Structured Co-reference Graph Attention for Video-grounded Dialogue

Junyeong Kim, Sunjae Yoon, Dahyun Kim, Chang D. Yoo
Korea Advanced Institute of Science and Technology (KAIST)
{junyeong.kim, sunjae.yoon, dahyun.kim, cd_yoo}@kaist.ac.kr

Abstract

A video-grounded dialogue system referred to as the Structured Co-reference Graph Attention (SCGA) is presented for decoding the answer sequence to a question regarding a given video while keeping track of the dialogue context. Although recent efforts have made great strides in improving the quality of the response, performance is still far from satisfactory. The two main challenging issues are as follows: (1) how to deduce co-reference among multiple modalities and (2) how to reason on the rich underlying semantic structure of video with complex spatial and temporal dynamics. To this end, SCGA is based on (1) Structured Co-reference Resolver that performs dereferencing via building a structured graph over multiple modalities, (2) Spatio-temporal Video Reasoner that captures local-to-global dynamics of video via gradually neighboring graph attention. SCGA makes use of pointer network to dynamically replicate parts of the question for decoding the answer sequence. The validity of the proposed SCGA is demonstrated on AVSD@DSTC7 and AVSD@DSTC8 datasets, a challenging video-grounded dialogue benchmarks, and TVQA dataset, a large-scale videoQA benchmark. Our empirical results show that SCGA outperforms other state-of-the-art dialogue systems on both benchmarks, while extensive ablation study and qualitative analysis reveal performance gain and improved interpretability.

Introduction

Understanding visual information along with the natural language appears to be a desiderata in our community. Thus far, notable progress has been made towards bridging the fields of computer vision and natural language processing that includes video moment retrieval (Ma et al. 2020), image-grounded question answering (Antol et al. 2015; Anderson et al. 2018) / dialogue (Das et al. 2017; de Vries et al. 2017), and video-grounded question answering (Tapaswi et al. 2016; Lei et al. 2018) / dialogue (Alamri et al. 2019a). Among those, we focus on video-grounded dialogue system (VGDS) that allows an AI agent to 'observe' (i.e., understand a video) and 'converse' (i.e., communicate the understanding in a dialogue). To be specific, given a video, dialogue history consisting of a series of QA pairs, and a follow-up question about the video, the goal is to infer a free-form natural language answer to the question. Video-grounded dialogue appears often in many real-world human-computer conversations, and VGDS can potentially provide assistance to various subsections of the population especially to those subgroups suffering from sensory impairments. Although recent years has witnessed impressive advancement in performance, current VGDSs are still struggling with the following two challenging issues: (1) how to fully co-reference among multiple modalities, and (2) how to reason on the rich underlying semantic structure of video with complex spatial and temporal dynamics.

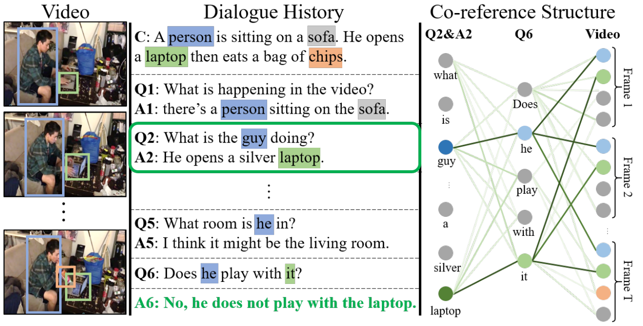

Video
Dialogue History
Co-reference Structure
C: A person is sitting on a sofa. He opens a laptop then eats a bag of chips.
Q1: What is happening in the video?
A1: there's a person sitting on the sofa.
Q2: What is the guy doing?
A2: He opens a silver laptop.
Q5: What room is he in?
A5: I think it might be the living room.
Q6: Does he play with it?
A6: No, he does not play with the laptop.
Q2&A2
Q6
Video
what
is
guy
a
silver
laptop
Does
he
play
with
it
Frame 1
Frame 2
Frame T

Figure 1: Illustration of the intuition behind SCGA as a video-grounded dialogue system. Left: video with detected objects, Middle: dialogue history, current Q&A. Most informative history is indicated by green box. Right: structured co-reference graph representing the underlying semantic dependencies between nodes (darker links indicate higher dependencies).

The first challenging issue of co-referencing among multiple modalities is illustrated in Figure 1: To the question in Q6 "Does he play with it?" contains pronouns for which the noun referent or antecedent must be identified. Henceforth, this task of identifying the pronoun's antecedent will be refer to as "dereferencing". Our study shows that nearly all dialogues and 61% of questions in the audio visual scene-aware dialogue (AVSD) dataset contains at least one pronoun (e.g., "it", "they") which makes "dereferencing" indispensable. Existing VGDSs have treated dialogue history just as any another input modality, but by recognizing and resolving its unique issue regarding pronoun reference, the quality of the output response can be enhanced significantly. For the first time, this paper proposes a VGDS that performs textual and visual co-reference via structured co-reference graph that identifies the pronoun's antecedents with nouns in the prior dialog and also with detected objects in video. This approach can be considered as an extension of prior efforts on visual dialogue (VisDial) to resolve visual co-reference issues via attention memory (Seo et al. 2017), reference pool (Kottur et al. 2018) and recursive attention (Niu et al. 2019).

The second challenging issue is to perform reasoning on rich underlying semantic structure of video with complex spatial and temporal dynamics. Majority of prior efforts on video-grounded dialogue task utilize holistic video feature from I3D model (Carreira and Zisserman 2017) that has been pre-trained on action recognition dataset, and they rely on a fully-connected transformer architecture to implicitly learn to infer the answer (Le et al. 2019b; Le and Chen 2020; Lee et al. 2020). These efforts underestimate the value of fine-grained visual representation from detector (Ren et al. 2015; Vu et al. 2019), consequently lacking the capability to apprehend relevant objects and their relationships and temporal evolution. We design a spatio-temporal video graph over object-level representation and perform graph attention for comprehensive understanding of the video.

