Article

CE-BART: Cause-and-Effect BART for Visual Commonsense Generation

Junyeong Kim [1], Ji Woo Hong [2], Sunjae Yoon [3], and Chang D. Yoo [4]*

[1] Chung-Ang University; junyeongkim@cau.ac.kr
[2] Korea Advanced Institute of Science and Technology; jiwoohong93@kaist.ac.kr
[3] Korea Advanced Institute of Science and Technology; sunjae.yoon@kaist.ac.kr
[4] Korea Advanced Institute of Science and Technology; cd_yoo@kaist.ac.kr
* Correspondence: cd_yoo@kaist.ac.kr; Tel.: +82-10-3774-1007

Abstract: "A Picture is worth a thousand words". Given an image, humans are able to deduce various cause-and-effect captions of past, current, and future events beyond the image. The task of visual commonsense generation aims at generating three cause-and-effect captions (1) what needed to happen *before*, (2) what is the current *intent*, and (3) what will happen *after* for a given image. However, such a task is challenging for machines owing to two limitations: existing approaches (1) directly utilize conventional vision-language transformers to learn relationships between input modalities, and (2) ignore relations among target cause-and-effect captions but consider each caption independently. We propose Cause-and-Effect BART (CE-BART) which is based on (1) Structured Graph Reasoner that captures intra- and inter-modality relationships among visual and textual representations, and (2) Cause-and-Effect Generator that generates cause-and-effect captions by considering the causal relations among inferences. We demonstrate the validity of CE-BART on VisualCOMET and AVSD benchmarks. CE-BART achieves SOTA performances on both benchmarks, while extensive ablation study and qualitative analysis demonstrate the performance gain and improved interpretability.

Keywords: Deep Learning; Visual-Language Reasoning; Visual Commonsense Generation; Video-grounded Dialogue; VisualCOMET; AVSD

1. Introduction

Visual Commonsense Generation (VCG) [1] is a challenging task that requires generating commonsense and cause-and-effect captions regarding visual and textual information. To be specific, given a still image and a description about the event shown in that image, the goal is to understand the cause-and-effect relations within the event and generate free-form natural language sentences that describe the inferred past/future events and the present intents of characters in the image. For example in Fig 1, given an image on the left of a woman approaching to a man at the table, agent generates three kinds of cause-and-effect captions: (1) sometime in the past, she walked into the room and have seen a man sitting at the table, (2) the intent of woman is to talk to the man, (3) sometime in the future, she will sit down at the table and speak with him about a serious topic. While reasoning about the rich dynamic story of the visual scene is easy for humans, it is difficult for machines since it requires higher-order cognitive-level understanding of the world.

In recent years, several visual reasoning tasks [2–4] were proposed and drew attention in computer vision and natural language processing communities. To elaborate a few, the visual question answering (VQA) task defines a question-answering paradigm as a test to measure a machine's reasoning abilities for a given image or video. Visual dialog (VisDial) task asks a series of questions in the form of dialogue grounded on image or video. Visual commonsense reasoning (VCR) task further requires the machine to provide a rationale explaining why its answer is correct. While above visual reasoning tasks are defined

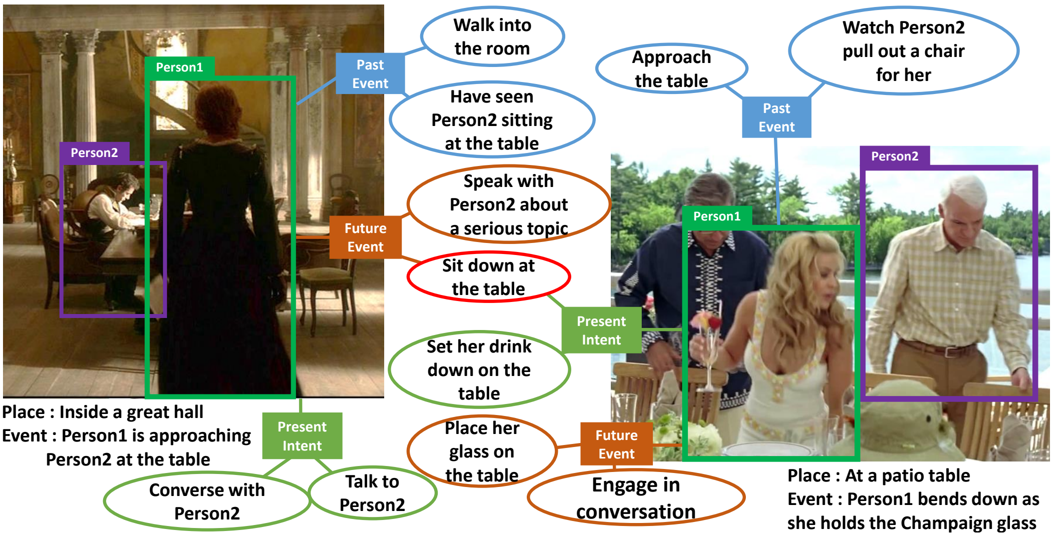
Person1
Person2
Past Event
Walk into the room
Have seen Person2 sitting at the table
Future Event
Speak with Person2 about a serious topic
Sit down at the table
Present Intent
Set her drink down on the table
Place her glass on the table
Place : Inside a great hall
Event : Person1 is approaching Person2 at the table
Present Intent
Converse with Person2
Talk to Person2
Approach the table
Watch Person2 pull out a chair for her
Past Event
Person2
Person1
Future Event
Engage in conversation
Place : At a patio table
Event : Person1 bends down as she holds the Champaign glass

Figure 1. Illustration on Visual Commonsense Generation. Given a person in image and corresponding textual event, agent is required to generate (1) what needed to happen *before*, (2) what is the current *intent* of person, and (3) what will happen *after*.

at recognition-level understanding and only consider the concepts and relations *within* the provided image or video, VCG focuses on reasoning about the rich cognitive-level dynamic story that goes beyond the directly visible contents by requiring the cause-and-effect caption generation. Piaget's cognitive development theory [5] describes the strive of human intelligence to know in two forms: as a form of states or as a form of transformations, suggesting that people must possess functions to represent both static and transformational aspects of realities. If former tasks represent reasoning in a stationary situation, then VCG represents reasoning in a transforming situation. Hence, the research on VCG opens the door for a major leap from recognition-level understanding to cognition-level reasoning.

