Two Heads are Better Than One: Hypergraph-Enhanced Graph Reasoning for Visual Event Ratiocination

Wenbo Zheng\textsuperscript{1,2} Lan Yan\textsuperscript{2,3} Chao Gou\textsuperscript{4} Fei-Yue Wang\textsuperscript{2}

Abstract

Even with a still image, humans can ratiocinate various visual cause-and-effect descriptions before, at present, and after, as well as beyond the given image. However, it is challenging for models to achieve such task—the visual event ratiocination, owing to the limitations of time and space. To this end, we propose a novel multimodal model, Hypergraph-Enhanced Graph Reasoning. First it represents the contents from the same modality as a semantic graph and mines the intra-modality relationship, therefore breaking the limitations in the spatial domain. Then, we introduce the Graph Self-Attention Enhancement. On the one hand, this enables semantic graph representations from different modalities to enhance each other and captures the inter-modality relationship along the line. On the other hand, it utilizes our built multi-modal hypergraphs in different moments to boost individual semantic graph representations, and breaks the limitations in the temporal domain. Our method illustrates the case of “two heads are better than one” in the sense that semantic graph representations with the help of the proposed enhancement mechanism are more robust than those without. Finally, we re-project these representations and leverage their outcomes to generate textual cause-and-effect descriptions. Experimental results show that our model achieves significantly higher performance in comparison with other state-of-the-arts.

1. Introduction

![Image](a)

Figure 1. Illustration of The Visual Event Ratiocination. Given a person in the image, the model is required to reason about what needed to happen before, what intents of the people at present, and what will happen next. Our model has the excellent abilities of visual inference and visual narrating, detailed in Section 3.4–3.5.

Visual Event Ratiocination is a novel challenging task about a combination of language and vision. Given an image and a description of the event in the image, it requires models to predict events that happen before/after and the present intents of the characters in the given image. In VisualCOMET benchmark dataset (Park et al., 2020), we show an example in Figure 1(a) given the image of the woman looking over at the man with a concerned look in a classroom, the model can infer and generate/narrate three kinds of event ratiocination (a.k.a., event’s cause-and-effect...
descriptions): ① sometime in the past, she might unconsciously say many things to him, and when coming back to her senses, she might realize he looks panicked and become suspicious of him; ② sometime at present, she might ask he what is going on; ③ sometime in the future, she might realize he isn’t listening. Besides, the event’s cause-and-effect descriptions (e.g., ‘ask what is going on’) in the current moment of one image might be the same as these in the past time of one another image (shown in Figure 1(b)).

“A picture is worth a thousand words”. The combination of two modalities (i.e., language and vision) is commonplace, so it is only natural to ask to what extent this combination may help machines understand the meaning. Some conventional tasks have been introduced for joint understanding from two modalities, e.g., visual question answering (Antol et al., 2015), referring expression reasoning (Liu et al., 2019). Different from these tasks, which only focus on visual recognition and inference about the current content of images, visual event ratiocination aims at reasoning the time-varying situation captured in the image, and pays attention to the model’s abilities of both visual inference and visual narrating, as shown in Figure 1. Therefore, the study of this task has scientific significance: it opens the door of a significant leap from recognition-level understanding to cognitive-level reasoning.

From the above observation, it is obvious that there are three kinds of vital relationships here for generating event’s cause-and-effect descriptions: ① relationships within the same modality, ② relationships between different modalities, and ③ spatio-temporal relationships among different images and their event ratiocination. Therefore, how to capture these three relationships from visual and semantic perspectives is essential for the task of visual event ratiocination.

Nevertheless, there has been little work on visual event ratiocination, while conventional visual-language tasks have been explored to a large extent. Park et al. (2020) proposed a dataset, VisualCOMET, which is the only one benchmark for this task at present. They also employ the Transformer (Vaswani et al., 2017) as a baseline model. Xing et al. (2021) propose the improved BART (Lewis et al., 2020) that incorporates textual information into the multi-modal model. Yet, these methods only look at conventional learning of visual and textual information while ignoring the link between modalities and among space-time. In short, current visual event ratiocination approaches have two main deficiencies:

① The existing models pay no attention to relationships from the same modality and different modalities.
② The current models ignore the spatio-temporal relationships from different samples with different moments.

To address the above two challenges, we propose a novel model, Hypergraph-Enhanced Graph Reasoning, to obtain a representation of the multi-modal contents in the task of visual event ratiocination. As shown in Figure 2.① we construct semantic graphs for the same modality through translating intra-modality relationships from the spatial domain to the graph domain; furthermore, we propose an enhancement mechanism between these graphs to capture relationships between different modalities; ② we construct hypergraphs from different modalities with different moments to capture spatio-temporal relationships, and make these high-order semantic relations enhance the multi-modal graph representations by the proposed enhancement mechanism between graphs and the hypergraph.

In summary, our main contributions are as follows:

☆ We propose a novel and effective hypergraph-enhanced graph reasoning model for visual event ratiocination. Our model captures intra- and inter-modality relationships as well as spatio-temporal relationships via learning in the graph domain. Experimental results show that our model has strong robustness and outperforms existing similar methods.