In this paper, we address the aforementioned challenging issues with our Structured Co-reference Graph Attention (SCGA) which is composed of (1) structured co-reference resolver that performs dereferencing via building a structured co-reference graph, (2) spatio-temporal video reasoner that captures local-to-global dynamics of video via gradually neighboring graph attention (GN-GAT). We first *select* a key dialog history that can resolve the pronoun's antecedent in the follow-up question. Gumbel-Softmax (Jang, Gu, and Poole 2016; Maddison, Mnih, and Teh 2016) enables us to perform *discrete* attention over dialogue history. We propose a *bipartite* structured co-reference graph over multiple modalities and perform graph attention to integrate informative semantics from the key dialog history to question and video, as shown in Figure 1. We then build spatio-temporal video graph that represents spatial and temporal relations among objects. Motivated by recent studies that each head in self-attention independently looks at same global context, learning redundant features (Voita et al. 2019; Kant et al. 2020), we propose *gradually neighboring* graph attention (GN-GAT) that is guided by constructed video graph. Rather than repeatedly calculating self-attention over the common neighborhood, each head looks at a different neighborhood defined by its unique adjacency matrix. Each adjacency matrix will have a unique connectivity that link nodes reached within a fix number of hops. Thus, heads associated with smaller number of hops will consider local context while heads associated a larger number of hops will be looking at the global context. Finally based on the observation that words used in response come from words used in the question (e.g., response to question “What did he do after closing the window?” can be “He [context verb] after closing the window”), a pointer network is incorporated into the response sequence decoder to either decode from a fixed vocabulary set or from words used in the question.

Related Work

Video-grounded Dialogues

Visual Question Answering (VQA) (Antol et al. 2015) has been considered as a proxy task to evaluate the model's understanding on vision and language. In recent years, video-grounded dialogue systems (Alamri et al. 2019a; Hori et al. 2019a) have been proposed to advance VQA to hold meaningful dialogue with humans, grounded on video. VGDS incorporating recurrent neural network to encode dialog history is considered in (Hori et al. 2019a; Nguyen et al. 2019; Le et al. 2019a; Sanabria, Palaskar, and Metze 2019). Transformer based VGDS has been considered with query-aware attention (Le et al. 2019b), word-embedding attention decoder (Lee et al. 2020), and pointer-augmented decoder (Le and Chen 2020). VGDS that generates scene-graph every frame and aggregates it over the temporal axis to model fine-grained information flow in videos has also been considered (Geng et al. 2020). These systems would perform better with (1)”dereferencing” capability and (2) capability to capture and reason on complex spatial and temporal dynamics of the video.

Co-reference Resolution

Co-reference resolution is a task that was first defined in the linguistic community (Bergsma and Lin 2006), whose objective is to build association between named entities and references. It would include the task of identifying or associating the pronoun's antecedent from the dialog history and detected objects. While none of the past VGDSs have explicitly considered co-reference resolution, a number of systems in the VisDial have attempted to resolve visual co-reference. In (Seo et al. 2017), attention memory stores a sequence of previous (attention, key) pairs and the most relevant previous pair for the current question is retrieved. In (Kottur et al. 2018), a neural module network refers back to entities from previous rounds of dialogue and reuses its associated entities. In (Niu et al. 2019), the dialog history is browses until the agent has sufficient confidence in the visual co-reference resolution, and refines the visual attention recursively. Taking co-reference resolution to another level from just focusing on visual co-reference, SCGA conducts both textual and visual co-reference via structured co-reference graph.

Graph-based Visual Reasoning

Prior works have used GCN (Kipf and Welling 2017) or GAT (Veličković et al. 2017) to enable relational reasoning for image captioning (Yao et al. 2018), VQA (Teney, Liu, and van den Hengel 2017; Li et al. 2019), and VideoQA (Jiang and Han 2020). Fully-connected graph between objects (Teney, Liu, and van den Hengel 2017), objects / words (Jiang and Han 2020) or spatial / semantic graph (Yao et al. 2018; Li et al. 2019) is constructed to link different objects in relationship. By contrast, we construct *bipartite* graph between multiple modalities to form structured co-reference graph. Further, we build video graph not only in spatial-axis but also in temporal axis. We also provide different role for each head in gradually neighborhood graph attention to prevent learning redundant features.

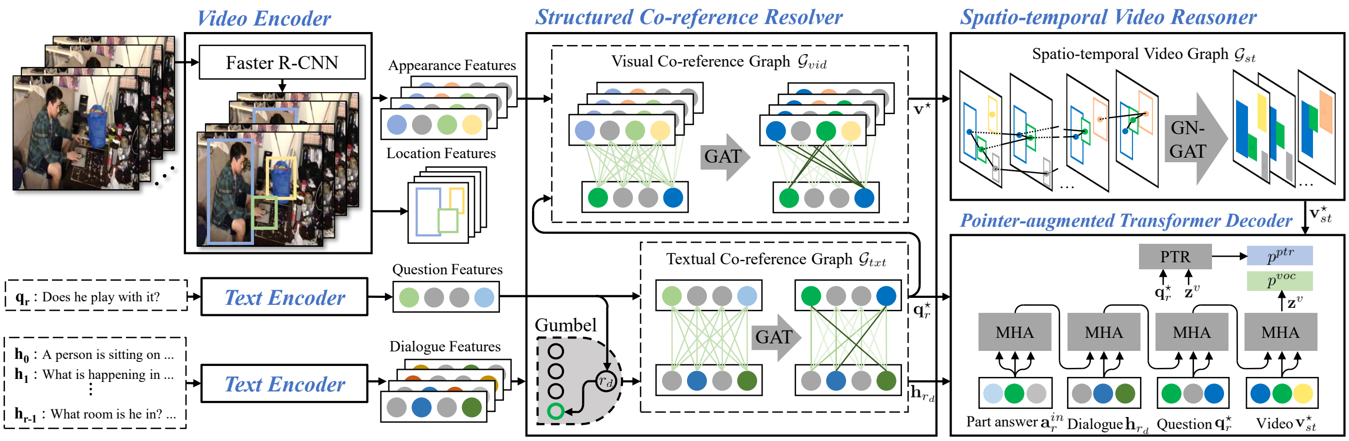

Video Encoder
Faster R-CNN
Appearance Features
Location Features
Question Features
Dialogue Features
$\mathbf{q_r}$: Does he play with it?
Text Encoder
$\mathbf{h_0}$: A person is sitting on ...
$\mathbf{h_1}$: What is happening in ...
$\mathbf{h_{r-1}}$: What room is he in? ...
Text Encoder
Structured Co-reference Resolver
Visual Co-reference Graph $\mathcal{G}_{vid}$
GAT
$\mathbf{v}^\star$
Textual Co-reference Graph $\mathcal{G}_{txt}$
GAT
Gumbel
r_d
$\mathbf{q}_r^\star$
$\mathbf{h}_{r_d}$
Spatio-temporal Video Reasoner
Spatio-temporal Video Graph $\mathcal{G}_{st}$
GN-GAT
$\mathbf{v}_{st}^\star$
Pointer-augmented Transformer Decoder
PTR
p^{ptr}
p^{voc}
$\mathbf{q}_r^\star$
$\mathbf{z}^v$
$\mathbf{z}^v$
MHA
MHA
MHA
MHA
Part answer $\mathbf{a}_r^{in}$
Dialogue $\mathbf{h}_{r_d}$
Question $\mathbf{q}_r^\star$
Video $\mathbf{v}_{st}^\star$