Only a few works on VCG have been published. Park *et al.* [1] constructed benchmark for visual commonsense generation, VisualCOMET, and proposed the baseline method. Xing *et al.* [6] proposed knowledge enhanced multimodal BART (KM-BART) that leverages the knowledge from external corpora to pre-train BART. Previous approaches only operate on conventional learning scheme of visual and textual information, overlooking the distinctiveness of cause-and-effect generation task. Two major limitations on previous approaches are: (1) directly utilize conventional vision-language transformers to learn relationships between input modalities, and (2) ignore relations among target captions, but consider each caption independently. Due to the former limitation, previous approaches ignore the intra- and inter-modality relationship that have proven to be beneficial to transformer-based generation [7]. Due to the latter limitation, the exiting models pay no attention to the intrinsic structure of the task or dataset. As the goal of the VCG is to generate the cause-and-effect captions, it is essential to consider causal relations among each inference for *before*, *intent* and *after*. While existing approaches consider these inferences as a separate case and train independently, we argue that the generation of three captions should be considered holistically.

In this paper, we address the aforementioned limitations with our novel cause-and-effect Bart (CE-BART) which is composed of (1) structured graph reasoner (SGR) and (2) cause-and-effect generator (CEG). SGR first builds semantic graphs for each modality to interpret the intra-modality relationships from spatial or token domain via graph structures, then it captures higher-order semantic relations among graphs (i.e., inter-modality relationships) via tripartite graph attention and strengthens the multi-modal graph representations. As SGR comprehends the intra- and inter-relationships interspersed in multi-modal representations beforehand, the latter workload of the transformer-based CEG is unburdened, allowing it to focus more on understanding the commonsense and cause-and-effect inference of the given input. CEG generate cause-and-effect captions for *before*, *intent* and *after* situations. While all existing approaches on visual commonsense generation are trained to generate each cause-and-effect captions separately (i.e., there are no connection between the generation of *before*, *intent*, *after* even for the same image), the proposed CEG infers all three cause-and-effect captions holistically by considering the causal relations. It consists of one transformer encoder for modeling multi-modal representations, and three transformer decoders each for the generating *before*, *intent*, and *after* captions. To consider causal relations among cause-and-effect inferences, decoders for *intent* and *after* are connected to that of *before* and *intent*, respectively. Through causal connections between three decoders, it can attend to hidden states of former decoder which take role of *cause* to generate *effect* captions (i.e., the proposed intent/after decoder can attend to not only the hidden states of transformer encoder, but also the hidden states of before/intent decoder).

2. Related Works

2.1. Commonsense Reasoning

Commonsense knowledge has been attracted lots of attention in both computer vision and natural language communities. Commonsense or causality knowledge refers to the basic level of practical knowledge and reasoning about everyday situations and events commonly shared among most people [6,8]. For example, if the sun is out, it's not likely to rain; if we drop a cup, it is likely to broke. Such causality knowledge has been shown to be beneficial for many tasks [9,10], thus it is essential for machines to learn to understand causality [11].

In the field of natural language processing, several commonsense knowledge base (KB) have been constructed to help machines better understand the causality commonsense. ConceptNet [12] and ATOMIC [13] are widely used commoonsense KBs that leverages human-annotations to provide high quality causality knowledge. These KBs are built based on tuples (s, r, o) where s, o are subject, object phrases, and r defines the relation between them. Relations in commonsense KB includes *causes*, *because*, *before*, *as a result*, *etc* which is essential for learning causality. Bosselut *et al.* [14] proposed COMET, a transformer-based architecture, for automatic commonsense knowledge base completion. COMET is trained to predict the object o, given subject s and relation r. In the field of computer vision, visual commonsense reasoning (VCR) task [4] has been proposed which is a visual question answering benchmark that requires the machine to provide a rationale explaining why its answer is correct.

2.2. Visual Commonsense Generation

Park *et al.* [1] proposes the task of visual commonsense generation and corresponding benchmark, VisaulCOMET, which aims at generating cause-and-effect descriptions for a given image and corresponding textual event and place. VisualCOMET is a visual commonsense knowledge base where image and corresponding textual event and place take place of the object in ATOMIC. There exists only a few works [1,6] dealing with the task of the visual commonsense generation. Park *et al.* [1] first proposed a baseline model based on GPT-2 [15]. The baseline model feeds visual and textual context as inputs and is trained to predict each of the cause-and-effect descriptions. Xing *et al.* [6] propose knowledge enhanced multimodal BART (KM-BART) utilizes BART [16] to pretrain on large external datasets and leverages knowledge from them. KM-BART is first pre-trained with knowledge-based commonsense generation by leveraging knowledge from COMET [14], attribute & relation prediction using Visual Genome benchmark [17], and masked language & region modeling using various pre-training benchmarks. It then is fine-tuned on VisualCOMET benchmark to achieve state-of-the-art performance on VisualCOMET benchmark. However, we argue that these systems operate on conventional learning scheme of visual and textual information, overlooking the distinctiveness of cause-and-effect generation task and possesses two major limitations: (1) conventional vision-language