☆ We explore how to enhance two semantic graphs with each other as well as semantic graphs with hypergraphs, and propose a novel graph self-attention enhancement. The qualitative experiment shows that this mechanism is effective.

☆ Our hypergraph-enhanced graph reasoning model has the outstanding abilities of visual inference and visual narrating. The qualitative discussion reveals that our model achieves better performance than other state-of-the-art approaches on the evaluation of both visual inference and visual narrating.

2. Hypergraph-Enhanced Graph Reasoning

In this section, we present the hypergraph-enhanced graph reasoning model in detail, as shown in Figure 2. Specifi-
Taking a given image as input, we detect the visual “person” using the Mask R-CNN [He et al. 2017], which extracts $N^V$ appearance features $\mathbf{Y}^V = \{v_i^V\}_{i=1}^{N^V}$, and their corresponding bounding-box $\mathbf{Y}^B = \{v_i^B\}_{i=1}^{N^B}$, where we encode the top-left position and the bottom-right position of the $i$-th bounding box using the 4-dimensional (-D) vector, i.e., $v_i = [x_i^{top}, y_i^{top}, x_i^{bot}, y_i^{bot}]$. To fuse image features, we calculate the visual features: $\mathbf{F}^V = \{v_i^V\}_{i=1}^{N^V}$, $v_i \in \mathbb{R}^{d_{X^V}}$, where $v_i = u^V v_i^V + u^B v_i^B$, $u^V$ and $u^B$ are learnable parameters, $d_{X^V}$ is the image feature dimension.

### 2.1. Multi-Modal Feature Embedding

In this subsection, we formalize the way with pre-trained models to extract multi-modal features.

#### Visual Features

Taking a given image as input, we detect the visual “person” using the Mask R-CNN [He et al. 2017], which extracts $N^V$ appearance features $\mathbf{Y}^V = \{v_i^V\}_{i=1}^{N^V}$, and their corresponding bounding-box $\mathbf{Y}^B = \{v_i^B\}_{i=1}^{N^B}$, where we encode the top-left position and the bottom-right position of the $i$-th bounding box using the 4-dimensional (-D) vector, i.e., $v_i = [x_i^{top}, y_i^{top}, x_i^{bot}, y_i^{bot}]$. To fuse image features, we calculate the visual features: $\mathbf{F}^V = \{v_i^V\}_{i=1}^{N^V}$, $v_i \in \mathbb{R}^{d_{X^V}}$, where $v_i = u^V v_i^V + u^B v_i^B$, $u^V$ and $u^B$ are learnable parameters, $d_{X^V}$ is the image feature dimension.

#### Textual Features

Following the work of VisualCOMET [Park et al. 2020], a given image corresponding $N^T$-word textual descriptions, including two kind of information (place, and events), is fed into the pre-trained BERT model [Devlin et al. 2019] to obtain the textual feature $\mathbf{F}^T = \{t_i\}_{i=1}^{N^T}$, where $t_i \in \mathbb{R}^{d_{X^T}}$ is the embedding of the $i$-th word, and $d_{X^T}$ is the dimension.

### 2.2. Graph Projection and Hypergraph Construction

In this subsection, in order to capture the intra-modality relationship from an individual modality, we build a multi-modal graph by the Graph Projection. Further, in order to have a picture of the whole situation that different images with descriptions of events at different moments, we construct a hyper-graph through the Hypergraph Construction.

#### Graph Projection

As shown in Figure 3, given an image and its corresponding textual descriptions, we construct a multi-modal graph composed of two sub-graphs, i.e., image-semantic graph $\mathbf{G}^V$, and text-semantic graph $\mathbf{G}^T$ for representing the information in two modalities. For simplicity, we uniformly denote two semantic graphs as $\mathbf{G}^{tag}$ and original feature embedding $\mathbf{F}^{tag}$, $tag \in \{V, T\}$. We project the feature map $\mathbf{F}^{tag}$ from given training samples into the graph $\mathbf{G}^{tag} \in \mathbb{R}^{N^{tag} \times d_{tag}}$, where $N^{tag}$ is the number of nodes.

### The Generator of Event Ratiocination

We employ the BART [Lewis et al. 2020] to build the generator $f^{CEG}_{\theta}(\cdot)$ of event ratiocination and obtain the event’s cause-and-effect descriptions in an autoregressive manner based on graph re-projected representations (i.e., refined sample features).
With the above equation, it is obvious that $H_{\text{Graph Convolution}}$ can be built as a lightweight fully-connected graph. We use the project function $f_{\text{proj}}(\cdot)$ that can be formulated as a linear combination with learnable weights for acquiring the $G^{\text{tag}}$:

$$G^{\text{tag}} = f_{\text{proj}}(f_{\text{red}}(G^{\text{tag}}; W_f^{\text{red}}))$$

$$= f_{\text{wise}}(G^{\text{tag}}; W_f^{\text{wise}}) \times f_{\text{red}}(G^{\text{tag}}; W_f^{\text{red}})$$

(1)

where $f_{\text{wise}}(\cdot)$ and $f_{\text{red}}(\cdot)$ are two convolution layers (Chen et al., 2019b; Liang et al., 2018) for graph projection and feature dimension reduction, respectively. $W_f^{\text{wise}}$ is weights of $f_{\text{wise}}(\cdot)$ and $W_f^{\text{red}}$ is the weights of $f_{\text{red}}(\cdot)$.