Figure 2: Illustration of Structured Co-reference Graph Attention (SCGA) which is composed of: (1) Input Encoder, (2) Structured Co-reference Resolver, (3) Spatio-temporal Video Reasoner, (4) Pointer-augmented Transformer Decoder.

Method

Here a formal definition of the video-grounded dialogue task is provided (Alamri et al. 2019a). We are given tuples of (v, h, q_r), consisting of a video v, the dialogue history $h = \{c, (q_1, a_1), \cdots, (q_{r-1}, a_{r-1})\}$, and a question q_r asked at current round $r \in \{1, \cdots, R\}$. The dialogue history itself is a set of question-answer pairs of previous rounds with a caption c in the beginning. The goal of video-grounded dialogue is to generate free-form natural language answer a_r to the question.

Figure 2 shows a schematic of Structured Co-reference Graph Attention (SCGA), consisting of a Input Encoder, Structured Co-reference Resolver, Spatio-temporal Video Reasoner, and Pointer-augmented Transformer Decoder. For the Video Encoder, Faster R-CNN (Ren et al. 2015) is used to extract sets of objects $v^t = \{v_o^t\}_{o=1}^{O}$ for each frame of $t \in \{1, \cdots, T\}$, where each object v_o^t is represented with an appearance feature vector $\mathbf{v}_o^t \in \mathbb{R}^{d_v}$ and location feature $\mathbf{b}_o^t \in \mathbb{R}^{d_b}$ ($T = 15$, $O = 6$, $d_v = 2048$, and $d_b = 4$) in our experiment. Each location feature $\mathbf{b}_o^t = [x, y, w, h]$ represents a spatial coordinate, where $[x, y]$ denotes the relative coordinate of top-left point of the b-box while $[w, h]$ denotes the width and height of the box. For the Text Encoder, we use a trainable token-level embedding layer to map sequence of token indices into d-dimensional feature representations. To incorporate ordering information of source tokens, we apply positional encoding (Vaswani et al. 2017) with layer normalization (Ba, Kiros, and Hinton 2016) on top of embedding layer. The encoded question (q_r) and each of dialogue history ($\{h_i\}_{i=1}^{r-1}$) are defined as:

$$\mathbf{q}_r = \text{LN}(\phi(q_r) + \text{PE}(q_r)) \in \mathbb{R}^{N_{q_r} \times d}, \quad (1)$$

$$\mathbf{h}_i = \text{LN}(\phi(h_i) + \text{PE}(h_i)) \in \mathbb{R}^{N_{h_i} \times d}, \quad (2)$$

where N_x denotes the number of tokens of sequence x. The followings sub-sections will explain the details of remaining model components.

Structured Co-reference Resolver

Textual Co-reference Resolution. We observed that there exists one key dialogue history that can resolve co-reference in the current question. To inject semantic information from the dialogue history into the question token representation for textual co-reference resolution, we first determine a key dialogue history, and then we let question tokens to attend to key dialogue history tokens. In our framework, we implement those functions via Gumbel-Softmax (Jang, Gu, and Poole 2016; Maddison, Mnih, and Teh 2016) to perform *discrete* attention over dialogue histories and graph attention over *bipartite* graph that connects all of the question tokens to all of the dialogue tokens. In this manner, question tokens learn to implicitly integrate informative semantics from the key dialogue history to its representation.

One can easily suppose that the current question simply follows from the latest dialogue history. However, sometimes the question requires looking back at an earlier dialogue, which means there are no relationships between the current question and recent dialogue histories. Inspired by Gumbel-Max trick with continuous softmax relaxations (Niu et al. 2019), we select a most relevant dialogue history h_{r_d} for current question q_r. Our approach is end-to-end trainable while making *discrete* decision, thanks to Gumbel-Softmax. We first calculate matching score $s_{r,i}$ between question feature $\mathbf{q}_r$ and each of dialogue history features $\{\mathbf{h}_0, \cdots, \mathbf{h}_{r-1}\}$:

$$e_{r,i} = f_e([f_q(\mathcal{A}(\mathbf{q}_r)) || f_h(\mathcal{A}(\mathbf{h}_i))]), \quad (3)$$

$$s_{r,i} = f_s([e_{r_i} || \Delta_{r,i}]), \quad (4)$$

where $[\cdot||\cdot]$ denotes concatenation operation, $\mathcal{A}$ represents average operation on word-axis, and f_x is a fully-connected layer with input x. Here, $\Delta_{r,i} = r - i$ provides distance information between q_r and h_i in the dialogue history. Gumbel-Softmax produces a r-dimensional one-hot vector

$\mathbf{g}_r$ for *discrete* attention over dialogue histories:

$$\mathbf{g}_r = \text{Gumbel\_Softmax}(\mathbf{s}_r), \tag{5}$$

$$\mathbf{h}_{r_d} = \sum_{i=0}^{r-1} g_{r,i} \cdot \mathbf{h}_i. \tag{6}$$

We formally define the textual co-reference graph $\mathcal{G}_{txt} = (\mathcal{V}_{txt}, \mathcal{E}_{txt})$ by treating each token from question q_r and related dialogue history h_{r_d} as graph nodes. In designing a co-reference resolver, we construct a *bipartite* graph and perform graph attention (Veličković et al. 2017) to inject useful semantic information from the related dialogue history into query representation. We first concatenate the $\mathbf{q}_r$ and $\mathbf{h}_{r_d}$ along word-axis to make a heterogeneous node matrix:

$$\boldsymbol{\mathcal{V}}_{txt} = [\mathbf{q}_r \mid\mid \mathbf{h}_{r_d}] \in \mathbb{R}^{N_{txt} \times d}, \tag{7}$$