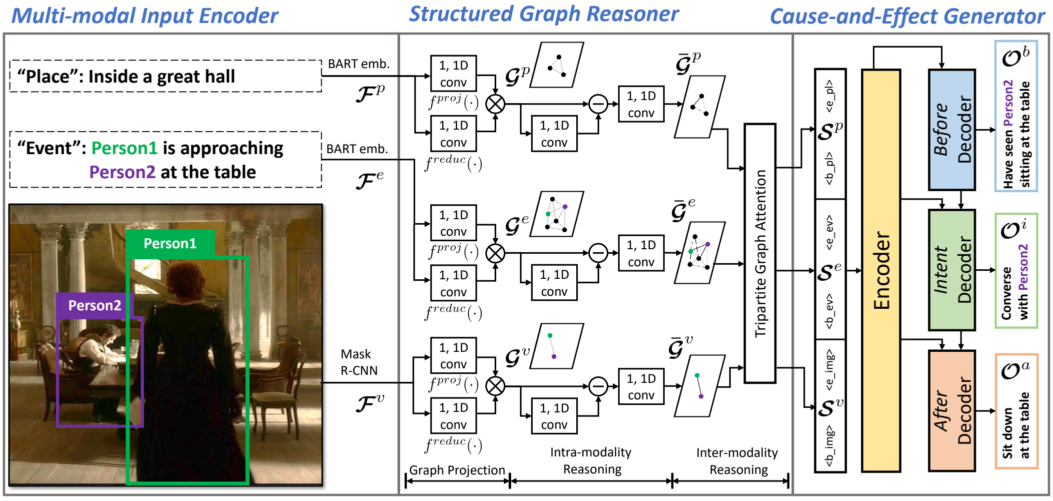

Figure 2. Illustration of Cause-and-Effect BART (CE-BART) which is composed of Multi-modal Input Encoder, Structured Graph Reasoner, and Cause-and-Effect Generator. (1) Multi-modal Input Encoder: we first obtain multi-modal features ($\mathcal{F}^v, \mathcal{F}^e, \mathcal{F}^p$) using pre-trained models; Mask R-CNN for visual, and BART word embedding layer for textual inputs, (2) Structured Graph Reasoner (SGR): we then build semantic graphs ($\mathcal{G}^v, \mathcal{G}^e, \mathcal{G}^p$) and strengthens their representations ($\mathcal{S}^v, \mathcal{S}^e, \mathcal{S}^p$) via capturing intra- and inter-modality relations, (3) Cause-and-Effect Generator (CEG): we finally generate cause-and-effect descriptions ($\mathcal{O}^b, \mathcal{O}^i, \mathcal{O}^a$) using BART-based transformer architecture which considers causal relations among inferences.

transformers are directly utilized to learn relationships between input modalities, (2) every training examples are trained independently without considering relations with others.

3. Cause-and-Effect BART

First, we provide a formal definition of the visual commonsense generation task [1] as follows. We are given tuples of (v, e, p), consisting of an image v, the event description e, and place description p. The goal of visual commonsense generation is to generate the three cause-and-effect captions corresponding to (1) what needed to happen *before*, (2) what is the current *intent* of the person, and (3) what will happen *after*.

3.1. Multi-modal Input Encoder

Following the previous work on visual commonsense generation, we use the Mask R-CNN [18] to detect the visual person, which extracts $\mathcal{N}^v$ number of appearance features $\mathcal{A} = \{a_i\}_{i=1}^{\mathcal{N}^v}$, and their corresponding location features $\mathcal{B} = \{b_i\}_{i=1}^{\mathcal{N}^v}$. Each location feature $b_i = [x_i, y_i, w_i, h_i]$ represents a spatial coordinate, where $[x_i, y_i]$ denotes the relative coordinate of top-left point the the bounding box while $[w_i, h_i]$ denotes the width and height of the box. We calculate the final visual feature as: $\mathcal{F}^v = \{v_i\}_{i=1}^{\mathcal{N}^v} \in \mathbb{R}^{\mathcal{N}^v \times d_v}$, where $v_i = w^a a_i + w^b b_i$, and w^a, w^b are learnable weights that embeds both features into visual feature dimension d_v.

We have two types of text for each image (i.e., event e and place p). Each sentence for event and place is fed into the word embedding layer of pre-trained BART to be further utilized. We obtain the textual feature as: $\mathcal{F}^e = \{e_i\}_{i=1}^{\mathcal{N}^e}$, $\mathcal{F}^p = \{p_i\}_{i=1}^{\mathcal{N}^p}$, where $\mathcal{N}^v, \mathcal{N}^p$ are the number of token features, and $e_i, p_i \in \mathbb{R}^{d_t}$ are the embedding of the i-th token in the event and place, respectively.

3.2. Structured Graph Reasoner

In order to capture the intra-modality relationships from individual modalities (i.e., image, event, and place) and inter-modality relationships among input modalities, structured graph reasoner first builds semantic graphs for each modality; image semantic graph $\mathcal{G}^v$, event semantic graph $\mathcal{G}^e$, and place semantic graph $\mathcal{G}^p$. Motivated by [19] that projects visual features in spatial domain into graph domain for relational reasoning over global context, structured graph reasoner performs graph convolutions to capture intra-modality relations. It then captures the higher-order semantic relations among graphs (i.e., inter-modality relationships) via tripartite graph attention to strengthens the multi-modal graph representations. The final strengthened semantic representations ($\boldsymbol{\mathcal{S}}^v, \boldsymbol{\mathcal{S}}^e, \boldsymbol{\mathcal{S}}^p$) are fed into the following cause-and-effect generator.

