# Joint Commonsense and Relation Reasoning for Image and Video Captioning

**Jingyi Hou,**[1] **Xinxiao Wu,**[1*] **Xiaoxun Zhang,**[2] **Yayun Qi,**[1] **Yunde Jia,**[1] **Jiebo Luo**[3]
[1]Lab. of IIT, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China
[2]Alibaba Group
[3]Department of Computer Science, University of Rochester, Rochester NY 14627, USA
{houjingyi, wuxinxiao, 1120163657, jiayunde}@bit.edu.cn, xiaoxun.zhang@alibaba-inc.com, jluo@cs.rochester.edu

## Abstract

Exploiting relationships between objects for image and video captioning has received increasing attention. Most existing methods depend heavily on pre-trained detectors of objects and their relationships, and thus may not work well when facing detection challenges such as heavy occlusion, tiny-size objects, and long-tail classes. In this paper, we propose a joint commonsense and relation reasoning method that exploits prior knowledge for image and video captioning without relying on any detectors. The prior knowledge provides semantic correlations and constraints between objects, serving as guidance to build semantic graphs that summarize object relationships, some of which cannot be directly perceived from images or videos. Particularly, our method is implemented by an iterative learning algorithm that alternates between 1) commonsense reasoning for embedding visual regions into the semantic space to build a semantic graph and 2) relation reasoning for encoding semantic graphs to generate sentences. Experiments on several benchmark datasets validate the effectiveness of our prior knowledge-based approach.

## Introduction

Most existing methods for image and video captioning (Donahue et al. 2015; Venugopalan et al. 2015b; 2015a; Pan et al. 2016) are based on the encoder-decoder framework which directly translates visual features into sentences, without exploiting high-level semantic entities (e.g., objects, attributes, and concepts) as well as relations among them. Recent work (Yao et al. 2018; Li and Jiang 2019; Yang et al. 2019) has shown promising efforts of using a scene graph that provides an understanding of semantic relationships for image captioning. These methods usually use pre-trained object and relationship detectors to extract a scene graph and then reason about object relationships in the graph. However, when facing detection challenges, such as heavy occlusion, tiny-size objects, and the long-tail problem, this paradigm might not accurately depict the objects and their relationships in images or videos, thus resulting in a degradation of captioning performance.

As we know, human beings can still describe images and videos by summarizing object relationships when some objects are not precisely identified or even absent, thanks to their remarkable reasoning ability based on prior knowledge. This inspires us to explore how to leverage prior knowledge to achieve relation reasoning in captioning, mimicking the human reasoning procedure. As an augmentation of the object relationships explicitly inferred from an image or a video, the prior knowledge about object relationships in the world provides information that is not available in the image or video. For example, as shown in Figure 1, the caption of "Several people waiting at a race holding umbrellas" will be generated via prior knowledge when describing a crowd of people standing along the road, even if the image shows no players or running actions (perhaps because the game is yet to begin). Clearly, the relationship of "people waiting race" is inferred from the commonsense relationship between "people" and "race" rather than from the image. Therefore, it is beneficial to integrate prior knowledge with visual information to reason relationships for generating accurate and reasonable captions.

In this paper, we utilize prior knowledge to guide the reasoning of object relationships for image and video captioning. The prior knowledge provides commonsense semantic correlations and constraints between objects to augment visual information extracted from images or videos. We employ external knowledge graphs in Visual Genome (Krishna et al. 2017) which represents a type of prior knowledge in that the nodes represent the objects and the edges denote the relations between nodes.

To effectively apply the prior knowledge into image and video captioning, we propose a joint commonsense and relation reasoning (C-R Reasoning) method that integrates both commonsense reasoning and relation reasoning, and implements them simultaneously. The commonsense reasoning selects local visual regions and maps them into a high-level semantic space to build a semantic graph by using the semantic constraints about relations in the knowledge graphs. The relation reasoning encodes the semantic graph by refining the representations of regions through a graph convolutional network (GCN) to generate textual descriptions. To be specific, we develop an iterative learning algorithm which

[*]Corresponding author.