where $N_{txt} = N_{q_r} + N_{h_{r_d}}$ is the number of nodes. Multi-head self-attention is then performed to model relations between each node and its neighboring nodes. For each head k, attention coefficient $\alpha^k_{i,j}$ denoting the relevance between two linked node $\boldsymbol{\mathcal{V}}_i$ and $\boldsymbol{\mathcal{V}}_j$ is calculated as:

$$\alpha^k_{i,j} = \frac{\exp(\sigma(a_k^\top [W^k \boldsymbol{\mathcal{V}}_i || W^k \boldsymbol{\mathcal{V}}_j])}{\sum_{n \in \mathcal{N}_i} \exp(\sigma(a_k^\top [W^k \boldsymbol{\mathcal{V}}_i || W^k \boldsymbol{\mathcal{V}}_n]))}, \tag{8}$$

where $\sigma(\cdot)$ is a nonlinear function such as LeakyReLU, $a_k \in \mathbb{R}^{2d}$ is the attention weight vector, $W^k \in \mathbb{R}^{d \times d}$ is the shared projection matrix, and $\mathcal{N}_i$ is the neighborhood of node i. Graph node features are updated by going through K independent attention mechanisms, and concatenating their output features:

$$\boldsymbol{\mathcal{V}}^\star_i = ||^K_{k=1} \sigma(\sum_{j \in \mathcal{N}_i} \alpha^k_{i,j} W^k \boldsymbol{\mathcal{V}}_j). \tag{9}$$

At the end, we add original $\boldsymbol{\mathcal{V}}_i$ to updated $\boldsymbol{\mathcal{V}}^\star_i$ and pick nodes corresponding to question tokens to serve as the final co-reference resolved question representation:

$$\mathbf{q}^\star_r = (\boldsymbol{\mathcal{V}}^\star_{txt} + \boldsymbol{\mathcal{V}}_{txt})[: N_{q_r}] \in \mathbb{R}^{N_{q_r} \times d}, \tag{10}$$

where $[: i]$ denotes slicing operation along node-axis.

Visual Co-reference Resolution. Pipeline of visual co-reference resolution is analogous to textual co-reference resolution. Video objects learn to implicitly integrate informative semantics from a co-reference resolved question to its representation. We first project each appearance feature $\mathbf{v}^t_k$ into d-dimensional space with linear transformation (where d is the same as in the text embedding). We also add layer normalization (Ba, Kiros, and Hinton 2016) on top of linear transform to ensure that the appearance feature has same scale as text representation $\mathbf{v}^t_o := \text{LN}(W\mathbf{v}^t_o)$. Again, we construct *bipartite* visual co-reference graph $\mathcal{G}_{vid} = (\mathcal{V}_{vid}, \mathcal{E}_{vid})$ by treating every objects from video and token from co-reference resolved question as graph nodes:

$$\boldsymbol{\mathcal{V}}_{vid} = [\mathbf{v} \mid\mid \mathbf{q}^\star_r] \in \mathbb{R}^{N_{vid} \times d}, \tag{11}$$

where $N_{vid} = N_v + N_{q_r}$, and $N_v = T \times O$. We perform graph attention over graph $\mathcal{G}_{vid}$ and obtain updated graph node $\boldsymbol{\mathcal{V}}^\star_{vid}$. Finally, we pick nodes corresponding to objects to serve as the final co-reference resolved video representation:

$$\mathbf{v}^\star = (\boldsymbol{\mathcal{V}}^\star_{vid} + \boldsymbol{\mathcal{V}}_{vid})[: N_v] \in \mathbb{R}^{N_v} \times d. \tag{12}$$

Spatio-temporal Video Reasoner

We first construct a spatio-temporal video graph $\mathcal{G}_{st} = (\mathcal{V}_{st}, \mathcal{E}_{st})$ that represents spatial and temporal among between detected objects, and perform our proposed *gradually neighboring* graph attention (GN-GAT) to reason on rich underlying semantic structure of video with complex spatial and temporal dynamics. Recent studies (Voita et al. 2019; Kant et al. 2020) show that multi-head self attention bares a limitation in learning redundant features due to repeated usage of same input context for every attention heads. Rather than repeatedly calculating self-attention over same neighborhood, we propose GN-GAT that each head considers different adjacency matrix whose connectivity gradually increases with distance with respect to the graph nodes. We can effectively model and reason on local-to-global context of spatial and temporal dynamics of video.

While using co-reference resolved video representation $\mathbf{v}^\star \in \mathbb{R}^{N_v \times d}$ as graph node $\mathcal{V}_{st}$, we define two sets of edge matrices, $\{E_t\}^T_{t=1}$ that capture spatial relations within each frame and $\{E^{t+1}_t\}^{T-1}_{t=1}$ that capture temporal relations between adjacent frames, to build graph edge $\mathcal{E}_{st}$. Here, we use location feature $\mathbf{b}^t_o$ to obtain E_t and E^{t+1}_t. Criterion for spatial relation matrix $E_t \in \mathbb{R}^{O \times O}$ is defined as $\max(\Delta_x, \Delta_y) < \tau_s$. More concretely, $E_t[i, j] = 1$ if i-th object and j-th object in frame t are close enough to match above criterion. Criterion for temporal relation matrix E^{t+1}_t is defined as $\max(\Delta_x, \Delta_y) < \tau_t$ with same object label. Again, $E^{t+1}_t[i, j] = 1$ if i-th object in frame t and j-th object in frame $t + 1$ are close enough and have same object label. Finally, graph edge $\mathcal{E}_{st} \in \mathbb{R}^{N_v \times N_v}$ is constructed as follows:

$$\mathcal{E}_{st} = \begin{bmatrix} E_1 & E^2_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ E^{2\top}_1 & E_2 & E^3_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & E^{3\top}_2 & E_3 & E^4_3 & \cdots & \mathbf{0} \\ \vdots & & & \ddots & & \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & E^{T\top}_{T-1} & E_T \end{bmatrix}. \tag{13}$$

The GN-GAT performs reasoning on constructed spatio-temporal video graph $\mathcal{G}_{st}$. Adjacency matrix for distance n among graph nodes can be calculated as boolean of n-th power of $\mathcal{E}_{st}$:

$$A_n = \text{Bool}(\mathcal{E}^n_{st}). \tag{14}$$

Each head of GN-GAT looks at adjacency matrix A_n with different distance n. In this way, we can effectively learn from local (i.e., lower distance n) to global (i.e., higher distance n) context of video gradually within a single graph attention layer. We assign more heads to higher n since global context consists much denser neighborhood compared to local context. Finally, we obtain $\mathbf{v}^\star_{st}$ by performing GN-GAT:

$$\mathbf{v}^\star_{st} = \text{GN-GAT}(\mathbf{v}^\star) \in \mathbb{R}^{N_v \times d}, \tag{15}$$

where GN-GAT is formulated as Equation 8,9 with different neighborhood defined by adjacency matrix for each head.