### 3. Hypergraph Construction

As shown in Figure 4, given $A^{\text{Hyper}}$ images and their corresponding textual statements including events, places and three event rationization from the training dataset, we construct a hyper-graph $H^{\text{Hyper}}$ that can be represented by the incidence matrix $H^{\text{Hyper}}$. For each hyper-graph vertex $F^{\text{Hyper}} \in (F^T_1 \times \ldots \times F^T_K) \cup (\bar{F}^T_1 \times \ldots \times \bar{F}^T_K)^{\text{Hyper}}$, where the visual feature $F^T$ and textual feature $F^\bar{T}$ are both from $i$-th image. Note that $F^T_i$ is obtained by Mask-RCNN and image, and $F^\bar{T}_i$ is obtained by BERT and textual statements. We find its $\mathcal{K}$ nearest neighbors (Chen et al., 2009), and then utilize each element $H_{ij}^k (i, j = 1, \ldots, 2 \times N^{\text{Hyper}})$ of the hyper-graph incidence matrix to connect these vertices:

$$H_{ij}^k = \begin{cases} 1 & f_{\text{KNN}}^k(\cdot) \in H^{\text{Hyper}}(F^{\text{Hyper}}) \\ 0 & \text{otherwise} \end{cases}$$

(2)

where $f_{\text{KNN}}^k(\cdot)$ is the nearest neighbor function resulting in the neighborhood set containing the top-$\mathcal{K}$ neighbors.

With the above equation, it is obvious that $H^{\text{Hyper}}$ is decided by $f_{\text{KNN}}^k(\cdot)$, which is dictated by the parameter $\mathcal{K}$. Therefore, according to the different values of $\mathcal{K}$, we can get different hyper-graph incidence matrices and thus different hyper-graphs. For the sake of writing, we denote the different incidence matrices as $H^{\text{Hyper}}_\mathcal{K}$. In our model, we adopt the average as the final incidence matrix:

$$H^{\text{Hyper}} = \frac{1}{\mathcal{K}} \sum_{\mathcal{K}} H^{\text{Hyper}}_\mathcal{K}$$

(3)

### 2.3. Graph and Hypergraph Representations

In this subsection, in order to capture semantic relations, we update the node representation of semantic graphs by the Graph Convolution. Similarly, we update the hyper-graph representation through the Hypergraph Convolution to automatically capture high-order semantic relations. Since the images and their statements are from different modalities and different moments, we want to analyze the relationship among these samples to enrich the graph representation from the individual modalities. To this end, we propose the Graph Self-Attention Enhancement, consisting of Graph-Graph Attention Unit and Graph-Hypergraph Attention Unit for the inter-modality relationship between two modalities as well as the spatio-temporal relationship among past, present, and future samples. Further, we employ the Graph Re-Project to transfer the reinforced graph representations into refined sample features.

#### 3.1. Graph Convolution

Based on the obtained graph $G^{\text{tag}}$, we make use of Graph Convolution (Kipf & Welling, 2017) to further propagate information and aims at correlations between the feature of the relative nodes by learning edge weights. In particular, a single graph convolution with its parameter $W_{f_{\text{conv}}} \in \mathbb{R}^{d_{\text{tag}} \times d_{\text{tag}}}$ is defined as:

$$G^{\text{tag}} = f_{\text{conv}}(G^{\text{tag}}; W_{f_{\text{conv}}})$$

$$= ((G^{\text{tag}} - A^{\text{tag}}(G^{\text{tag}}))G^{\text{tag}})W_{f_{\text{conv}}}$$

(4)

where $A^{\text{tag}}$ is the $N^{\text{tag}} \times N^{\text{tag}}$ adjacency matrix of graph $G^{\text{tag}}$ for cross-nodes diffusion, $f_{\text{conv}} \in \mathbb{R}^{d_{\text{tag}} \times d_{\text{tag}}}$ is the identity matrix. A Laplacian smoothing operator (Li et al., 2018) is performed to propagate the node features over the graph. Considering its own representation of each node, the adjacency matrix is added with self-connection. The graph convolution is implemented by two convolution layers along with channel-wise and node-wise directions as shown in Figure 3. The identity matrix $I^{\text{tag}}$ is also a residual connection for every node. The adjacency matrix and its parameter $W_{f_{\text{conv}}}$ can be optimized by gradient descent.

