stage framework first utilizes the off-the-shelf object detectors (Yu et al. 2018a) to generate a bunch of region proposals from the image, and then select the top-ranked semantic-related proposal which mostly matches the language description. However, this framework heavily hinge on the quality of the pre-trained proposal detector. Recently, one-stage framework directly predicts the coordinate of referred regions without generating sets of candidate proposals in advance. These lines of work focus on devising diverse architectures, such as modular attention network (Yu et al. 2018a), various kinds of graph structure (Wang et al. 2019a; Yang, Li, and Yu 2019a), and multi-modal tree (Liu et al. 2019b), to establish fine-grained multimodal relationship for referred localization. Most recently, transformer-based methods (Yang et al. 2022; Deng et al. 2021) have been introduced to strengthen the plenitudinous interaction between the visual-linguistic context and further facilitate this task.

Preliminaries

For brevity, we formulate these two reasoning-hungry tasks as a generic form as follows. Given the training set $\Phi =$

$\{\mathcal{U}_b, \mathcal{T}_b, e_b\}_{b=1}^{B}$, where e_b can be the answer a_b in answer set $\mathcal{A}$ in the matter of VQA task or localization coordinate vector $\mathbf{b}_b$ for REC task, $\mathcal{U}_b$ corresponds to the b-th image with related text $\mathcal{T}_b$, where $\mathcal{T}_b$ can be questions and expressions, and B denotes the number of total training samples. During training, we are required to learn a mapping function $\mathbf{z}_b = g(\mathbf{U}_b, \mathbf{T}_b)$ that obtains the representation $\mathbf{z}_b$ for answer distribution prediction or grounding coordinate prediction, where $\mathbf{U}_b = [\mathbf{u}_{b1}, \mathbf{u}_{b2}, ..., \mathbf{u}_{bR}]$ and $\mathbf{T}_b = [\mathbf{t}_{b1}, \mathbf{t}_{b2}, ..., \mathbf{t}_{bW}]$ are obtained from $\mathcal{U}_b$ and $\mathcal{T}_b$ via image and text feature extractor, respectively, and $\mathbf{u}_{br}$ and $\mathbf{t}_{bw}$ denotes the feature of the r-th region and the w-th word, respectively.

Transformer with Social-Like Prior

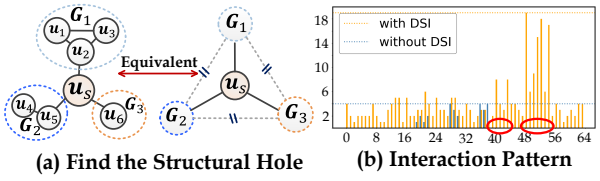
To highlight our contribution, we provide the illustration of architecture of our proposed Social-Like Transformer Block (STB) in Figure 1. In STB, we replace the traditional SA with our SSA, which consists of two key parts: Dynamic Structured Interaction (DSI) and Discriminative Pillar Verification (DPV). In what follows, we elaborate the designed motivation and details of these two components sequentially.

Dynamic Structured Interaction (DSI)

Before introducing our DSI, we first briefly review the traditional calculation pipeline of SA. The traditional SA function $f_{SA}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ first operates on query $\mathbf{Q} \in \mathbb{R}^{R\times d}$, key $\mathbf{K} \in \mathbb{R}^{R\times d}$, and value $\mathbf{V} \in \mathbb{R}^{R\times d}$, which are obtained by three separate linear projections $\varphi_q(\cdot)$, $\varphi_k(\cdot)$, $\varphi_v(\cdot)$ using the visual regional representation $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_R] \in \mathbb{R}^{R\times d}$ as input, where R is the number of regions in image, and the interaction matrix between any pair-wise regions are calculated as follows,

$$\mathbf{A}_{ir} = \varphi_q(\mathbf{u}_i)\varphi_k(\mathbf{u}_r)^\top/\sqrt{d} = \mathbf{q}_i\mathbf{k}_r^\top/\sqrt{d}, \tag{1}$$

where $\mathbf{A}_{ir}$ considers all the long-range correlations and assigns a fixed number of interactive neighbors for each region to strengthen the reasoning-hungry contextual representation. However, this strategy would unavoidably trigger some redundant connection that the irrelevant regions bring. Based on the analysis in the literatures (Dong, Cordonnier, and Loukas 2021; Dong et al. 2021), the deeper SA would result in more common neighbors, and similarly lead to the representation assimilation as the over-smoothing phenomenon in Graph Neural Network (GNN) (Velickovic et al. 2018), which further incurs rank collapse. To tackle this issue, we harness *structural holes* in SNT to make the interaction pattern between regions more structured. As depicted in Figure 2, *structural hole* $\mathbf{u}_s$ can be served as a hub that establishes connections between several isolated social groups $\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$. Interestingly, we observe that identifying the *structural holes* $\mathbf{u}_s$ is equivalent to exactly removing the possible redundant relationships from the complete graph that built by regions in different groups, i.e, $\overline{\mathbf{u}_i\mathbf{u}_r}|_{\mathbf{u}_i,\mathbf{u}_r\in\mathbf{G}_1,\mathbf{G}_2,\mathbf{G}_3}$. As concluded in SNT, the appearance of *structural holes* hinges on their relative location surroundings and their own features (Tabassum et al. 2018). In light of this, two modulation networks $\varphi_\alpha(\cdot)$ and $\varphi_\beta(\cdot)$ (i.e., learnable fully connected networks) are employed, using the region feature $\mathbf{u}_i$ and its connection preference $\mathbf{A}_{i,:}$ as control information to generate the effective mask that helps to form *structural hole*,