Pointer-augmented Transformer Decoder

Answer response is decoded by incorporating the question and video representations from the preceding model components. Following prior work (Hori et al. 2019a), we decode answer tokens in autoregressive manner. We design a Transformer decoder consists of 4 attention layers: masked self-attention to partially generated answer so far ($\mathbf{a}_r^{in}$), guided-attention to selected history ($\mathbf{h}_{r_d}$), guided-attention to question ($\mathbf{q}_r^\star$) from structured co-reference resolver, and guided attention to video ($\mathbf{v}_{st}^\star$) from spatio-temporal video reasoner. We further augment decoder with dynamic pointer network (Vinyals, Fortunato, and Jaitly 2015; Hu et al. 2020) to either decode token from fixed vocabulary or copy from question words based on the intuition that question tokens can form a structure of answer (e.g., "He [context verb] after closing the window" for question "What does he do after closing the window?").

Each attention on transformer decoder is multi-head attention (Vaswani et al. 2017) on query, key, and value tensors: Attention(Q, K, V). Our transformer decoder can be formulated as:

$$\mathbf{z}^a = \text{Attention}(\mathbf{a}_r^{in}, \mathbf{a}_r^{in}, \mathbf{a}_r^{in}), \tag{16}$$

$$\mathbf{z}^h = \text{Attention}(\mathbf{z}^a, \mathbf{h}_{r_d}, \mathbf{h}_{r_d}), \tag{17}$$

$$\mathbf{z}^q = \text{Attention}(\mathbf{z}^h, \mathbf{q}_r^\star, \mathbf{q}_r^\star), \tag{18}$$

$$\mathbf{z}^v = \text{Attention}(\mathbf{z}^q, \mathbf{v}_{st}^\star, \mathbf{v}_{st}^\star), \tag{19}$$

where $\mathbf{a}_r^{in} \in \mathbb{R}^{j \times d}$ is partially generated answer at j-th decoding step embedded with text encoder. Note that we mask the first self-attention layer to ensure causality in answer decoding. At j-th decoding step, we either choose word index from fixed vocabulary distribution p_j^{voc} or dynamic pointer distribution p_j^{ptr} through argmax operation on $p_j = [p_j^{voc} || p_j^{ptr}]$:

$$p_j^{voc} = g^{voc}(\mathbf{z}_j^v) \in \mathbb{R}^{||V||}, \tag{20}$$

$$p_j^{ptr} = g_q^{ptr}(\mathbf{q}_r^\star)^\top g_z^{ptr}(\mathbf{z}_j^v) \in \mathbb{R}^{N_{q_r}}, \tag{21}$$

where g^{voc} is a linear layer to vocabulary size $||V||$-dimension, and g_x^{ptr} is a linear layer to d-dimension. Logits from dynamic pointer network is obtained through bilinear interaction between question token representation (i.e., $g_q^{ptr}(\mathbf{q}_r^\star)$) and decoder output (i.e., $g_z^{ptr}(\mathbf{z}_j^v)$).

Optimization

During training, we use teacher-forcing (Lamb et al. 2016) to supervise each decoding steps, i.e., ground-truth tokens are used as decoder input. We train the model with multi-label binary cross entropy loss over concatenated token distribution p_j, since answer token can appear on both fixed vocabulary and question tokens. We add two special tokens to our fixed answer vocabulary, `<bos>` and `<eos>`, where `<bos>` is used as first step of decoder to indicates the beginning of sentence and `<eos>` denotes the end of sentence to stop the decoding process.

Experiments

Datasets

We validate our proposed SCGA on two recent datasets.

AVSD (Alamri et al. 2019a) is a widely used benchmark dataset for video-grounded dialogue, which are collected on the Charades (Sigurdsson et al. 2016) human-activity dataset. It contains 7,659, 1,787, 1,710 dialogues for training, validation and test, respectively. Each dialogue contains 10 dialogue turns, and each turn consists of a question and target response. For evaluation, 6 reference responses are provided. We provide experimental results on both AVSD@DSTC7 (Alamri et al. 2019b) and AVSD@DSTC8 (Hori et al. 2020) challenge benchmark.

TVQA (Lei et al. 2018) is a large-scale benchmark dataset for multi-modal video question answering, which consists multiple-choice QA pairs for short video clips and corresponding subtitles. It contains 122,039, 15,252, 7,623 QAs for training, validation and test, respectively. To fit our problem setting, we made some modifications to TVQA. Among multiple answer candidates, we select correct one to be a target response. We split training set of TVQA into training and validation set, and serve official validation set as test set in our experiments since test labels are not publicly available.

Experimental Details

Metrics. We follow official objective metrics for AVSD benchmark, including BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE-L (Lin 2004), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). The metrics are formulated to compute the word overlapping between each generated response and reference responses.

Model Hyperparameters. The dimension of hidden layer is set to $d = 512$, the number of attention heads for GAT and decoder is set to $K = 8$. Criterions for edge $\mathcal{E}_{st}$ are set to $\tau_s = 0.4$, $\tau_t = 0.2$ for sparse local connection. For GN-GAT, we set distance $n = 1, 2, 3, 4$, and $1, 1, 2, 4$ heads are assigned to each distance, respectively. All the hyperparameters were tuned via grid-search over validation set.

Training Details. Our model is trained on NVIDIA TITAN V (12GB of memory) GPU with Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$, and $\epsilon = 10^{-9}$. We adopt a learning rate strategy similar to (Vaswani et al. 2017), and set the learning rate warm-up strategy to $10,000$ training steps and trained model up to 20 epochs. We select the batch size of 32 and dropout rate of 0.3. For all experiments, we select the best model that achieves the lowest perplexity on the validation set. During inference, we adopt a beam search with a beam size of 5 and a length penalty of 1.0. The maximum length of output tokens are set to 30. The entire framework is implemented with PyTorch.