#### 3.2. Hypergraph Convolution

To capture high-order semantic relations automatically, we use the Hypergraph Convolution (Feng et al., 2019) with the hyper-graph to propagate hypergraph information and update hypergraph embeddings, as illustrated in Figure 3:

$$E^{\text{Hyper}} = f^\Phi(H^{\text{Hyper}}; W_f^\Phi) \odot f^\Omega(H^{\text{Hyper}}; W_f^\Omega)$$

(5)

where $f^\Phi(\cdot)$ is the $1 \times 1$ convolution with its weights $W_f^\Phi$ followed by a non-linear activation function (in our case ReLU function (Goodfellow et al., 2016)); $f^\Omega(\cdot)$ with its weights $W_f^\Omega$ is channel-wise global average pooling (GAP) (Goodfellow et al., 2016) followed by a $1 \times 1$ convolution similar to (Hu et al., 2018), and it plays a role in a diagonal matrix, which helps in learning a better distance metric among the nodes for the incidence matrix $H^{\text{Hyper}}$; $f^\Omega(\cdot)$ with its weights $W_f^\Omega$ is a single-layer convolution (Lin et al., 2014) that is used to capture the global relationship of the features to develop better hyper-edges (Wu et al., 2020); $(\cdot)^T$ means the matrix transpose operation; $\odot$ means the matrix dot product calculation operation.
Graph Self-Attention Enhancement

Through the graph convolution process, we get two graph embeddings from two modalities: $\mathbf{X}^T$ and $\mathbf{X}^S$. Then, by the hyper-graph convolution process, we get the hyper-graph embedding from obtained hyper-graphs: $\mathbf{X}^{Hyper}$. We want to (1) use each graph embedding of the two modalities to enhance the graph representation of each individually, (2) use the hyper-graph embedding to enhance the graph representation of each modality. To this end, we propose the cascaded multi-modal structure of stacked attention layers, each of which contains our Graph Self-Attention Enhancement, consisting of two kinds of self-attentions, as shown in Figure 5. At last, we get final output $\mathbf{X}^T$ and $\mathbf{X}^S$.

We propose the graph self-attention enhancements that are an extension of multi-head attention consisting of some parallel heads, in which we replace original scaled dot product attention in classical multi-head attention (Vaswani et al., 2017) with the non-local attention block (Wang et al., 2018b; Zhu et al., 2019; Dong Zhang & Sun, 2020; Yin et al., 2020; Zhu et al., 2020) in each head. Our non-local attention block contains two kinds of attention units: (1) Graph-Graph Attention Unit and (2) Graph-HyperGraph Attention Unit.

Graph-Graph Attention Unit

The graph-graph attention unit focuses on enriching the graph embedding of one modality with the graph embedding of the other modality. In particular, for $len$-th layer, suppose there are two sliced graph embeddings from two modalities as the input of our non-local attention block: $\mathbf{X}_{len}^{m_1}$ and $\mathbf{X}_{len}^{m_2}$, where $m_1, m_2 \in \{V, S\}$. The graph-graph attention unit can be represented as:

$$\mathbf{Q} \mathbf{u}^{m_1}_{len} = \text{query}^{m_1}_{len} (\mathbf{X}^{m_1}_{len}),$$

$$\mathbf{K} \mathbf{e}^{m_2}_{len} = \text{key}^{m_2}_{len} (\mathbf{X}^{m_2}_{len}),$$

$$\mathbf{V} \mathbf{c}^{m_2}_{len} = \text{value}^{m_2}_{len} (\mathbf{X}^{m_2}_{len});$$

$$\mathbf{V} \mathbf{c}^{m_2}_{len} = \text{softmax} ((\mathbf{Q} \mathbf{u}^{m_1}_{len})^T \times \mathbf{X}^{m_2}_{len}) \times (\mathbf{V} \mathbf{c}^{m_2}_{len})^T;$$

$$\mathbf{X}^{m_1}_{len} = \text{cat}((\mathbf{V} \mathbf{c}^{m_2}_{len})^T, \mathbf{Q} \mathbf{u}^{m_1}_{len}; \mathbf{W}_{cat}).$$

where $\text{query}^{m_1}_{len}()$, $\text{key}^{m_2}_{len}()$, and $\text{value}^{m_2}_{len}()$ are three linear transformations; we use the softmax function ([Goodfellow et al., 2016] to get the embeddings $\mathbf{V} \mathbf{c}^{m_2}_{len}$; by referring to the design of the non-local block ([Wang et al., 2018b], $\text{cat}()$ is implemented by a $1 \times 1$ convolution, with $\mathbf{W}_{cat}$ that acts as a weighting parameter to adjust the impor-

Figure 4. Snapshot of Our Built HyperGraphs. Images and their statements with different moments are connected by the same semantics. This way can capture spatio-temporal relationships.

1. Graph-Graph Attention Unit

The graph-graph attention unit focuses on enriching the graph embedding of one modality with the graph embedding of the other modality. In particular, for $len$-th layer, suppose there are two sliced graph embeddings from two modalities as the input of our non-local attention block: $\mathbf{X}^{m_1}_{len}$ and $\mathbf{X}^{m_2}_{len}$, where $m_1, m_2 \in \{V, S\}$. The graph-graph attention unit can be represented as:

$$\mathbf{Q} \mathbf{u}^{m_1}_{len} = \text{query}^{m_1}_{len} (\mathbf{X}^{m_1}_{len}),$$

$$\mathbf{K} \mathbf{e}^{m_2}_{len} = \text{key}^{m_2}_{len} (\mathbf{X}^{m_2}_{len}),$$

$$\mathbf{V} \mathbf{c}^{m_2}_{len} = \text{value}^{m_2}_{len} (\mathbf{X}^{m_2}_{len});$$

$$\mathbf{V} \mathbf{c}^{m_2}_{len} = \text{softmax} ((\mathbf{Q} \mathbf{u}^{m_1}_{len})^T \times \mathbf{X}^{m_2}_{len}) \times (\mathbf{V} \mathbf{c}^{m_2}_{len})^T;$$