(a) Find the Structural Hole (b) Interaction Pattern

Figure 2: (a) shows the equivalence between localizing the possible *structural hole* $\mathbf{u}_s$ and removing the redundant relations, and (b) shows the interaction number (y-axis) with other regions of each region (x-axis), where the region index with red circle signifies *structural hole*.

$$\mathbf{M}_{i,:} = \varphi_\alpha(\mathbf{F}) \odot (\mathbf{A}_{i,:} - \mathcal{M}(\mathbf{A}_{i,:})) + \varphi_\beta(\mathbf{F}), \tag{2}$$

where $\mathbf{F} = [\mathbf{A}_{i,:}, \mathbf{u}_i] \in \mathbb{R}^{d+R}$, $\mathcal{M}(\cdot)$ is used to calculated the average value of vector, and $\odot$ denotes the element-wise multiplication operation. We then further impose a non-negative constraint to obtain the discrete interactive indicator for each region,

$$\mathbf{H}_{i,:} = \mathbb{I}(\mathit{ReLU}(\mathbf{M}_{i,:})), \tag{3}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, where its value will be set to 1 if the input is larger 0, and set to 0 otherwise. After obtaining $\mathbf{H}_{i,:}$, we achieve the structured interaction calibration as follows,

$$\widetilde{\mathbf{A}}_{ij} = \mathit{Softmax}(\mathbf{H}_{ij} \odot \mathbf{A}_{ij}). \tag{4}$$

Guided by the structured interaction matrix $\widetilde{\mathbf{A}}_{ij}$, the feature update process can be summarized as,

$$\widetilde{\mathbf{u}}_i = \sum_j \widetilde{\mathbf{A}}_{ij}\ \varphi_v(\mathbf{u}_j). \tag{5}$$

By this means, $\widetilde{\mathbf{A}}_{ij}$ has the ability to control the structured interaction pattern, which effectively avoids redundant information aggregation and global feature homogenization.

Discriminative Pillar Verification (DPV)

As shown in Figure 2, the localized *structural hole* $\mathbf{u}_s$ has greater degree and larger possibility to get in touch with the representative regions within different groups. They are responsible for receiving more sources of information when performing context representation enhancement. Therefore, these holes, called *pillar* candidates, convey more global information of image than other local-ware regions after DSI, which are supposed to be assigned more attention. Based on these observations, we take two key concepts, *degree centrality* and *transitivity*, into consideration to quantify the reasonable attention allocation, that is, finding the *pillar*. The former characterizes the importance of each region according to its degree, and the latter models the interaction preference towards different neighbors. Inspired by this, we make

attempts to reveal the possibility of identifying the *pillar* by leveraging the degree information and connection pattern, which leads to two nature standards of being pillar: *one region has larger possibility to be pillar when 1) it obtains the focus from more other regions in image or 2) it has larger possibility to be pillar when it establish closer connection to other pillars.* **It is worth emphasizing that** unlike the traditional top-down approaches (Anderson et al. 2018), our DPV supports to harness the intrinsic structure information in visual graph rather than semantic-guided strategies to refine the visual discriminative representation.

In what follows, we illustrate the details regarding how to verify *pillar* based on above philosophy. We first generate the region importance score s_i of the i-th region by computing $s_i = \sum_r \mathbf{A}_{ir}$, where $\mathbf{s}_i$ denotes the obtained focus from other regions. By concatenating all the focus score of regions $\mathbf{s} = [s_1, s_2, ..., s_R]$, we can obtain the global focus vector $\mathbf{s}$ of whole visual graph. Taking the *transitivity* and *degree centrality* in visual graph structure into consideration, these region importance information would be transferred alongside the weight of connected relationship, and we combine the transitivity information $(\mathbf{H}_{i,:} \odot \mathbf{A}_{i,:})$ with centrality information $\mathbf{s}$ to obtain the final focus score,

$$\mathbf{w}_i = \mathbf{H}_{i,:} \odot (\mathbf{s} + \mathbf{A}_{i,:}). \tag{6}$$

In Eqn.(6), $(\mathbf{H}_{i,:} \odot \mathbf{A}_{i,:})$ characterizes the structured interaction preference of the i-th region to other regions, $(\mathbf{H}_{i,:} \odot \mathbf{s}$ represents the calibrated importance distribution of regions, and the informative vector $\mathbf{w}_i$ could comprehensively convey the possibility to be *pillar*. To further endow pillar verification process with more flexibility, we adopt the learnable form as follows,

$$p_i = \textit{Sigmoid}(\varphi_p([\mathbf{w}_i, s_i])), \tag{7}$$

where $[\cdot, \cdot]$ denotes the concatenation operation, $\varphi_p(\cdot)$ represents the two-layer linear projection. By concatenating the score of each region, the learned score vector $\mathbf{p} = [p_1, p_2, ..., p_R]$ is obtained. Thereafter, the learned visual representation in SA is further re-calibrated by regularized vector $\mathbf{p}$, which is updated as follows,