Results on AVSD Benchmark

Table 1 summarizes the experimental results on AVSD dataset. We compare SCGA with several baseline methods (please refer to Related Work for description on baseline methods). For fair comparison, we report the performances

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
AVSD@DSTC7							
Baseline (Hori et al. 2019a)	0.621	0.480	0.379	0.305	0.217	0.481	0.733
HMA (Le et al. 2019a)	0.633	0.490	0.386	0.310	0.242	0.515	0.856
RMFF (Yeh et al. 2019)	0.636	0.510	0.417	0.345	0.224	0.505	0.877
EE-DMN (Lin et al. 2019)	0.641	0.493	0.388	0.310	0.241	0.527	0.912
JMAN (Chu et al. 2020)	0.667	0.521	0.413	0.334	0.239	0.533	0.941
FA-HRED (Nguyen et al. 2019)	0.695	0.553	0.444	0.360	0.249	0.544	0.997
CMU (Sanabria, Palaskar, and Metze 2019)	0.718	0.584	0.478	0.394	0.267	0.563	1.094
MSTN (Lee et al. 2020)	-	-	-	0.377	0.275	0.566	1.115
JSTL (Hori et al. 2019b)	0.727	0.593	0.488	0.405	0.273	0.566	1.118
MTN (Le et al. 2019b)	0.731	0.597	0.490	0.406	0.271	0.564	1.127
MTN-P (Le and Chen 2020)	**0.750**	0.619	0.514	0.427	0.280	**0.580**	1.189
SCGA w/o caption	0.702	0.588	0.481	0.398	0.265	0.546	1.059
SCGA	0.745	**0.622**	**0.517**	**0.430**	**0.285**	0.578	**1.201**
AVSD@DSTC8							
MDMN (Xie and Iacobacci 2020)	-	-	-	0.296	0.214	0.496	0.761
JMAN (Chu et al. 2020)	0.645	0.504	0.402	0.324	0.232	0.521	0.875
STSGR (Geng et al. 2020)	-	-	-	0.357	0.267	0.553	1.004
MSTN (Lee et al. 2020)	-	-	-	0.385	0.270	0.564	1.073
MTN-P (Le and Chen 2020)	0.701	0.587	0.494	**0.419**	0.263	0.564	1.097
SCGA w/o caption	0.675	0.559	0.459	0.377	0.269	0.555	1.024
SCGA	**0.711**	**0.593**	**0.497**	0.416	**0.276**	**0.566**	**1.123**

Table 1: Experimental results on the test split of AVSD benchmark at DSTC7 and DSTC8 challenges.

of official six reference evaluation on AVSD@DSTC7 and AVSD@DSTC8, without using external data to pretrain the model. SCGA achieves the state-of-the-art performance against all baseline methods on majority of metrics. The result indicates that resolving co-reference amongst multiple modalities and capturing fine-grained local-to-global dynamics of video can help to generate quality response to boost model performance. We also provide experimental results without using caption, which simulates real-word video-grounded dialogue situation; we are only given video context and dialogue history. Competitive performance of SCGA w/o caption indicates that SCGA is able to reason on contextual cues from video.

Ablation Study

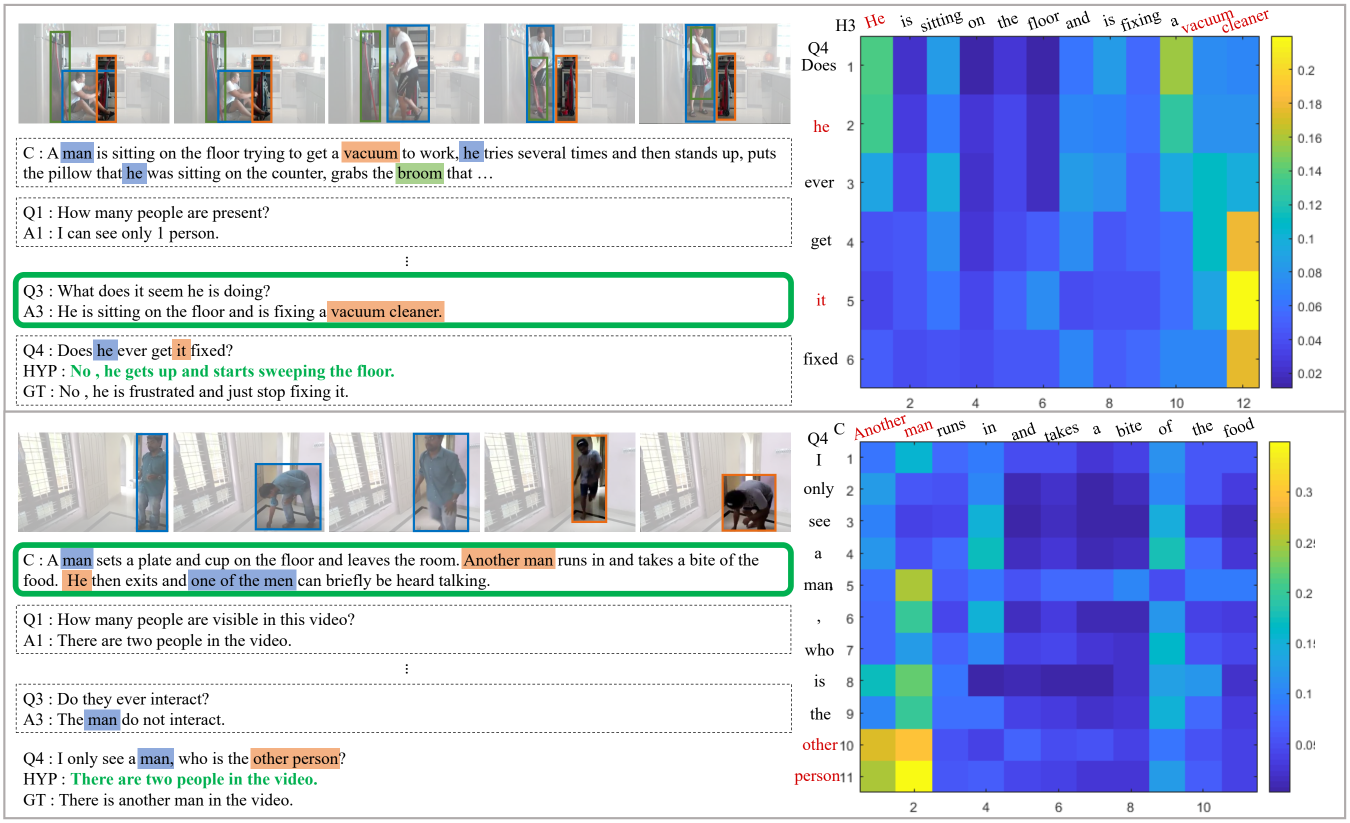
We experiment with several variants of SCGA in order to measure the effectiveness of the proposed key components. The first block of Table 2 provides the ablation results of structured co-reference resolver. While structured co-reference resolver boosts performance significantly, we can see that textual co-reference resolver is more important to integrate informative semantics from key dialogue history, improving CIDEr from 1.161 to 1.201. Without textual co-reference resolver, visual co-reference resolver also cannot work properly since co-reference in question tokens are not resolved. The second block of Table 2 provides the ablation results on spatio-temporal video reasoner. Without this module, spatial and temporal dynamics of video are implicitly learned through transformer decoder, which shows performance drop of 0.034 in CIDEr. We further provide results on different distance n for GN-GAT.