$$\mathbf{X}^{m_1}_{len} = \text{cat}((\mathbf{V} \mathbf{c}^{m_2}_{len})^T, \mathbf{Q} \mathbf{u}^{m_1}_{len}; \mathbf{W}_{cat}).$$

where $\text{query}^{m_1}_{len}()$, $\text{key}^{m_2}_{len}()$, and $\text{value}^{m_2}_{len}()$ are three linear transformations; we use the softmax function ([Goodfellow et al., 2016] to get the embeddings $\mathbf{V} \mathbf{c}^{m_2}_{len}$; by referring to the design of the non-local block ([Wang et al., 2018b], $\text{cat}()$ is implemented by a $1 \times 1$ convolution, with $\mathbf{W}_{cat}$ that acts as a weighting parameter to adjust the impor-
At the last layer, we use the proposed graph self-attention where

\[
\text{Qu} \cdot \text{Klen} = \text{Vlen}, \quad \text{Vlen} = \text{softmax}(\text{Qlen} \cdot \text{Klen})
\]

Therefore, we re-denote these vectors as

\[
\text{len} \in \{1, 2, 3, 4, 5, 6, 7, 8\}
\]

Graph-HyperGraph Attention Unit

Similarly, the graph-hypergraph attention unit aims at the aggregation of hypergraph embedding and the results of the former unit. There are two kinds of attention layers with the stack fashion. The graph-graph attention unit focuses on aggregating the graph embedding of one modality and the graph embedding of the other modality. Similarly, the graph-hypergraph attention unit aims at improving the graph embedding of each modality with hypergraph embedding from obtained hyper-graph, under the condition of the previous step in the graph-graph attention unit. In particular, for \(len\)-th layer, we employ another non-local attention block to process the result \(Z_{len}^{m_1} \, Z_{len}^{m_2}\) of the graph-graph attention unit and the sliced hyper-graph embeddings \(Z_{len}^{\text{Hyper}}\) from obtained hyper-graphs, as follows:

\[
\begin{align*}
\mathcal{V}^{len} &:= \text{softmax}(\langle \mathcal{Q} \mathcal{W}^{len} \rangle^{T} \times \mathcal{K}_{len}^{\text{Hyper}}) \\
Z_{len}^{\text{Hyper}} &:= f_{\text{cat}}(\langle \mathcal{V}^{len} \rangle^{T} \mathcal{V}_{len}^{m_1} Z_{len}^{m_2}) \quad \text{(7)}
\end{align*}
\]

where \(\mathcal{Q} \mathcal{W}^{len}\) is from \(Z_{len}^{m_1} \cdot Z_{len}^{\text{Hyper}}\) and \(\mathcal{V}^{len} \cdot Z_{len}^{\text{Hyper}}\) is from \(Z_{len}^{\text{Hyper}}\). For rich contextual information, we use the strip pooling \(f_{\text{cat}}(\cdot)\) to enhance models.

At the last layer, we use the proposed graph self-attention enhancements for the vectors \(Z_{len}^{\text{Hyper}}\) and \(Z_{len}^{\text{Hyper}}\), respectively. To inform the model of different modalities of input, we add two sets of special tokens: for images, we use \(<\text{img}>\) and \(<\text{/img}>\). For refined linguistic features as final refined features \(\mathcal{F}^{\text{Vv}}\) and \(\mathcal{F}^{\text{Vf}}\).

2.4. The Generator of Event Ratiocination

We develop a cause-and-effect generator (CEG) as shown in Figure 3 based on Lewis et al. (2020)'s work and with the refined features \(\mathcal{F}^{\text{Vv}}\) and \(\mathcal{F}^{\text{Vf}}\) in the sequence-to-sequence task (Bao et al., 2020) to transfer into the predicted output \(\tilde{\mathcal{O}}\). Formally, CEG \(f_{\text{CEG}}(\cdot)\) with its parameter \(\mathcal{W}_{f_{\text{CEG}}}\) is:

\[
\tilde{\mathcal{O}} = f_{\text{CEG}}(\mathcal{F}^{\text{Vv}}, \mathcal{F}^{\text{Vf}}, \mathcal{W}_{f_{\text{CEG}}}) \quad \text{(8)}
\]

Encoder

Following the BART (Lewis et al., 2020) and its variant (Xing et al., 2021), the encoder of CEG is based on a multi-layer bidirectional Transformer (Dai et al., 2019), as shown in Figure 3. We use \(<\text{before}>\), \(<\text{after}>\), or \(<\text{intent}>\) as the starting special token. To inform the model of different modalities of input, we add two sets of special tokens: for images, we use \(<\text{img}>\) and \(<\text{/img}>\) to indicate the start and the end of refined visual feature \(\mathcal{F}^{\text{Vv}}\), respectively. To inform the model textual inputs, we use \(<\text{text}>\) and \(<\text{/text}>\) for refined linguistic feature \(\mathcal{F}^{\text{Vf}}\).