$$\bar{\mathbf{U}} = \widetilde{\mathbf{U}} \odot \mathbf{p}, \tag{8}$$

where $\bar{\mathbf{U}} = [\bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2, ..., \bar{\mathbf{u}}_R]$ and $\widetilde{\mathbf{U}} = [\widetilde{\mathbf{u}}_1, \widetilde{\mathbf{u}}_2, ..., \widetilde{\mathbf{u}}_R]$, each of vector $\widetilde{\mathbf{u}}_i$ can be obtained by Eqn.(5), and $\bar{\mathbf{u}}_i$ is accordingly calculated by $\bar{\mathbf{u}}_i = \widetilde{\mathbf{u}}_i \cdot p_i$. After obtaining $\bar{\mathbf{U}}$, we input these representation to the subsequent blocks, such as FFN (Vaswani et al. 2017) or Cross-Modality Attention (Vaswani et al. 2017), which is the same as the modules in traditional transformer block. After T iterations of our STB, we derive the final refined output $\bar{\mathbf{U}}^{(T)}$.

Training Objective

Output Head. In VQA task, we aggregate $\bar{\mathbf{U}}^{(T)}$ with original question representation $\mathbf{T}$ by attentional reduction (Tan and Bansal 2019) to obtain the joint global representation for the whole image, which are followed by a multi-layer predictor for multi-label classification,

$$\mathbf{z} = \mathcal{F}_{d \to |\mathcal{A}|}(\bar{\mathbf{U}}^{(T)}, \mathbf{T}), \tag{9}$$

where $\mathcal{A}$ denotes the candidate answer set, $|\mathcal{A}|$ represents the number of element in $\mathcal{A}$, d denotes the feature dimension before predictor, and $\mathbf{z} \in \mathbb{R}^{|\mathcal{A}|}$ represents the predicted answer distribution. For REC task, we adopt the pooling operator on $\bar{\mathbf{U}}^{(T)}$, which is followed by the box regression layer,

$$\mathbf{b} = \mathcal{F}_{d \to 4}(\bar{\mathbf{U}}^{(T)}), \tag{10}$$

where $\mathbf{b} \in \mathbb{R}^4$ denotes the predicted 4-dim bounding box vector, each of element represents four coordinates of the grounding region, respectively.

Loss Function. In VQA task, we use the binary cross-entropy as the loss function for training following (Anderson et al. 2018), which is defined as follows:

$$\mathcal{L}_{vqa} = -\sum_{j}^{|\mathcal{A}|} \mathbf{y}_j log(\mathbf{z}_j) - (1 - \mathbf{y}_j) log(1 - \mathbf{z}_j), \tag{11}$$

where $\mathbf{z}_j$ denotes the predicted probability on the j-th answer slot, and $\mathbf{y}_j = 1$ if the ground-truth answer is a_j, and other slot is set to 0. For REC task, we denote the prediction as $\mathbf{b} = (x, y, w, h)$, and the normalized ground-truth box is denoted as $\tilde{\mathbf{b}} = \left(\tilde{x}, \tilde{y}, \tilde{w}, \tilde{h}\right)$. The training objective is,

$$\mathcal{L}_{rec} = \mathcal{L}_{smooth}(\tilde{\mathbf{b}}, \mathbf{b}) + \mathcal{L}_{giou}(\tilde{\mathbf{b}}, \mathbf{b}), \tag{12}$$

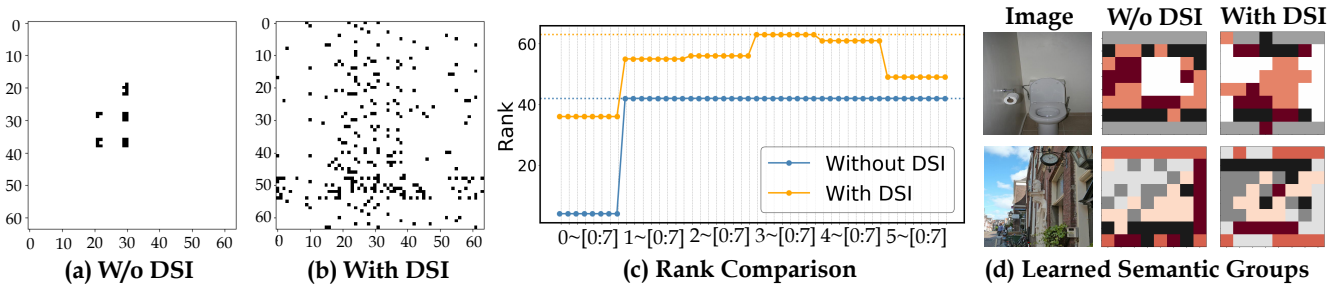
where $\mathcal{L}_{smooth}(\cdot)$ and $\mathcal{L}_{giou}(\cdot)$ represents the smooth ℓ-1 loss and GIoU loss (Rezatofighi et al. 2019), respectively.

Experiments

To demonstrate the effectiveness of our method, we apply them to two considerably competitive reasoning tasks, VQA and REC. Five benchmarks are involved, namely, VQA 2.0, CLEVR, RefCOCO, RefCOCO+ and RefCOCOg.