Model Variants	CIDEr
Full SCGA	1.201
w/o Textual Co-ref Resolver	1.161
w/o Visual Co-ref Resolver	1.189
w/o Structured Co-ref Resolver	1.152
w/o ST Video Reasoner	1.167
w/ distance $n = 1$	1.182
w/ distance $n = [1, 2]$	1.194
w/ distance $n = [1, 6]$	1.197
w/ distance $n = [1, 8]$	1.187

Table 2: Ablation study on model variants of SCGA on the test split of AVSD@DSTC7 benchmark.

Results on TVQA Benchmark

Other than AVSD results, we also report results to TVQA benchmark on our modified setting. We consider the subtitle corresponding to each video as dialogue history in our experiments on TVQA. We compare SCGA with two baseline methods that were reproduced using public codebase. Table 3 shows SCGA outperforms baseline methods on CIDEr metric. Subtitles plays an important role in providing clue for answering the question. Previous approaches to TVQA in multiple-choice setting (Kim et al. 2019a,b, 2020) attempted to locate key sentences through temporal attention or temporal localization. Our results on TVQA demonstrates that SCGA is able to not only resolve co-reference, but also locates a key sentence from subtitle.


C : A man is sitting on the floor trying to get a vacuum to work, he tries several times and then stands up, puts the pillow that he was sitting on the counter, grabs the broom that …

Q1 : How many people are present?
A1 : I can see only 1 person.

⋮

Q3 : What does it seem he is doing?
A3 : He is sitting on the floor and is fixing a vacuum cleaner.

Q4 : Does he ever get it fixed?
HYP : **No , he gets up and starts sweeping the floor.**
GT : No , he is frustrated and just stop fixing it.

C : A man sets a plate and cup on the floor and leaves the room. Another man runs in and takes a bite of the food. He then exits and one of the men can briefly be heard talking.

Q1 : How many people are visible in this video?
A1 : There are two people in the video.

⋮

Q3 : Do they ever interact?
A3 : The man do not interact.

Q4 : I only see a man, who is the other person?
HYP : **There are two people in the video.**
GT : There is another man in the video.

Figure 3: Visualization of structured co-reference graph attention (SCGA) in the test split of the AVSD@DSTC7 benchmark.

Methods	CIDEr
Baseline (Hori et al. 2019a)	0.781
MTN (Le et al. 2019b)	0.973
SCGA	1.062

Table 3: Experimental results on the TVQA benchmark.

Qualitative Analysis

Figure 3 visualizes the intermediate functionality of SCGA with samples from test split of the AVSD@DSTC7 benchmark. Each example is provided with a selected dialogue history (indicated by green box), learned attention weights for textual and visual co-reference graph. In figure 3, the question of upper example: 'Does he ever get it fixed' has a semantic relevance to dialogue history H3: 'He is sitting on the floor and fixing a vacuum cleaner' in the green box. The attention map on the right shows similarity between tokens of Q4 and tokens of H3. Specifically, 'it' token in Q4 refers to 'vacuum cleaner' tokens in H3, showing high attention values. Through the Visual Co-reference Resolver, the co-referenced objects in video are highlighted from other detected objects where the textual co-reference resolved question makes it easier to find the vacuum cleaner in the video. The figure below also shows that the high relevant sentence is selected and textual co-reference tokens are enhanced.

Conclusion

In this paper, VGDS referred to as Structured Co-reference Graph Attention (SCGA) is presented to consider two major challenging issues: (1) How to deduce co-reference among multiple modalities; (2) How to reason on the rich underlying semantic structure of video with complex spatial and temporal dynamics. SCGA is based on (1) Structured Co-reference Resolver that performs dereferencing via building a structured graph over multiple modalities, (2) Spatio-temporal Video Reasoner that captures both global and local dynamics of video via segmented self-attention layer. Furthermore, SCGA makes use of pointer network to dynamically replicate parts of the question for decoding the answer sequence. Our empirical results on AVSD@DSTC7, AVSD@DSTC8 and TVQA benchmarks show that SCGA achieves state-of-the-art performance.

Acknowledgments

This work was partly supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2017-0-01780) and partly supported by the MSIT(Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Problem(2020-0-01649) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and it has been supported by Microsoft Research.

References

Alamri, H.; Cartillier, V.; Das, A.; Wang, J.; Cherian, A.; Essa, I.; Batra, D.; Marks, T. K.; Hori, C.; Anderson, P.; Lee, S.; and Parikh, D. 2019a. Audio Visual Scene-Aware Dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alamri, H.; Hori, C.; Marks, T. K.; Batra, D.; and Parikh, D. 2019b. Audio Visual Scene-aware dialog (AVSD) Track for Natural Language Generation in DSTC7. In *DSTC7 at AAAI2019-W*.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* .

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W05-0909.