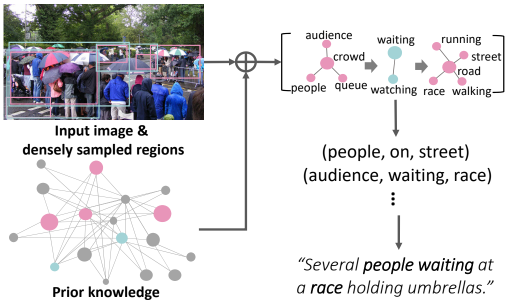


Figure 1: An example of how commonsense reasoning facilitates image and video captioning in our work. The concept "race" is absent from the image but can be inferred from prior knowledge via commonsense reasoning.

alternates between building semantic graph via commonsense reasoning and generating captions via relation reasoning.

Our method does not rely on any pre-trained detectors and does not require any annotations of semantic graphs for training. By discovering the inherent relationships guided by prior knowledge, our method can identify objects that are difficult to detect or even absent from images or videos. Another merit of our method lies in the ability of reaching semantic coherency within a video or image for captioning, which alleviates the problem of semantic inconsistency between the pre-defined object or relationship categories and the target lexical words in existing methods (Yao et al. 2018; Li and Jiang 2019; Yang et al. 2019; Aditya et al. 2018; Zhou, Sun, and Honavar 2019).

## Related Work

Recently, exploiting relationships between objects for image captioning has received increasing attention. Yao et al. (2018) employed two graph convolutional networks (GCNs) to reason semantic and spatial correlations among visual features of detected objects and their relationships to boost image captioning. Li et al. 2019 generated scene graphs of images by detectors, and built a hierarchical attention-based model to reason visual relationships for image captioning. Yang et al. 2019 incorporated language inductive bias into a GCN based image captioning model to not only reason relationship via GCN but also represent visual information in language domain via a scene graph auto-encoder for easier translation. These methods explicitly exploit high-level semantic concepts via the pre-defined scene graph of each image and the annotations of object and relationship locations in the image. Quite different from their methods, our method utilizes prior knowledge to generate a graph of latent semantic concepts in an image or a video, without requiring any pre-trained detectors. Moreover, our iterative algorithm enables the scene graph generation and captioning to be trained in an end-to-end manner, thus alleviates the semantic inconsistency between the pre-defined object/relation categories and the target lexical words.

Some recent methods apply external knowledge graphs for image captioning. In (Aditya et al. 2018), the commonsense reasoning is used to detect the scene description graph of an image, and the graph is directly translated into a sentence via a template-based language model. CNet-NIC (Zhou, Sun, and Honavar 2019) incorporates knowledge graphs to augment information extracted from images for captioning. Different from these methods that directly extract explicit semantic concepts from external knowledge, our method uses external knowledge to reason relationships between semantic concepts via joint commonsense and relation reasoning, without facing the "hallucinating" problem as stated by (Rohrbach et al. 2018).

Some Visual Question Answering (VQA) methods (Berant et al. 2013; Fader, Zettlemoyer, and Etzioni 2014; Su et al. 2018; Mao et al. 2019) apply commonsense or relation reasoning. In these methods, almost the entire semantic graph is given in terms of the question sentences, while the semantic graph is built only by using the input visual cues for image and video captioning with reasoning. The reasoning problem in image and video captioning is thus more challenging. To tackle this problem, we leverage the prior knowledge to help reasoning and propose a joint learning method to implement the reasoning.

## Our Method

Our method consists of three modules: visual mapping and knowledge mapping, commonsense reasoning, and relation reasoning, as shown in Figure 2. In the visual mapping and knowledge mapping module, the candidate proposals of semantic entities are generated, and then the visual feature vectors of the proposals are learned via visual mapping, and the knowledge vectors of the proposals via knowledge mapping. In the commonsense reasoning module, given the candidate proposals, a semantic graph is built under the guidance of the prior knowledge graphs. In the relation reasoning module, given the semantic graph, textual descriptions are generated via GCN and a sequence-based language model.