Experimental Setup

Datasets. **VQA 2.0** is the most commonly used benchmark dataset for VQA, which is developed based on VQA 1.0. The images stems from Microsoft COCO (Lin et al. 2014). The overall dataset has about ∼1000K examples, which are splited into *train*, *val* and *test*, respectively. **CLEVR** is a synthetic diagnostic dataset, which is used to examine a range of visual reasoning abilities. In CLEVR, there are ∼70K/∼15K images and ∼700K/∼150K questions in the *train/val* set, where questions are parsed as several compositional functional programs. **RefCOCO**, **RefCOCO+**, and **RefCOCOg** are three commonly used benchmarks for REC. RefCOCO has ∼20K images with ∼142K referring expressions for 50K referred objects, which is split into *train*, *val*, *testA*, and *testB* set. The expressions are generally represented as short sentences with an average length of 3.5 words. RefCOCO+ provides ∼20K images and ∼142K expressions regarding 50K referred objects. RefCOCOg has ∼260K images with ∼960K expressions to describe 50K objects. The expressions in RefCOCOg are generally longer than those in the other two datasets, with 8.4 words on average. In particular, RefCOCOg can be splitted into two parts, which are RefCOCOg-google (Mao et al. 2016b) (*val-g*) and RefCOCOg-umd (Nagaraja, Morariu, and Davis 2016) (*val-u*, *test-u*).



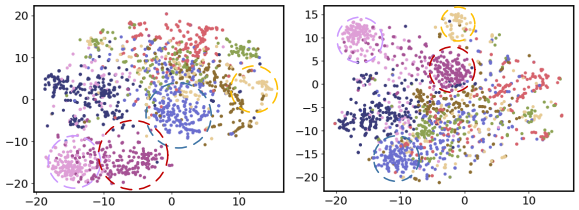
(a) W/o DSI (b) With DSI (c) Rank Comparison (d) Learned Semantic Groups

Figure 3: (a) and (b) show the learned interaction matrix (discretization via threshold 0.1 and 0.6 for better illustration, respectively) with DSI and without DSI in the last SA layer, respectively. (c) shows how our DSI facilitates the rank learning in SA. For x-axis, we define $L \sim [H_0 : H_7]$ as the H_h-th head of the L-th layer in transformer. (d) shows the semantic groups obtained by clustering the learned grid features using spectral clustering.

Models	VQA 2.0 *test-dev*			
	All	*Y/N*	*Num.*	*Others*
Multimodal Embedding:				
MCB (Fukui et al. 2016)	62.3	78.8	38.3	53.4
MLB (Kim et al. 2017)	66.3	83.6	44.9	56.3
MUTAN (Hedi et al. 2017)	66.0	82.9	44.5	56.5
BLOCK (Hedi et al. 2019)	67.6	83.6	47.3	58.5
Attentional Modeling:				
CapsAtt (Zhou et al. 2019)	65.5	82.6	45.1	55.5
UpDn (Anderson et al. 2018)	65.3	81.8	44.2	56.1
BAN (Kim and Jun 2018)	69.5	85.3	50.9	60.3
Visual Relation Modeling:				
ODA (Wu et al. 2018)	68.2	84.7	48.0	58.7
MuRel (Cadène et al. 2019)	68.2	84.7	48.0	58.7
ReGAT (Li et al. 2019)	70.3	86.0	**54.4**	60.3
CRA-Net (Peng et al. 2019)	68.6	84.9	49.5	59.0
VCTREE (Tang et al. 2020)	68.2	84.3	47.8	59.1
Transformer-based:				
MMNAS (Yu et al. 2020)	71.2	87.3	55.7	61.0
LENA (Han et al. 2021)	70.3	86.6	54.3	60.2
AGAN (Zhou et al. 2020)	71.2	86.9	54.3	61.6
ReATT (Guo et al. 2021)	70.4	87.0	53.1	60.2
SUPER (Han et al. 2023)	70.3	86.6	51.5	60.7
Ours:				
MCAN* (Yu et al. 2019)	71.3	87.2	53.7	61.7
+Ours	**71.9**	**87.8**	**54.8**	**62.4**
TRAR* (Zhou et al. 2021)	71.5	87.5	54.5	61.7
+Ours	**72.0**	**87.8**	**54.8**	**62.2**
TRAR† (Zhou et al. 2021)	71.9	87.5	54.1	62.6
+Ours†	**72.5**	**88.0**	**55.1**	**63.0**

Table 1: Performance comparison on VQA 2.0 *test-dev*, all the model is trained on the *train* + *val* + *vg* splits. "*" denotes our re-implementation results, and "†" represents the versions that finetuned with only *train* + *val* after 10 epochs.

Implementation Details. In VQA task, the model configuration for VQA 2.0 and CLEVR are similar. Following (Zhou et al. 2021), the input text words are initialized by 300-dimension GLOVE embeddings, and LSTM is utilized to the encode language information, and the dimension is set to 512. We use ResNext152 pre-trained on Visual Genome as the basic visual backbone to extract grid image features, these features are first padded to 16×16 scale, and then pooled by a kernel size of 2×2 with a stride of 2, obtaining the 8×8 resolution for visual input. The numbers of training epochs for VQA 2.0 and CLEVR are set to 13 and 16, respectively, and warming-up strategy is adopted in the first three epochs. The learning rates is initialized by 1e-4, which are decayed by 0.2 on the 10-th, 13-th and 15-th epochs. The batch size is set to 64. In REC task, following (Yang et al. 2022), we use DarkNet-152 or ResNet-50 as our visual feature extraction backbone followed by 6-layer transformer encoder for visual information encoding, and the textual embedding extraction branch is initialized with basic BERT model (Devlin et al. 2019). On three datasets, our model is trained for 90 epochs with a initial 1e-4 learning rate dropped by a factor of 10 after 60 epochs, except RefCOCOg with training for 60 epochs and dropping after 40 epochs, and we set the batch size to 16.