Bergsma, S.; and Lin, D. 2006. Bootstrapping Path-Based Pronoun Resolution. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chu, Y.-W.; Lin, K.-Y.; Hsu, C.-C.; and Ku, L.-W. 2020. Multi-step Joint-Modality Attention Network for Scene-Aware Dialogue System. In *DSTC8 at AAAI2020-W*.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M. F.; Parikh, D.; and Batra, D. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

de Vries, H.; Strub, F.; Chandar, S.; Pietquin, O.; Larochelle, H.; and Courville, A. 2017. GuessWhat?! Visual Object Discovery Through Multi-Modal Dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Geng, S.; Gao, P.; Marks, T.; Hori, C.; and Cherian, A. 2020. Spatio-Temporal Scene Grpah Reasoning for Audio Visual Scene-Aware Dialog at DSTC8. In *DSTC8 at AAAI2020-W*.

Hori, C.; Alamri, H.; Wang, J.; Wichern, G.; Hori, T.; Cherian, A.; Marks, T. K.; Cartillier, V.; Lopes, R. G.; Das, A.; Essa, I.; Batra, D.; and Parikh, D. 2019a. End-to-end Audio Visual Scene-aware Dialog Using Multimodal Attention-based Video Features. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Hori, C.; Cherian, A.; Hori, T.; and Marks, T. K. 2020. Audio Visual Scene-Aware Dialog (AVSD) Track for Natural Language Generation in DSTC8. In *DSTC8 at AAAI2020-W*.

Hori, C.; Cherian, A.; Marks, T. K.; and Hori, T. 2019b. Joint Student-Teacher Learning for Audio-Visual Scene-Aware Dialog. In *Proceedings of the Interspeech*.

Hu, R.; Singh, A.; Darrell, T.; and Rohrbach, M. 2020. Iterative Answer Prediction With Pointer-Augmented Multimodal Transformers for TextVQA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical Reparameterization with Gumbel-Softmax. In *arXiv preprint arXiv:1611.01144*.

Jiang, P.; and Han, Y. 2020. Reasoning with Heterogeneous Graph Alignment for Video Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Kant, Y.; Batra, D.; Anderson, P.; Schwing, A.; Parikh, D.; Lu, J.; and Agrawal, H. 2020. Spatially Aware Multimodal Transformers for TextVQA. *Proceedings of the European Conference on Computer Vision (ECCV)* .

Kim, J.; Ma, M.; Kim, K.; Kim, S.; and Yoo, C. D. 2019a. Gaining Extra Supervision via Multi-task Learning for Multi-Modal Video Question Answering. In *International Joint Conference on Neural Networks (IJCNN)*.

Kim, J.; Ma, M.; Kim, K.; Kim, S.; and Yoo, C. D. 2019b. Progressive Attention Memory Network for Movie Story Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kim, J.; Ma, M.; Pham, T.; Kim, K.; and Yoo, C. D. 2020. Modality Shifting Attention Network for Multi-Modal Video Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.

Kottur, S.; Moura, J. M. F.; Parikh, D.; Batra, D.; and Rohrbach, M. 2018. Visual Coreference Resolution in Visual Dialog using Neural Module Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Lamb, A. M.; ALIAS PARTH GOYAL, A. G.; Zhang, Y.; Zhang, S.; Courville, A. C.; and Bengio, Y. 2016. Professor Forcing: A New Algorithm for Training Recurrent Networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Le, H.; and Chen, N. F. 2020. Multimodal Transformer with Pointer Network for the DSTC8. In *DSTC8 at AAAI2020-W*.

Le, H.; Hoi, S. C.; Sahoo, D.; and Chen, N. F. 2019a. End-to-End Multimodal Dialog Systems with Hierarchical Multimodal Attention on Video Features. In *DSTC7 at AAAI2019-W*.

Le, H.; Sahoo, D.; Chen, N.; and Hoi, S. C. 2019b. Multimodal Transformer Networks for End-to-End Video-Grounded Dialogue Systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lee, H.; Yoon, S.; Dernoncourt, F.; Kim, D. S.; Bui, T.; and Jung, K. 2020. DSTC8-AVSD: Multimodal Semantic Transformer Network with Retrieval Style Word Generator. In *DSTC8 at AAAI2020-W*.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. TVQA: Localized, Compositional Video Question Answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Li, L.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Relation-Aware Graph Attention Network for Visual Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Lin, K.-Y.; Hsu, C.-C.; Chen, Y.-N.; and Ku, L.-W. 2019. Entropy-Enhanced Multimodal Attention Model for Scene-Aware Dialogue Generation. In *DSTC7 at AAAI2019-W*.

Ma, M.; Yoon, S.; Kim, J.; Lee, Y.; Kang, S.; and Yoo, C. D. 2020. VLANet: Video-Language Alignment Network for Weakly-Supervised Video Moment Retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* .

Nguyen, D. T.; Sharma, S.; Schulz, H.; and Asri, L. E. 2019. From FiLM to Video: Multi-turn Question Answering with Multi-modal Context. In *DSTC7 at AAAI2019-W*.

Niu, Y.; Zhang, H.; Zhang, M.; Zhang, J.; Lu, Z.; and Wen, J.-R. 2019. Recursive Visual Attention in Visual Dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. doi:10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Sanabria, R.; Palaskar, S.; and Metze, F. 2019. CMU Sinbad's Submissino for the DSTC7 AVSD Challenge. In *DSTC7 at AAAI2019-W*.

Seo, P. H.; Lehrmann, A.; Han, B.; and Sigal, L. 2017. Visual Reference Resolution using Attention Memory for Visual Dialog. In *Advances in Neural Information Processing Systems (NIPS)*.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. MovieQA: Understanding Stories in Movies Through Question-Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Teney, D.; Liu, L.; and van den Hengel, A. 2017. Graph-Structured Representations for Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems (NIPS)*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* .

Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer Networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; and Titov, I. 2019. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Vu, T.; Jang, H.; Pham, T.; and Yoo, C. D. 2019. Cascade RPN: Devling into High-Quality Region Proposal Network with Adaptive Convolution. In *Advances in Neural Information Processing Systems (NIPS)*.

Xie, H.; and Iacobacci, I. 2020. Audio Visual Scene-Aware Dialog System Using Dynamic Memory Networks. In *DSTC8 at AAAI2020-W*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring Visual Relationship for Image Captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yeh, Y.-T.; Lin, T.-C.; Cheng, H.-H.; Deng, Y.-H.; Su, S.-Y.; and Chen, Y.-N. 2019. Reactive Multi-Stage Feature Fusion for Multimodal Dialogue Modeling. In *DSTC7 at AAAI2019-W*.