For simplicity, we denote the feature representation as $\boldsymbol{\mathcal{F}}^x$ and semantic graph as $\boldsymbol{\mathcal{G}}^x$ for each modality $x \in \{v, e, p\}$. We first project the feature representation $\boldsymbol{\mathcal{F}}^x$ into semantic graph $\boldsymbol{\mathcal{G}}^x$ which is a lightweight *fully-connected* graph. Basically, the projection into graph domain is formulated as a linear combination (i.e., weighted global pooling):

$$\boldsymbol{\mathcal{G}}^x = f^{proj}(\boldsymbol{\mathcal{F}}^x; W_{f_{proj}}) times f^{reduc}(\boldsymbol{\mathcal{F}}^x; W_{f_{reduc}}) \in \mathbb{R}^{\mathcal{N}^x \times d_g}, \tag{1}$$

where the dimension reduction function f^{reduc} parameterized by $W_{f_{reduc}}$ projects each feature into graph feature dimension d_g and graph projection function f^{proj} parameterized by $W_{f_{proj}}$ produces the weights for linear combination. Here, both function f^{reduc}, f^{proj} are 1D convolution layers with a kernel size of 1.

In order to capture intra-modality relations in individual semantic graph, we utilize graph convolution [20] to update node representations and obtain $\bar{\boldsymbol{\mathcal{G}}}^x$. Given a fully connected graph $\boldsymbol{\mathcal{G}}^x$, graph convolution learns edge weights that correspond to the correlations between node representations. A single layer of graph convolution is formulated as:

$$\bar{\boldsymbol{\mathcal{G}}}^x = \boldsymbol{\Lambda} \boldsymbol{\mathcal{G}}^x W_x = ((\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{A}}^x)\boldsymbol{\mathcal{G}}^x) W_x, \tag{2}$$

where $\boldsymbol{\Lambda}$ and $\boldsymbol{\mathcal{A}}^x$ are $\mathcal{N}^x \times \mathcal{N}^x$ adjacency matrix for diffusing information across nodes of $\boldsymbol{\mathcal{G}}^x$, $W_x \in \mathbb{R}^{d_g \times d_g}$ denotes the state update weight, and $\boldsymbol{\mathcal{I}} \in \mathbb{R}^{\mathcal{N}^x \times \mathcal{N}^x}$ is the identity matrix. Here, adjacency matrix $\boldsymbol{\mathcal{A}}^x$ is randomly initialized and learned during training, together with W_x, and the identity matrix serves as a shortcut connection. We can implement Equation 2 using two consecutive 1D convolution layers along different directions: channel-wise convolution (i.e., modeling $(\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{A}}^x)$) and node-wise convolution (i.e., modeling W_x).

Finally, we capture inter-modality relations among three semantic graphs ($\bar{\boldsymbol{\mathcal{G}}}^v, \bar{\boldsymbol{\mathcal{G}}}^e, \bar{\boldsymbol{\mathcal{G}}}^p$) via tripartite graph attention and calculates the strengthened semantic representations ($\boldsymbol{\mathcal{S}}^v, \boldsymbol{\mathcal{S}}^e, \boldsymbol{\mathcal{S}}^p$). We perform graph attention over *tripartite* graph that connects all of the nodes in individual modalities to all of the nodes belonging to the other modalities. By doing so, every node in each modality learns to integrate informative semantics from the other modality to its representation effectively to capture inter-modality relations. First, we concatenate $\bar{\boldsymbol{\mathcal{G}}}^v, \bar{\boldsymbol{\mathcal{G}}}^e, \bar{\boldsymbol{\mathcal{G}}}^p$ along node-axis to make a tripartite graph structure $\bar{\boldsymbol{\mathcal{G}}}^T$, which each node is connected to all the other nodes that belonged to different modality: $\bar{\boldsymbol{\mathcal{G}}}^T = [\bar{\boldsymbol{\mathcal{G}}}^v || \bar{\boldsymbol{\mathcal{G}}}^e || \bar{\boldsymbol{\mathcal{G}}}^p] \in \mathbb{R}^{\mathcal{N}^T \times d_g}$. We perform graph attention [21] over $\bar{\boldsymbol{\mathcal{G}}}^T$ that calculates the multi-head attention to capture relations between each node and its neighboring nodes (i.e., inter-modality relations):

$$\boldsymbol{\mathcal{S}}^T = \mathrm{GAT}(\bar{\boldsymbol{\mathcal{G}}}^T), \tag{3}$$

$$\boldsymbol{\mathcal{S}}^v, \boldsymbol{\mathcal{S}}^e, \boldsymbol{\mathcal{S}}^p = slice(\boldsymbol{\mathcal{S}}^T), \tag{4}$$

where $slice(\cdot)$ operation slices the multi-modal representations along node-axis with corresponding length of each modality.

3.3. Cause-and-Effect Generator

A cause-and-effect generator (CEG) is proposed to generate cause-and-effect captions by considering the causal relationships among inferences. It is a sequence-to-sequence transformer architecture that feeds the strengthened semantic graph ($\boldsymbol{\mathcal{S}}^v, \boldsymbol{\mathcal{S}}^e, \boldsymbol{\mathcal{S}}^p$) and decodes cause-and-effect captions (*before* $\boldsymbol{\mathcal{O}}^b$, *intent* $\boldsymbol{\mathcal{O}}^i$ and *after* $\boldsymbol{\mathcal{O}}^a$) in autoregressive manner. Different from existing approaches that treat the generation of each caption as separate objectives, the cause-and-effect generator infers all three cause-and-effect captions holistically. Formally, we have the function of CEG f_{CEG} with its parameter $\mathcal{W}_{f_{CEG}}$ whose goal is:

$$\mathcal{O}^b, \mathcal{O}^i, \mathcal{O}^a = f_{CEG}(\mathcal{S}^v, \mathcal{S}^e, \mathcal{S}^p; \mathcal{W}_{f_{CEG}}). \tag{5}$$