(a) Without DPV (b) With DPV

Figure 4: t-SNE visualization of representation from the last SA layer in MCAN. (a) and (b) show the vectors obtained by performing average pooling on $\widetilde{\mathbf{U}}$ and $\bar{\mathbf{U}}$, respectively.

Evaluation Metric. For VQA task, we adopted the standard accuracy metric for evaluation following (Antol et al. 2015). Given an image and a corresponding question, for a predicted answer a, the accuracy is computed as follows,

$$Acc_a = \min(1, \frac{\#\text{humans that provide } a}{3}). \quad (13)$$

For REC task, we follow the same $Acc@0.5$ evaluation protocol in prior work (Yang et al. 2022). Given a language ex-

Models	Program?	CLEVR *val* (Acc%)					
		Overall	*Count*	*Exist*	*Comp_Num*	*Query_Attr*	*Comp_Attr*
Human (Hu et al. 2017)	None	92.60	86.70	96.60	86.40	95.00	96.00
Supervised:							
DDRProg (Zhuang et al. 2018)	Used	98.30	96.50	98.80	98.40	99.10	99.00
NS-VQA (Yi et al. 2018)	Used	99.80	99.70	99.90	99.90	99.80	99.80
NS-CL (Mao et al. 2019)	Used	98.90	98.20	99.00	98.80	99.30	99.10
OCCAM (Wang et al. 2021)	Used	99.40	98.10	99.80	99.00	99.90	99.90
Unsupervised:							
FILM(Zhang, Niu, and Chang 2018)	Non-used	97.60	94.50	99.20	93.80	99.20	99.00
MAC (Zhuang et al. 2018)	Non-used	98.90	97.20	99.50	99.40	99.30	99.50
TBD (Zhuang et al. 2018)	Non-used	98.70	96.80	98.90	99.10	99.40	99.60
XNM-Net (Shi, Zhang, and Li 2019)	Non-used	97.80	96.00	98.10	98.60	98.70	97.80
Ours:							
MCAN* (Yu et al. 2019)	Non-used	98.32	95.24	98.06	98.53	98.85	98.42
+Ours	Non-used	**98.75**	**96.38**	**99.51**	**99.28**	**99.65**	**99.13**
TRAR* (Zhou et al. 2021)	Non-used	98.83	96.73	99.50	99.18	**99.65**	99.24
+Ours	Non-used	**99.15**	**97.73**	**99.65**	**99.33**	99.61	**99.35**

Table 2: Comparison with the state-of-the-arts on CLEVR *val*. The program option "Non-used" means totally without program annotations, and "Used" means using ground-truth programs. "*" denotes our re-implementation results.

Models	VQA 2.0 *val*			
	All	*Others*	*Yes/No*	*Num.*
MCAN	67.27	58.70	84.88	49.04
+DSI	**67.58**	**59.21**	**85.25**	**49.95**
+DPV	**67.69**	**59.13**	**85.15**	**49.82**
+DSI+DPV	**68.01**	**59.30**	**85.47**	**50.71**
TRAR	67.62	58.87	85.32	49.81
+DSI	**67.89**	**59.38**	**85.34**	**50.28**
+DPV	**67.97**	**59.08**	**85.42**	**50.19**
+DSI+DPV	**68.16**	**59.54**	**85.56**	**50.42**

Table 3: Ablative results on VQA 2.0 *val*.

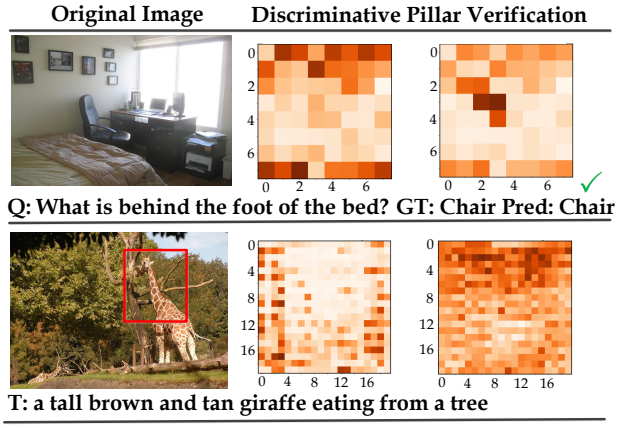


Figure 5: Case studies of the learned pillar score from the 1-th and 5-th layer on VQA 2.0 and RefCOCOg. The index within $[0-M]$ (M = 8, 20) represents the scale of images.

pression query, the predicted region is considered as correct if its covered region has at least 0.5 overlap with the ground-truth bounding box.