### Visual Mapping

The goal of visual mapping is to generate candidate proposals of semantic entities such as objects, attributes and relationships. Specifically, the proposals of objects and attributes are represented by visual features of local regions. The relationship proposals are represented by visual features of the union areas of two local regions. The local region refers to a 2D patch in images or a 3D cuboid in videos. We densely sample local regions from input images or videos, and then features of the regions are extracted using pre-trained CNNs. We cluster on the sampled regions to obtain typical candidate proposals that are represented by the cluster centers. Let $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{N_v}] \in \mathbb{R}^{L_v \times N_v}$ denote the visual feature vectors of the candidate proposals, where $\boldsymbol{v}_i \in \mathbb{R}^{L_v \times 1}$ is the visual feature vector of the $i$-th candidate proposal and $N_v$ is the number of candidate proposals.

### Knowledge Mapping

Knowledge mapping aims at learning knowledge vectors of the candidate proposals by projecting the visual feature vectors $\boldsymbol{V}$ of the candidate proposals onto a semantic concept

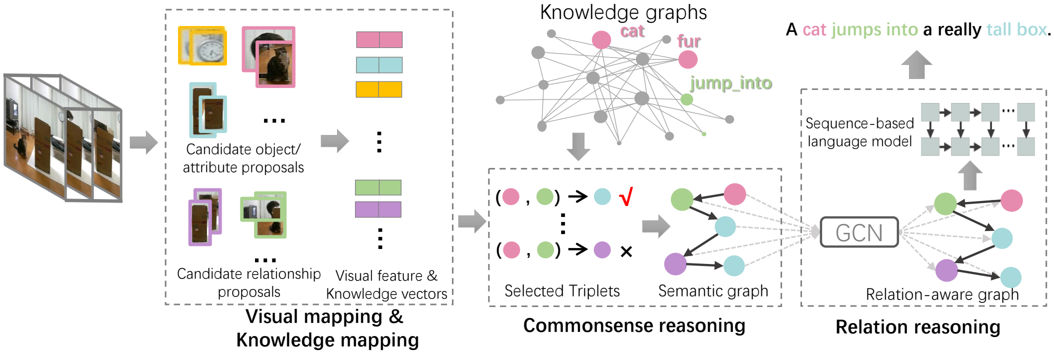

Knowledge graphs
cat
fur
jump_into
A cat jumps into a really tall box.
Sequence-based language model
Candidate object/ attribute proposals
Candidate relationship proposals
Visual feature & Knowledge vectors
√
×
Selected Triplets
Semantic graph
GCN
Relation-aware graph
Visual mapping & Knowledge mapping
Commonsense reasoning
Relation reasoning

Figure 2: Overview of our method for video captioning. We first densely sample spatial-temporal regions from the input videos and cluster the regions according to their visual appearances to obtain the candidate proposals. Then we project the proposals into semantic spaces via visual and knowledge mappings and build a semantic graph via commonsense reasoning guided by prior knowledge graphs. Finally, we learn the relation-aware graph via relation reasoning for captioning, and the generated sentence in turn refines the knowledge graphs. For image captioning, the spatial image regions are first densely sampled.