3.3.1. Encoder

The encoder of CEG is based on a multi-layer bidirectional Transformer as in the BART and its variant in the visual commonsense generation, KM-BART. Different from the encoder in KM-BART whose input sequence starts with one of the special tokens `<before>`, `<intent>`, `<after>` to indicate the model which cause-and-effect caption should be generated, CEG only takes the three sets of semantic graph representations (i.e., $\mathcal{S}^v, \mathcal{S}^e, \mathcal{S}^p$) as it infers all three captions holistically. To inform the start and end of different input modalities to the encoder, we add three sets of special tokens; `<b_img>`, `<e_img>` for image embedding $\mathcal{S}^v$, `<b_ev>`, `<e_ev>` for event embedding $\mathcal{S}^e$ and `<b_pl>`, `<e_pl>` for place embedding $\mathcal{S}^p$.

3.3.2. Decoder

The decoders of CEG are based on multi-layer unidirectional Transformer as it works in an autoregressive manner during generation. There are total three of decoders for CEG, each for the generation of *before, intent* and *after* captions. To inform each decoder about the start of generation, we add three special starting tokens for each decoder `<before>`, `<intent>` and `<after>`. Further, we add a special end inference token `<e_inf>` at the end of the target sequence to indicating the stop of a decoding process. During training, we use teacher-forcing [22] to supervise each decoding steps, i.e., ground truth tokens are used as decoder input. The decoders of CEG only take a right-shifted target token sequence as input.

To consider causal relations among cause-and-effect captions, decoders for *intent* and *after* are connected to that of *before* and *intent*, respectively. Through causal connections between three decoders, it can attend to hidden states of former decoder which take role of *cause* to generate *effect* captions (i.e., the proposed *intent/after* decoder can attend to not only the hidden states of transformer encoder, but also the hidden states of *before/intent* decoder), as shown in Figure 2. Formally, we divide the function of CEG f_{CEG} in Equation 5 as encoder E_{CEG} and a set of decoders D^x_{CEG} where $x \in \{b, i, a\}$. The conventional approaches [1,6] generate the cause-and-effect captions separately without considering causal relations:

$$\mathcal{O}^b = D_{con}(E_{con}(v, e, p)), \tag{6}$$
$$\mathcal{O}^i = D_{con}(E_{con}(v, e, p)), \tag{7}$$
$$\mathcal{O}^a = D_{con}(E_{con}(v, e, p)), \tag{8}$$

where E_{con} and D_{con} represents the function of encoder and decoder of existing approaches which feeds the image v, event e, and place p. On the other hand, the proposed CEG has sequential connections among decoders to keep causal relations among cause-and-effect captions:

$$\mathcal{O}^b = D^b_{CEG}(E_{CEG}(\mathcal{S}^v, \mathcal{S}^e, \mathcal{S}^p)), \tag{9}$$
$$\mathcal{O}^i = D^i_{CEG}(E_{CEG}(\cdot), D^b_{CEG}(\cdot)), \tag{10}$$
$$\mathcal{O}^a = D^a_{CEG}(E_{CEG}(\cdot), D^i_{CEG}(\cdot)). \tag{11}$$

4. Experiments

4.1. Benchmark Dataset

VisualCOMET [1] is a large-scale benchmark dataset for visual commonsense generation, which is the only available dataset of its kind at present. It consists of over 1.4 million

doi:10.20944/preprints202211.0134.v1

Table 1. Comparison with State-of-the-art methods in VisualCOMET benchmark. Here, "Proj-SGR" denotes the graph projection (i.e., Equation 1), "Intra-SGR" denotes the intra-modality reasoning (i.e., Equation 2), and "Iner-SGR" stands for the inter-modality reasoning (i.e., Equation 3). "CEG" stands for the cause-and-effect generator with three decoder (i.e., Equations 9-11).

Methods	Validation set			Test set		
	BLEU2	METEOR	CIDEr	BLEU2	METEOR	CIDEr
Baseline [1]	13.50	11.55	18.27	12.71	11.13	17.36
KM-BART [6]	23.47	15.02	39.76	-	-	-
Variants on CE-BART						
BART-base	22.51	14.73	37.86	-	-	-
+ Proj-SGR	22.47	14.97	38.91	-	-	-
+ Intra-SGR	23.85	15.72	39.59	-	-	-
+ Inter-SGR	25.07	18.24	41.07	-	-	-
+ CEG	28.60	19.32	43.58	-	-	-
CE-BART	**28.60**	**19.32**	**43.58**	**28.14**	**18.91**	**42.64**

Table 2. Analysis conducted on validation split of VisualCOMET. We provide an analysis of the behavior of CEG by observing separate performance evaluation for with and wihtout CEG.

Methods	Before			Intent			After		
	B2	M	C	B2	M	C	B2	M	C
w/o CEG	29.7	20.4	45.1	19.4	15.4	40.7	26.1	18.9	37.7
w/ CEG	30.9	20.9	45.9	25.5	16.6	42.4	29.6	20.2	41.9

textual captions of visual commonsense inferences carefully annotated over a diverse set of 59,000 images paired with 139,000 event descriptions. Visual commonsense inferences are divided into 1,174K, 146K, 145K examples for training, validation, and test, respectively.

Audio Visual Scene-aware Dialogue (AVSD) [23] provides video, caption, and dialogue history consisting of a series of textual QA pairs, and follow-up question about the video. The goal is to generate a free-form natural language answer to the question. As both tasks share similar input-output relations, CE-BART can be easily applied to video-grounded dialogue, and transfer causal knowledge learned from a visual commonsense generation for better video understanding.