Overall Performance Comparison

We compare the performance of our method against state-of-the-art baselines on five popular benchmarks. (1) *Results on VQA 2.0:* The quantitative results are illustrated in Table 1. It can be seen that our model not only achieves the significant improvement based on several modern architectures, but also outperforms a bunch of SOTA by a noticeable margin. Specifically, we integrate our social-like prior into TRAR, obtaining the new SOTA performance on this highly competitive benchmark. Our best single model delivers **72.5%** overall accuracy on the *test-dev* set. In addition, with a deeper look at the *others* type (involves more "what ..." or "why ..." types) that deeply requires for reasoning ability, we observe that our method yields a clear-cut improvement over MCAN and TRAR by 0.6% and 0.67% on VQA 2.0 *val*, respectively, and 0.7% and 0.5% on VQA 2.0 *test-dev*, respectively. This adequately shows the power of our social-like mechanism on reasoning ability. (2) *Results on CLEVR 1.0:* As can be observed in Table 2, we also report the results on CLEVR. This dataset mainly focuses on the diagnosis about visual reasoning ability, and the questions are usually longer and more complicated compared to VQA 2.0. On CLEVR, the performance gain shows obvious at *Count* and *Exist* based on MCAN, this probably due to that answering this type of questions requires more sensitivity to spatial information in image. Our DSI excells in the structural modeling and dynamism of interaction, which better verifies this phenomenon. (3) *Results on RefCOCOg, RefCOCO, and RefCOCO+:* To further verify the gain on the grounding task, we also report our performance in Table 4. With our social-like modeling, our method consistently achieves the new SOTA performance among almost all the subsets and splits. Remarkably, when performing grounding on longer expressions (on the RefCOCOg

Models	Backbone	RefCOCO			RefCOCO+			RefCOCOg		
		val	*testA*	*testB*	*val*	*testA*	*testB*	*val-g*	*val-u*	*test-u*
Two-stage:										
MAttNet (Yu et al. 2018b)	ResNet-101	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27
LGRANs (Wang et al. 2019b)	VGG16	-	76.60	66.40	-	64.00	53.40	61.78	-	-
DGA (Yang, Li, and Yu 2019b)	VGG16	-	78.42	65.53	-	69.07	51.99	-	-	63.28
RvG-Tree (Hong et al. 2019)	ResNet-101	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51
NMTree (Liu et al. 2019a)	ResNet-101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
Ref-NMS (Chen et al. 2021)	ResNet-101	80.70	84.00	76.04	68.25	73.68	59.42	-	70.55	70.62
One-stage:										
SSG (Chen et al. 2018)	DarkNet-53	-	76.51	67.50	-	62.14	49.27	47.47	58.80	-
FAOA (Yang et al. 2019)	DarkNet-53	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.36
RCCF (Liao et al. 2020)	DLA-34	-	81.06	71.85	-	70.35	56.32	-	-	65.73
ReSC-Large (Yang et al. 2020)	DarkNet-53	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
LBYL-Net (Huang et al. 2021)	DarkNet-53	79.67	82.91	74.15	68.64	73.38	59.49	62.70	-	-
Ours:										
TransVG* (Zhou et al. 2021)	ResNet-50	79.89	83.64	74.53	62.28	63.56	53.12	63.89	64.51	**65.67**
+Ours	ResNet-50	**80.78**	**84.66**	**75.09**	**63.34**	**65.71**	**55.01**	**64.76**	**65.69**	65.43
VLTVG* (Yang et al. 2022)	ResNet-50	82.45	**84.98**	76.45	71.08	75.71	60.45	69.20	71.08	70.64
+Ours	ResNet-50	**83.09**	84.87	**76.98**	**71.52**	**76.46**	**61.46**	**70.73**	**72.24**	**70.79**

Table 4: Comparison of our method with other state-of-the-art methods on RefCOCO, RefCOCO+, and RefCOCOg. "*" denotes our re-implementation results. Due to that transformer implementation of TransVG and VLTVG adopts toolkit (i.e., nn.MultiheadAttention), which ceases to provide the open interface for our design, and we can only resort to the bootstrapped implementation in MCAN. This may cause $\sim$2% drop compared with the official results.

dataset), we achieve 70.73% and 72.24% on the *val* set and *val-u*, respectively, obtaining 1.53% absolute improvement over the previous SOTA VLVTG. In a deep analysis, longer expression with complicated semantic like "a tall brown and tan giraffe eating from a tree" tends to localize more details, such as "tan giraffe", "eating", "tall brown", and "tree" in the image, which may be more brittle to the noise from redundant regions when performing context learning. Benefiting from social-like prior, our method could learn more delicate visual context and discriminative region representation.