Decoder

The decoder of our model is also a multi-layer Transformer, similar to Wang et al. (2019b). Different from the encoder,
which is bidirectional, the decoder is unidirectional as it is supposed to be autoregressive when generating texts. The decoder does not take as inputs the visual embeddings.

**Pre-Training**

We pre-train a CEG model with 12 layers in each of the encoder and decoder, and a hidden size of 1024. Following RoBERTa (Liu et al., 2019), we use a batch size of 8000, and train the model for 500,000 steps. Documents are tokenized with the same byte-pair encoding as GPT-2 (Radford et al., 2019). To pretrain our model, we use three image-text datasets: Conceptual Captions Dataset (Sharma et al., 2018), Im2Text Dataset (Ordonez et al., 2011) and Microsoft COCO Dataset (Lin et al., 2014b). We pre-train the encoder in two steps, in both cases back-propagating the cross-entropy loss from the output of the CEG model. In the first step, we freeze the parameter \( \mathbf{W}_{fCEG} \) and only update the randomly initialized encoder, and the self-attention input projection matrix of encoder first layer. In the second step, we pre-train all CEG model parameters for a small number of iterations.

### 3. Experiments and Results

In this section, we experimentally evaluate the proposed model on the benchmark datasets and compare its performance with other state-of-the-arts.

#### 3.1. Benchmark Dataset Description

VisualCOMET dataset (Park et al., 2020) consists of over 1.4 million textual statements of visual event ratiocination carefully annotated over a diverse set of 59,000 images, each paired with short video summaries of before and after.

#### 3.2. Experimental Setup

In this subsection, we outline the used evaluation metrics and implementation details.

**Evaluation Metrics**

VisualCOMET dataset includes 1174K training examples and 146K validation examples. Some examples in the dataset share the same images or events, but with different ratiocination for events before/after or intents at present. Following Park et al. (2020), we report three metrics: BLEU-2 (Papineni et al., 2002), METEOR (Denkowski & Lavie, 2014), and CIDEr (Vedantam et al., 2015). Following the work of (Xing et al., 2021), we report our model performance on the validation set as the test set is not available yet.

**Implementation Details**

In our training process, Adam optimizer (Kingma & Ba, 2015) is used with momentum parameters setting \( \beta_1 \) and \( \beta_2 \) to 0.9 and 0.999. The learning rate is initially set to 0.0001. The training batch size is set to 64. For our built hypergraph, we set the \( K \) to 8. For our graph self-attention enhancements, the number of head in multi-head attention is 8. We set the number of the stacked graph-graph attention unit and graph-hypergraph attention unit to 2.

### 3.3. Comparison with State-of-The-Arts

We compare the state-of-the-art methods with our model on the VisualCOMET benchmark.

**Effect of Graph Reasoning.** From Table 1, #1 is better than Vision&Lang Transformer and KM-BART. This suggests that the generator of event ratiocination is effective. Compared to #1, ours and other variants (i.e., #2∼#5) have better performance. Graph reasoning can help the model capture the intra-modality relationship, and the model with graph reasoning is better than without this process. This implies that the design of graph reasoning is effective.

**Effect of Graph-Graph Attention Unit.** The graph-graph attention unit strives to reinforce the graph embedding of one modality with the graph embedding of the other modality. In essence, the graph embedding as the query’s input exploits this unit with the information of another modality to enhance itself. In this way, the model can capture the inter-modality relationship. From Table 1 it is clear that the method (i.e., #5) with the graph-graph attention unit is better than without this unit (i.e., #2∼#4). It shows the proposed graph-graph attention unit effectively improves the task of visual event ratiocination.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU-2</th>
<th>METEOR</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vision&amp;Lang Transformer (Park et al., 2020)</td>
<td>13.50</td>
<td>11.55</td>
<td>18.27</td>
</tr>
<tr>
<td>KM-BART (Xing et al., 2021)</td>
<td>23.47</td>
<td>15.02</td>
<td>39.76</td>
</tr>
<tr>
<td>Ours</td>
<td>32.97</td>
<td>23.99</td>
<td>49.19</td>
</tr>
</tbody>
</table>

**Effect of Graph-HyperGraph Attention Unit.** Similarly, the graph-hypergraph attention unit is the extension of the graph-graph attention unit and integrates the graph embedding from two modalities with the hypergraph. In other words, the graph embedding from the hypergraph is employed with the graph-hypergraph attention to strengthen the
Figure 6. The Comparison Results of Visual Inference. (a) Results on The VIOLIN Dataset; (b) Results on The VLEP Dataset.

The visual inference is one of the key to the task of visual event ratiocination and requires the model that devotes itself to the judgment of the temporal event’s ratiocination for a given image. In order to further evaluate the performance of the proposed model, we perform two tasks: video-and-language inference (Liu et al., 2020) and video-and-language future event prediction (Lei et al., 2020).

Video-and-Language Future Event Prediction

Video-and-language future event prediction focuses on future event prediction from videos. In particular, given a video with dialogue, and two possible future events, the model is required to understand both visual and language semantics from this video, and choose the more likely event from two provided possible future events. Similarly, to fit this task, for our model, we split the given video clip into frames, and compare the ratiocination generated by our model with the candidates to obtain the final results by the Bi-LSTM (Huang et al., 2015). We use accuracy as evaluation metric.