4.2. Experimental Details

We initialize the cause-and-effect generator with a pre-trained BART-base model with 6 transformer layers in each of the encoder and decoder, and a hidden size of 768. For tripartite graph attention in structured graph reasoner, the number of heads in multi-head attention is set to 8. We trained using 4 NVIDIA Quadro RTX 8000 (48GB of memory) and Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initially set to 0.0001 and trained the model up to 30 epochs with an effective training batch size of 512. During inference, we adopt a beam search and for each set of input, we decode *before*, *intent* and *after* captions sequentially.

4.3. Experimental Results on VisualCOMET

We compare our proposed Cause-and-Effect BART (CE-BART) with state-of-the-art methods on VisualCOMET benchmark. Table 1 summarizes the experimental results on VisualCOMET benchmark on both validation and test split, since the current state-of-the-art method, KM-BART [6], only provides the results on validataion split. Also, we provide the results on ablation study in Table 1 with several variants of CE-BART in order to measure the effectiveness of the proposed key components of CE-BART. All the reported

Table 3. Comparison with State-of-the-art methods in AVSD benchmark. We compare CE-BART with various state-of-the-art systems on AVSD benchmark; Baseline [24], STSGR [25], MTN[26], MTN-TMT [27], VX2TEXT [28].

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
Baseline [24]	-	-	-	0.078	0.113	0.277	0.727
STSGR [25]	-	-	-	0.133	0.165	0.362	1.272
MTN [26]	0.356	0.242	0.174	0.135	0.165	0.365	1.366
MTN-TMT [27]	-	-	-	0.142	0.171	0.371	1.357
VX2TEXT [28]	0.361	0.260	0.197	0.154	0.178	0.393	1.605
CE-BART	0.364	0.266	0.203	0.158	0.181	0.400	1.681
CE-BART w/ pre-train	**0.365**	**0.268**	**0.205**	**0.161**	**0.183**	**0.404**	**1.721**

performances in Table 1 are the average value of 5 independently trained models with different seed.

Starting from direct fine-tuning of BART-base model in VisualCOMET which shows slightly lower performance compared to previous SOTA method, KM-BART, every component of proposed CE-BART boosts performance on all three metrics. The results of the ablation study suggest that the limitations of existing approaches that we have introduced are valid; (1) conventional vision-language transformers are directly utilized to learn relationships between input modalities, (2) every training examples are trained independently without considering relations with others.

Structured graph reasoner is proposed to capture intra- and inter-modality relations among visual and textual representations. Inclusion of graph reasoning shows 3.21 point boost in CIDEr metric compared to BART-base model. Among the components of structured graph reasoner, intra-modality reasoning provides 0.68 point and inter-modality reasoning provides 1.48 point gain in CIDEr. As our inter-modality reasoning module performs multi-head attention over tripartite graph, whose neighborhood is defined as the nodes of heterogeneous modality, each node reinforces its representation with the information from other modalities, thus it is able to comprehend inter-modality relations effectively. Our design of structured graph reasoning is effective in capturing intra- and inter-modality relations which is essential in visual commonsense generation.

Cause-and-effect generator is proposed to generate cause-and-effect descriptions holistically by considering the causal relationships among inferences. It improves CIDEr score by 2.51 points. Through causal connections between three decoders, cause-and-effect generator looks at the former decoder which take role of cause to generate effect description. Our design of cause-and-effect generator is effective in modeling causal relations among generated descriptions which is essential in visual commonsense generation.

Through ablation study, we suggest that proposed CE-BART can effectively capture intra- and inter-modality relationships interspersed in multi-modal input representations, and effectively generate cause-and-effect descriptions holistically by considering causal relations through cause-and-effect generator with causal connections between decoders. CE-BART surpasses the other state-of-the-art methods on both validation and test split of VisualCOMET benchmark. Compared to KM-BART [6], which is state-of-the-art method in validation split, CE-BART reaches a CIDEr score of 43.58, which improves almost 4 points. Compared to baseline [1], which is state-of-the-art method in test split, CE-BART achieves a CIDEr score of 42.64, which is more than double compared to 17.36. In the meantime, CE-BART also managed to improve BLEU-2 score by almost 16 points and METEOR score by more than 7 points.

We also provide an in-depth quantitative analysis of CEG in Table 2 to show the significance of dependencies among three decoders. The motivation behind capturing relations between training samples is to consider the relations among *before*, *intent*, *after* descriptions while generating them for an image: current *intent* is related to the situation *before* and situation *after* is related to current *intent*. As CEG has connections among