Analysis on Social-Like Prior

To take a closer look on our social-like prior, we devise three variants: (1) **+DSI** refers to the method using only DSI. (2) **+DPV** is the variant that solely equipped with DPV. (3) **+DSI+DPV** represents the full model with collaborative learning of two modules. In what follows, we shed light on their behaviors from two perspectives: (1) *Ablative Quantitative Results:* The ablative results are reported in Table 3, which dissects the effectiveness of DSI and DPV. It is evident that the performance of all part-equipped variants exhibits a noticeable improvement compared to the base model across all types of questions. Furthermore, our full method achieves superior gain on several basic versions, which can be observed in Table 1, 2, and 4. These observations convincingly verified the effectiveness of our social-like modules. (2) *Interpretability:* As depicted in Figure 3 (a), (b), and (c), *structural holes* in DSI, serve as a link for communication with different semantic groups and also have more sources of information, are naturally required to have more diverse interactions. In Figure 2 (b), it makes information interaction more structured, which is reflected in better continuity within semantic groups and mostly approximate the semantic distribution of the original image. In (c), we notice that the rank of interaction matrix with DSI is consistently larger than that without DSI in each SA layer, nearly approaching full rank, which verifies its advantages of retarding the rank collapse. To obtain deeper insights into of DPV, we use the t-SNE (van der Maaten and Hinton 2008) algorithm to project the pooling representation obtained from SA into a two-dimensional Euclid space. Afterwards, we randomly sample nearly $1K$ instances from the original distribution, and cluster these 2D vectors into 8 answer groups in 8 colors. As shown in Figure 4, we can see that after DPV, the representation distribution is prone to be sphere-gathered and relatively detached. These observations underline the role of DPV in enhancing representation discriminative learning. Beyond that, we showcase the behavior DPV in Figure 5. Taking the first case as example, the question asks for the objects located at foot of bed, while being undisturbed by information mess derived from desk or printer, we observe that the learned pillar score distribution is basically consistent with the location of answer clues, which shows the rationality and effectiveness of DPV.

Conclusion

In this paper, we launch social-like transformer based on *social network theory*, which facilitates dynamic structured interaction and enhance the discriminative representation by pillar verification. To verify their gains, we incorporate them into several transformer-based methods, and the extensive experiments convincingly demonstrate the superiority of our method based on the observation on quantitative results, interaction pattern, and representation visualization.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*, 6077–6086. IEEE.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*, 2425–2433. IEEE.

Cadène, R.; Ben-younes, H.; Cord, M.; and Thome, N. 2019. MUREL: Multimodal Relational Reasoning for Visual Question Answering. In *CVPR*, 1989–1998. IEEE.

Chen, L.; Ma, W.; Xiao, J.; Zhang, H.; and Chang, S.-F. 2021. Ref-NMS: Breaking Proposal Bottlenecks in Two-Stage Referring Expression Grounding. In *AAAI*, volume 35, 1036–1044.

Chen, X.; Ma, L.; Chen, J.; Jie, Z.; Liu, W.; and Luo, J. 2018. Real-time referring expression comprehension by single-stage grounding network. *arXiv preprint arXiv:1812.03426*.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. Transvg: End-to-end visual grounding with transformers. In *CVPR*, 1769–1779.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186. Association for Computational Linguistics.

Dong, C.; Wang, G.; Xu, H.; Peng, J.; Ren, X.; and Liang, X. 2021. EfficientBERT: Progressively Searching Multilayer Perceptron via Warm-up Knowledge Distillation. In *EMNLP*, 1424–1437. ACL.

Dong, Y.; Cordonnier, J.; and Loukas, A. 2021. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In *ICML*, volume 139, 2793–2803. PMLR.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *EMNLP*, 457–468. ACL.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C. H.; Wang, X.; and Li, H. 2019. Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering. In *CVPR*, 6639–6648. IEEE.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *ICCV*, 6325–6334. IEEE.

Guo, W.; Zhang, Y.; Yang, J.; and Yuan, X. 2021. Re-Attention for Visual Question Answering. *TIP*, 30: 6730–6743.

Han, Y.; Guo, Y.; Yin, J.; Liu, M.; Hu, Y.; and Nie, L. 2021. Focal and Composed Vision-semantic Modeling for Visual Question Answering. In *ACM Multimedia*, 4528–4536. ACM.

Han, Y.; Yin, J.; Wu, J.; Wei, Y.; and Nie, L. 2023. Semantic-Aware Modular Capsule Routing for Visual Question Answering. *TIP*, 32: 5537–5549.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778. IEEE.

Hedi; Cadène, R.; Cord, M.; and Thome, N. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *ICCV*, 2631–2639. IEEE.

Hedi; Cadène, R.; Thome, N.; and Cord, M. 2019. BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection. In *AAAI*, 8102–8109. AAAI Press.

Hong, R.; Liu, D.; Mo, X.; He, X.; and Zhang, H. 2019. Learning to compose and reason with language tree structures for visual grounding. *TPAMI*.

Hu, R.; Andreas, J.; Darrell, T.; and Saenko, K. 2018. Explainable Neural Computation via Stack Neural Module Networks. In *ECCV*, 55–71. Springer.

Hu, R.; Rohrbach, M.; Andreas, J.; Darrell, T.; and Saenko, K. 2017. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 1115–1124.

Huang, B.; Lian, D.; Luo, W.; and Gao, S. 2021. Look Before You Leap: Learning Landmark Features for One-Stage Visual Grounding. In *CVPR*, 16888–16897.

Jiang, H.; Misra, I.; Rohrbach, M.; Learned-Miller, E. G.; and Chen, X. 2020. In Defense of Grid Features for Visual Question Answering. In *CVPR*, 10264–10273. IEEE.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. B. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *CVPR*, 1988–1997. IEEE.

Kim, J.; and Jun, J. 2018. Bilinear Attention Networks. In *NeurIPS*, 1571–1581. MIT Press.

Kim, J.; On, K. W.; Lim, W.; Kim, J.; Ha, J.; and Zhang, B. 2017. Hadamard Product for Low-rank Bilinear Pooling. In *ICLR*.

Li, L.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Relation-Aware Graph Attention Network for Visual Question Answering. In *ICCV*, 10312–10321. IEEE.