Dataset. VIOLIN benchmark dataset (Liu et al., 2020) consists of 95,322 video hypothesis pairs from 15,887 video clips, spanning over 582 hours of video. These video clips contain rich content with event shifts, and interaction.

Ours is better than LXMERT (Tan & Bansal, 2019; Liu et al., 2020), HERO (Li et al., 2020). This suggests that our model has better performance than other state-of-the-art approaches for the task of video-and-language inference.

Video-and-Language Future Event Prediction

Video-and-language event prediction focuses on future event prediction from videos. In particular, given a video with dialogue, and two possible future events, the model is required to understand both visual and language semantics from this video, and choose the more likely event from two provided possible future events. Similarly, to fit this task, for our model, we split the given video clip into frames, and compare the ratiocination generated by our model with the candidates to obtain the final results by the Bi-LSTM (Huang et al., 2015). We use accuracy as evaluation metric.

Dataset. VLEP benchmark dataset (Lei et al., 2020) contains 28,726 examples from 10,234 short video clips. Each example consists of a short video clip with its dialogue and text summary, and two potential future event ratiocination.

Ours is better than RoBERTa (Liu et al., 2019; Lei et al., 2020). As a results, our model has state-of-the-art performance for the task of video-and-language event prediction.

3.4. Discussion on The Ability of Visual Inference

The visual inference is one of the key to the task of visual event ratiocination, and requires the model that devotes itself to the judgment of the temporal event’s ratiocination for a given image. In order to further evaluate the performance of the proposed model, we perform two tasks: video-and-language inference (Liu et al., 2020) and video-and-language future event prediction (Lei et al., 2020).

Video-and-Language Inference

Video-and-language inference aims at the joint understanding of video and text. Given a video clip with aligned subtitles as the premise, paired with a natural language hypothesis, a model needs to infer whether the hypothesis is entailed or contradicted by the given video clip. To fit this task, for our model, we replace the textual descriptions in our original task (i.e., visual event ratiocination) with subtitles and statements, split the given video clip into frames. We compare the ratiocination generated by our model with the candidates to obtain the final results. Specifically, we input the generated ratiocination and the candidate ones to the Bi-LSTM (Huang et al., 2015; Liu et al., 2020) to make the prediction. We use accuracy as evaluation metric.

Ours is better than HERO (Tan & Bansal, 2019; Liu et al., 2020), and compared to the proposal on the VisualCOMET dataset.

3.5. Discussion on The Ability of Visual Narrating

Visual narrating is another key for the task of visual event ratiocination and focuses on generating semantic descriptions from images or videos, e.g., video captioning (Shetty & Laaksonen, 2016) and visual storytelling (Huang et al., 2016). To further evaluate our model, we perform these two tasks in this subsection. To fit these two tasks, for our model, we split the given video clip into frames and fix the starting special token as \(<\text{intent}>\). We use BLEU-4 (Papineni et al., 2002), METEOR, and CIDEr as evaluation metrics.

Video Captioning

The goal of video captioning is to generate a sentence to describe video content accurately. Here, we introduce the used dataset followed by the comparison analysis.

Dataset. MSVD (Chen & Dolan, 2011) contains 1,970 video clips with multiple descriptions for each video clip. Following the work of Venugopalan et al. (2015), we use
1, 200 video clips for training, 100 video clips for validation, and 670 video clips for testing.

Performance of Different Methods. On the MSVD dataset, we compare ours with 29 state-of-the-arts. From Table 2, the proposed model significantly outperforms existing state-of-the-arts. From above, our model is more robust than other state-of-the-arts on the task of video captioning.

Visual Storytelling

Visual storytelling requires the model to understand the event flow in these photos deeply. Here, we introduce the used dataset followed by the comparison analysis.

Dataset. The VIST dataset (Huang et al., 2016) is used for solving visual storytelling, which includes 10, 117 Flickr albums with 210, 819 unique images. After filtering the broken images, there are 40, 098 training, 4, 988 validation, and 5, 050 testing samples.

Performance of Different Methods. On the VIST dataset, we compare ours with 13 state-of-the-arts. From Table 3, the proposed model significantly outperforms existing state-of-the-arts. From above, our model is more robust than other state-of-the-arts on the task of visual storytelling.