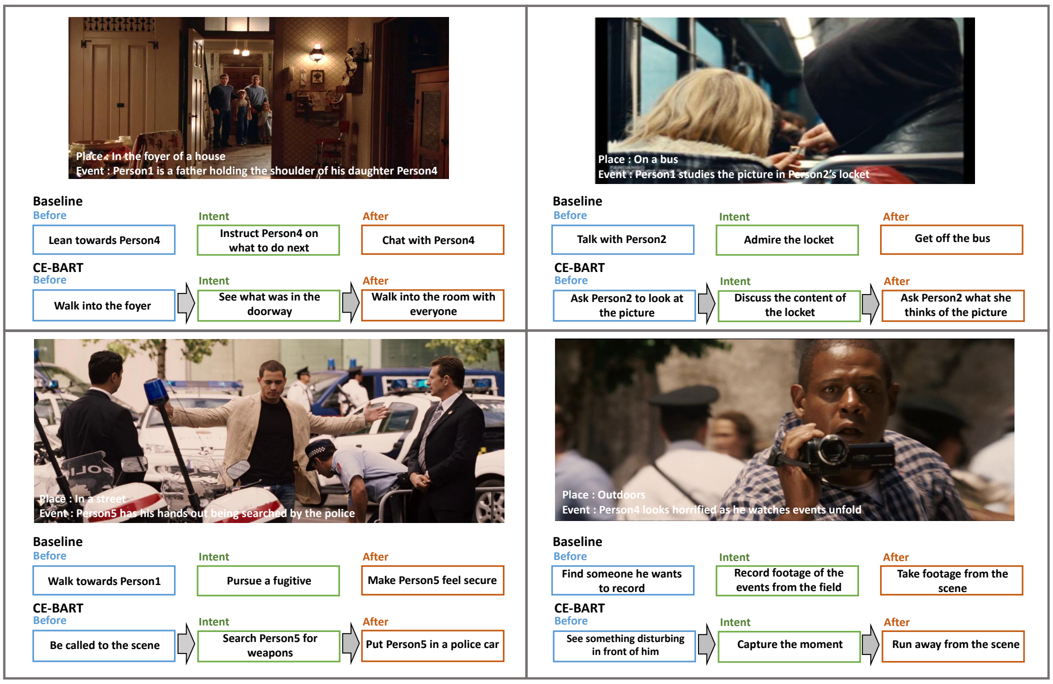
Place : In the foyer of a house
Event : Person1 is a father holding the shoulder of his daughter Person4
Baseline
Before
Lean towards Person4
Intent
Instruct Person4 on what to do next
After
Chat with Person4
CE-BART
Before
Walk into the foyer
Intent
See what was in the doorway
After
Walk into the room with everyone
Place : On a bus
Event : Person1 studies the picture in Person2's locket
Baseline
Before
Talk with Person2
Intent
Admire the locket
After
Get off the bus
CE-BART
Before
Ask Person2 to look at the picture
Intent
Discuss the content of the locket
After
Ask Person2 what she thinks of the picture
Place : In a street
Event : Person5 has his hands out being searched by the police
Baseline
Before
Walk towards Person1
Intent
Pursue a fugitive
After
Make Person5 feel secure
CE-BART
Before
Be called to the scene
Intent
Search Person5 for weapons
After
Put Person5 in a police car
Place : Outdoors
Event : Person4 looks horrified as he watches events unfold
Baseline
Before
Find someone he wants to record
Intent
Record footage of the events from the field
After
Take footage from the scene
CE-BART
Before
See something disturbing in front of him
Intent
Capture the moment
After
Run away from the scene

Figure 3. Four examples from the test split of VisualCOMET benchmark.

decoders, *intent* decoder can operate by using not only the image but also the information from *before* decoder. Similarly, *after* decoder can generate predictions with the help of *intent* decoder. We conduct separate evaluations for *before*/*intent*/*after* captions. Without connections among decoders (i.e., w/o CEG), *before* prediction shows superior performance compared to *intent* and *after* predictions. But with connections among decoders (i.e., w/ CEG), we can observe performance boost in *intent* and *after* predictions as expected via considering relations.

4.4. Experimental Results on AVSD

We conducted additional experiments to validate the generalizability of the proposed CE-BART in other VL tasks. We have conducted experiments in the task of video-grounded dialogue, which is a multi-turn question answering task. Formal definition of video-grounded dialogue is as follows: we are given a video, a dialogue history consisting of a series of textual QA pairs, and a follow-up question about the video, and goal is to generate a free-form natural language answer. As we can see, both tasks share similar input-output relations; therefore, CE-BART can be easily applied to the new task, video-grounded dialogue, and transfer causal knowledge learned from a visual commonsense generation for better video understanding. We trained CE-BART for video-grounded dialogue in two settings: (1) CE-BART without pre-training in VisualCOMET, (2) CE-BART with pre-training in VisualCOMET. Through this comparative experiment, we are able to observe that causal information learned from VisualCOMET can help with understanding the video. Table 3 summarizes the results on AVSD benchmark. It is observed that CE-BART improves over existing methods and achieves SOTA performance on all of the metrics. By pre-training CE-BART on VisualCOMET, we can further boost performance which indicates that causal knowledge trained from VisualCOMET can be successfully transferred for better video understanding in AVSD. We could validate that our proposed CE-BART can benefit other VL tasks by effectively transferring causal knowledge learned from VisualCOMET.

4.5. Qualitative Analysis

Figure 3 visualizes examples from test split of VisualCOMET and compare predictions of CE-BART and Baseline. CE-BART successfully utilizes SGR that captures intra- and inter-modality relationship, which is flexible in selecting important information regardless of rather it is textual or visual information. In first example, CE-BART notices that text of father holding his daughter's shoulder is not the crucial information and focuses more on the image. With this ability to consider and select crucial information, in second example, CE-BART focuses on the key object in the scene and people's interaction centered around that object and avoid stating the obvious. Further, as CEG is trained successively generate captions through causal connections, it has the information from the former decoder which take the role of *cause* to generate *effect* caption. In third example, CE-BART is continuous regarding the contents, where baseline model produces discontinuous caption. In the lower-right example, the baseline model produces the overlapping captions, while CE-BART can effectively generate the rich dynamic story of the visual scene.

5. Limitations

We believe that our proposed CE-BART contains several limitations that can be removed through further experiments in the future. First, its scalability is limited due to requirement of large GPU resources. We have conducted experiments using 4 NVIDIA Quadro RTX 8000 (48GB of memory) which are extremely expensive. Second, its scalability to control time scale is limited. There is no factor in current task setting that selects how much of a past / future situation it requires. We will further develop our methods to overcome several limitations.