Liao, Y.; Liu, S.; Li, G.; Wang, F.; Chen, Y.; Qian, C.; and Li, B. 2020. A real-time cross-modality correlation filtering method for referring expression comprehension. In *CVPR*, 10880–10889.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, 740–755. Springer.

Liu, D.; Zhang, H.; Wu, F.; and Zha, Z.-J. 2019a. Learning to assemble neural module tree networks for visual grounding. In *CVPR*, 4673–4682.

Liu, D.; Zhang, H.; Zha, Z.; and Wu, F. 2019b. Learning to Assemble Neural Module Tree Networks for Visual Grounding. In *ICCV*, 4672–4681. IEEE.

Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *ICLR*. OpenReview.net.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016a. Generation and Comprehension of

Unambiguous Object Descriptions. In *ICCV*, 11–20. IEEE Computer Society.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016b. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 11–20.

Nagaraja, V. K.; Morariu, V. I.; and Davis, L. S. 2016. Modeling context between objects for referring expression understanding. In *ECCV*, 792–807. Springer.

Peng, L.; Yang, Y.; Wang, Z.; Wu, X.; and Huang, Z. 2019. CRA-Net: Composed Relation Attention Network for Visual Question Answering. In *MM*, 1202–1210. ACM.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I. D.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *CVPR*, 658–666. IEEE.

Shi, J.; Zhang, H.; and Li, J. 2019. Explainable and Explicit Visual Reasoning Over Scene Graphs. In *CVPR*, 8376–8384. IEEE.

Tabassum, S.; Pereira, F. S. F.; Fernandes, S.; and Gama, J. 2018. Social network analysis: An overview. *Wiley Interdisciplinary Reviews Data Mining Knowledge Discovery*, e1256.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *CVPR*, 5099–5110. ACL.

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2020. Learning to Compose Dynamic Tree Structures for Visual Contexts. In *CVPR*, 6619–6628. IEEE.

van der Maaten, L.; and Hinton, G. 2008. Viualizing data using t-SNE. *JMLR*, 9: 2579–2605.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *CVPR*, 5998–6008.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *ICLR*.

Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and Hengel, A. v. d. 2019a. Neighbourhood Watch: Referring Expression Comprehension via Language-guided Graph Attention Networks. In *CVPR*.

Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and Hengel, A. v. d. 2019b. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *CVPR*, 1960–1968.

Wang, Z.; Wang, K.; Yu, M.; Xiong, J.; Hwu, W.; Hasegawa-Johnson, M.; and Shi, H. 2021. Interpretable Visual Reasoning via Induced Symbolic Space. In *ICCV*, 1858–1867. IEEE.

Wu, C.; Liu, J.; Wang, X.; and Dong, X. 2018. Object-Difference Attention: A Simple Relational Attention for Visual Question Answering. In *MM*, 519–527. ACM.

Yang, L.; Xu, Y.; Yuan, C.; Liu, W.; Li, B.; and Hu, W. 2022. Improving Visual Grounding with Visual-Linguistic Verification and Iterative Reasoning. In *CVPR*.

Yang, S.; Li, G.; and Yu, Y. 2019a. Dynamic Graph Attention for Referring Expression Comprehension. *Cornell University - arXiv,Cornell University - arXiv*.

Yang, S.; Li, G.; and Yu, Y. 2019b. Dynamic graph attention for referring expression comprehension. In *CVPR*, 4644–4653.

Yang, Z.; Chen, T.; Wang, L.; and Luo, J. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *ECCV*, 387–404. Springer.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 2019. A fast and accurate one-stage approach to visual grounding. In *CVPR*, 4683–4693.

Yi, K.; Wu, J.; Gan, C.; Torralba, A.; Kohli, P.; and Tenenbaum, J. 2018. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In *NeurIPS*, 1039–1050.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018a. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *CVPR*.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018b. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 1307–1315.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling Context in Referring Expressions. In *ECCV*, volume 9906 of *Lecture Notes in Computer Science*, 69–85. Springer.

Yu, Z.; Cui, Y.; Yu, J.; Wang, M.; Tao, D.; and Tian, Q. 2020. Deep Multimodal Neural Architecture Search. In *MM*, 3743–3752. ACM.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep Modular Co-Attention Networks for Visual Question Answering. In *CVPR*, 6281–6290. IEEE.

Zhang, H.; Niu, Y.; and Chang, S.-F. 2018. Grounding referring expressions in images by variational context. In *CVPR*, 4158–4166.

Zhou, Y.; Ji, R.; Su, J.; Sun, X.; and Chen, W. 2019. Dynamic Capsule Attention for Visual Question Answering. In *AAAI*, 9324–9331. AAAI Press.

Zhou, Y.; Ji, R.; Sun, X.; Luo, G.; Hong, X.; Su, J.; Ding, X.; and Shao, L. 2020. K-armed Bandit based Multi-Modal Network Architecture Search for Visual Question Answering. In *ACMMM*, 1245–1254. ACM.

Zhou, Y.; Ren, T.; Zhu, C.; Sun, X.; Liu, J.; Ding, X.; Xu, M.; and Ji, R. 2021. TRAR: Routing the Attention Spans in Transformer for Visual Question Answering. In *ICCV*, 2054–2064. IEEE.

Zhuang, B.; Wu, Q.; Shen, C.; Reid, I.; and Van Den Hengel, A. 2018. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *CVPR*, 4252–4261.