Comparison Results on The Task of Video Captioning. Best is pointed in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2VT (Venugopalan et al., 2015)</td>
<td>42.1</td>
<td>30.0</td>
<td>58.8</td>
</tr>
<tr>
<td>MP-LSTM (Venugopalan et al., 2015)</td>
<td>50.40</td>
<td>32.50</td>
<td>71.00</td>
</tr>
<tr>
<td>LSTM-E (Pan et al., 2016)</td>
<td>45.30</td>
<td>31.00</td>
<td>-</td>
</tr>
<tr>
<td>p-RNN (Yu et al., 2016)</td>
<td>47.40</td>
<td>30.30</td>
<td>53.60</td>
</tr>
<tr>
<td>Tempor-attention (Yao et al., 2015)</td>
<td>41.92</td>
<td>29.60</td>
<td>51.67</td>
</tr>
<tr>
<td>Bi-GRU-RCN (Ballas et al., 2016)</td>
<td>48.42</td>
<td>31.70</td>
<td>65.38</td>
</tr>
<tr>
<td>hLSTMat (Song et al., 2017)</td>
<td>48.50</td>
<td>31.90</td>
<td>-</td>
</tr>
<tr>
<td>MAMRNN (Li et al., 2017)</td>
<td>41.40</td>
<td>32.20</td>
<td>53.90</td>
</tr>
<tr>
<td>PickNet (Chen et al., 2018)</td>
<td>46.10</td>
<td>33.10</td>
<td>76.00</td>
</tr>
<tr>
<td>MCF (Wu &amp; Han, 2018)</td>
<td>46.46</td>
<td>33.72</td>
<td>75.46</td>
</tr>
<tr>
<td>RecNet (Wang et al., 2018a)</td>
<td>52.30</td>
<td>34.10</td>
<td>80.30</td>
</tr>
<tr>
<td>GRU-EVE (Xiao et al., 2019)</td>
<td>45.60</td>
<td>33.70</td>
<td>74.20</td>
</tr>
<tr>
<td>Middle-out-self MemNet &amp; Sigal, 2018</td>
<td>47.00</td>
<td>34.10</td>
<td>79.50</td>
</tr>
<tr>
<td>TDCOnED (Chen et al., 2019)</td>
<td>48.30</td>
<td>32.90</td>
<td>72.30</td>
</tr>
<tr>
<td>MenNet (Wu et al., 2020)</td>
<td>51.62</td>
<td>34.85</td>
<td>84.27</td>
</tr>
</tbody>
</table>

Comparison Results on The Task of Visual Storytelling. Best is pointed in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU-4</th>
<th>ROUGE</th>
<th>METEOR</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSRL (Huang et al., 2019)</td>
<td>13.4</td>
<td>35.2</td>
<td>30.8</td>
<td>-</td>
</tr>
<tr>
<td>hattn-rank (Yu et al., 2017)</td>
<td>-</td>
<td>34.1</td>
<td>29.5</td>
<td>7.5</td>
</tr>
<tr>
<td>CIDEr-RL (Wang et al., 2018a)</td>
<td>13.8</td>
<td>34.9</td>
<td>29.7</td>
<td>8.1</td>
</tr>
<tr>
<td>GAN (Wang et al., 2018a)</td>
<td>14.0</td>
<td>35.0</td>
<td>29.5</td>
<td>9.0</td>
</tr>
<tr>
<td>VSCMR (Li et al., 2019)</td>
<td>14.3</td>
<td>35.5</td>
<td>30.2</td>
<td>9.0</td>
</tr>
<tr>
<td>ARLE-RL (Wang et al., 2018a)</td>
<td>14.1</td>
<td>35.0</td>
<td>29.5</td>
<td>9.4</td>
</tr>
<tr>
<td>MenNet (Wu et al., 2020)</td>
<td>14.1</td>
<td>35.5</td>
<td>29.5</td>
<td>9.2</td>
</tr>
<tr>
<td>StoryAnchor (Zhang et al., 2020)</td>
<td>14.0</td>
<td>35.5</td>
<td>30.2</td>
<td></td>
</tr>
<tr>
<td>SGVST (Wang et al., 2020b)</td>
<td>14.7</td>
<td>35.5</td>
<td>30.2</td>
<td></td>
</tr>
<tr>
<td>K-Storyteller (Yang et al., 2019)</td>
<td>12.8</td>
<td>29.9</td>
<td>35.2</td>
<td></td>
</tr>
<tr>
<td>TAVST (Wang et al., 2020a)</td>
<td>14.6</td>
<td>31.0</td>
<td>35.7</td>
<td></td>
</tr>
<tr>
<td>Net (Ji et al., 2020)</td>
<td>14.7</td>
<td>35.6</td>
<td>29.7</td>
<td></td>
</tr>
<tr>
<td>S2VT-RL+DRPN (Xu et al., 2020)</td>
<td>49.2</td>
<td>32.6</td>
<td>86.4</td>
<td></td>
</tr>
<tr>
<td>MemNet (Wu et al., 2020)</td>
<td>51.62</td>
<td>34.85</td>
<td>84.27</td>
<td></td>
</tr>
</tbody>
</table>

4. Conclusion and Future Work

In this paper, we present a novel multi-modal model Hypergraph-Enhanced Graph Reasoning for the task of visual event rationcination. The model firstly represents the image with multi-modal contents as two semantic graphs, where each graph represents one modality. Then, the proposed Graph Self-Attention Enhancement in our model, rises up each graph representations with the help of our built hypergraph, followed by re-projecting back into the original spatial feature domain. Finally, we obtain the cause-and-effect descriptions with these finer representations of elements about the image. Experimental results show that our model achieves state-of-the-art performance. Moving forward, we will take rich structured information especially effective knowledge graphs as the guidance for our model.

Acknowledgements

We would like to thank all anonymous reviewers for their useful feedback. This work is supported in part by MOST and NNSF of China (2008AAA0101502, 61806198, U1811463), and Squirrel AI Learning.
References


Hypergraph-Enhanced Graph Reasoning for Visual Event Ratiocination


Hypergraph-Enhanced Graph Reasoning for Visual Event Ratiocination