6. Conclusion

We proposed a novel Cause-and-Effect BART for the task of visual commonsense generation. The proposed CE-BART consists of two major components: (1) Structured Graph Reasoner, and (2) Cause-and-Effect Generator. Structured graph reasoner builds semantic graphs for individual modalities and strengthens their representations via capturing intra- and inter-modality relations among graph structures. Cause-and-effect generator is a transformer architecture with three decoder, each for generating before, intent, and after captions. The experimental results on VisualCOMET and AVSD benchmark shows that CE-BART achieves new state-of-the-art performance.

Author Contributions: Conceptualization, J. K.; methodology, J. K.; software, J. K., J. H. and S. Y.; validation, J. K. and S. Y.; formal analysis, J. K. and C. Y.; writing, J. K., J. H., S. Y. and C. Y.; supervision, C. Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Chung-Ang University Research Grants in 2022, and supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant through the Ministry of Science and ICT (MSIT), Government of Korea (Development of Previsional Intelligence Based on Long-Term Visual Memory Network) under Grant 2020-0-00004.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Not applicable

Acknowledgments: The authors would like to thank to Jinwoo Park, Jangyoung Jeon, Jungeun Kim and Zhang Nan for providing help on proofreading this paper and researchers at ETRI for providing constructive comments on this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VCG	Visual Commonsense Generation
VQA	Visual Question Answerig
VCR	Visual Commonsense Reasoning
CE-BART	Cuase-and-Effect BART
SGR	Structured Graph Reasoner
CEG	Cause-and-Effect Generator

References

1. Park, J.S.; Bhagavatula, C.; Mottaghi, R.; Farhadi, A.; Choi, Y. VisualCOMET: Reasoning about the Dynamic Context of a Still Image. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2020.
2. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; Parikh, D. Vqa: Visual question answering. In Proceedings of the The IEEE International Conference on Computer Vision (ICCV), 2015.
3. Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J.M.; Parikh, D.; Batra, D. Visual Dialog. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
4. Zellers, R.; Bisk, Y.; Farhadi, A.; Choi, Y. From Recognition to Cognition: Visual Commonsense Reasoning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
5. Piaget, J. The role of action in the development of thinking. In *Knowledge and development*; Springer, 1977; pp. 17–42.
6. Xing, Y.; Shi, Z.; Meng, Z.; Ma, Y.; Wattenhofer, R. KM-BART: Knowledge Enhanced Multimodal BART for Visual Commonsense Generation. In Proceedings of the arXiv preprint arXiv:2101.00419, 2021.
7. Kim, J.; Yoon, S.; Kim, D.; Yoo, C.D. Structured Co-reference Graph Attention for Video-grounded Dialogue. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2021.
8. Sap, M.; Shwartz, V.; Bosselut, A.; Choi, Y.; Roth, D. Commonsense Reasoning for Natural Language Processing. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computatioonal Linguistics (ACL), Tutorial, 2020.
9. Hashimoto, C.; Torisawa, K.; Kloetzer, J.; Sano, M.; Varga, I.; Oh, J.H.; Kidawara, Y. Toward Future Scenario Generation: Extracting Event Causality Exploiting Semantic Relation, Context, and Association Features. In Proceedings of the Proceedings of the 52th Annual Meeting of the Association for Computatioonal Linguistics (ACL), 2014.
10. Ning, Q.; Feng, Z.; Wu, H.; Roth, D. Joint Reasoning for Temporal and Causal Relations. In Proceedings of the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2018.
11. Pearl, J.; Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*, 1st ed.; Basic Books, Inc.: USA, 2018.
12. Li, X.; Taheri, A.; Tu, L.; Gimpel, K. Commonsense Knowledge Base Completion. In Proceedings of the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), 2016.
13. Sap, M.; Bras, R.L.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N.A.; Choi, Y. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computatioonal Linguistics (ACL), 2019.
14. Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Çelikyilmaz, A.; Choi, Y. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019.
15. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
16. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020.
17. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. 2016.
18. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
19. Chen, Y.; Rohrbach, M.; Yan, Z.; Shuicheng, Y.; Feng, J.; Kalantidis, Y. Graph-Based Global Reasoning Networks. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
20. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations (ICLR), 2017.
21. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903* **2017**.
22. Lamb, A.M.; ALIAS PARTH GOYAL, A.G.; Zhang, Y.; Zhang, S.; Courville, A.C.; Bengio, Y. Professor Forcing: A New Algorithm for Training Recurrent Networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2016.
23. Alamri, H.; Cartillier, V.; Das, A.; Wang, J.; Cherian, A.; Essa, I.; Batra, D.; Marks, T.K.; Hori, C.; Anderson, P.; et al. Audio Visual Scene-Aware Dialog. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

24. Hori, C.; Alamri, H.; Wang, J.; Wichern, G.; Hori, T.; Cherian, A.; Marks, T.K.; Cartillier, V.; Lopes, R.G.; Das, A.; et al. End-to-end Audio Visual Scene-aware Dialog Using Multimodal Attention-based Video Features. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.
25. Geng, S.; Gao, P.; Chatterjee, M.; Hori, C.; Roux, J.; Zhang, Y.; Li, H.; Cherian, A. Dynamic Graph Representation Learning for Video Dialog via Multi-modal Shuffled Transformers. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2021.
26. Le, H.; Sahoo, D.; Chen, N.; Hoi, S.C. Multimodal Transformer Networks for End-to-End Video-Grounded Dialogue Systems. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019.
27. Li, W.; Jiang, D.; Zou, W.; Li, X. TMT: A Transformer-based Modal Translator for Improving Multimodal Sequence Representations in Audio Visual Scene-aware Dialog. In Proceedings of the Proceedings of the Interspeech, 2020.
28. Lin, X.; Bertasius, G.; Wang, J.; Chang, S.F.; Parikh, D.; Torresani, L. VX2TEXT: End-to-End Learning of Video-Based Text Generation From Multimodal Inputs. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.