space with knowledge embedding vectors of prior knowledge. The knowledge embedding vectors are calculated by using knowledge graphs on the Visual Genome[1] via complEX (Trouillon et al. 2016). Supposing that there are totally $C$ semantic concepts in the knowledge graphs, let $\boldsymbol{E} = [\boldsymbol{e}_1, ..., \boldsymbol{e}_C] \in \mathbb{C}^{L_k \times C}$ represent the knowledge embedding vectors where $\boldsymbol{e}_i \in \mathbb{C}^{L_k \times 1}$ denotes the $i$-th semantic concept and $\mathbb{C}$ is the complex domain that enables the knowledge embedding vectors to represent directed knowledge graphs. The knowledge vectors of the candidate proposals are derived from the aggregation of the knowledge embedding vectors weighted by a soft-assignment that is implemented by a non-linear mapping network. Let $\boldsymbol{K} = [\boldsymbol{k}_1, \dots, \boldsymbol{k}_{N_v}] \in \mathbb{C}^{L_k \times N_v}$ represent the knowledge vectors of the candidate proposals, where $\boldsymbol{k}_i$ denotes the knowledge vector of the $i$-th candidate proposal, $\boldsymbol{k}_i = \boldsymbol{E}\boldsymbol{p}_i$ and $\boldsymbol{p}_i \in \mathbb{R}^{C \times 1}$ represent the weights of knowledge embedding vectors. Since there are three kinds of semantic concepts (object, relationship, and attribute), we build three non-linear mapping networks to soft-assign the visual feature vectors with concept labels of object, relationship, and attribute, respectively. The ground-truth labels of objects (resp. relationships and attributes) are simply derived from the nouns (resp. verbs and adjectives) of the ground-truth sentences via POS tagging using the NLTK toolkit (Xue 2011). The training and inference procedures of the three networks are similar, so we only describe details of the network for the object below.

During training, we apply a multiple self-attention mechanism to the visual feature vectors $\boldsymbol{V}$ of an image or a video to make the network focus on the relevant candidate proposals to the ground-truth. Specifically, $K$ attention operations are used to obtain vectors $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_K] \in \mathbb{R}^{L_v \times K}$, where $\boldsymbol{z}_k \in \mathbb{R}^{L_v \times 1}$ represents the vector after the $k$-th attention operation, given $\boldsymbol{z}_k = \boldsymbol{V}\boldsymbol{a}_k^\top$, and $\boldsymbol{a}_k \in \mathbb{R}^{N_v \times 1}$ represents the $k$-th attention weights calculated by a non-linear mapping with the sparsemax (Martins and Astudillo 2016) activation. The predicted object class probabilities of $\boldsymbol{V}$ are calculated as $\sigma(\sum_{k=1}^{K} f(\boldsymbol{z}_k)) \in \mathbb{R}^{C \times 1}$, where $f(\cdot)$ is a linear mapping function that maps $\boldsymbol{z}_k$ to a $C$ dimensional space and $\sigma(\cdot)$ is a sigmoid operation. Given the predicted class probabilities and ground-truth class labels of objects, the network for the object is trained with a binary cross-entropy loss function. Moreover, to encourage the model to focus on diverse objects in each image or video, we set a constraint $C$ to regularize $f(\boldsymbol{Z})$, formulated by

$$C = -\sum_{i \neq j} \mathrm{KL}(p(f(\boldsymbol{z}_i))||p(f(\boldsymbol{z}_j))), \tag{1}$$

where $\mathrm{KL}(\cdot)$ denotes the Kullback–Leibler divergence and $p(\cdot)$ is a softmax function.

During inference, the visual feature vector of each proposal is directly fed into the object network without attention operations, i.e., $\boldsymbol{z}_i = \boldsymbol{v}_i$. A sparsemax operation is used to normalize $f(\boldsymbol{z}_i)$ to generate the weights of embedding vectors of the knowledge graphs, and thus the knowledge vector $\boldsymbol{k}_i$ is given by

$$\boldsymbol{k}_i = \boldsymbol{E}\boldsymbol{p}_i, \quad \boldsymbol{p}_i = \mathrm{sparsemax}(f(\boldsymbol{z}_i)). \tag{2}$$

## Joint Commonsense and Relation Reasoning

After learning the visual feature vectors $\mathcal{V}$ and the knowledge vectors $\mathcal{K}$ of all the candidate proposals from training data, we implement image and video captioning by alternatively conducting commonsense reasoning and relation reasoning, as illustrated in Figure 3. The commonsense reasoning constructs the semantic graph of candidate proposals with the guidance of triplet constraints summarized in the knowledge graphs. The relation reasoning learns the relation-aware features via a GCN and generates textual descriptions using a sequence-based language model.

**Commonsense reasoning.** Taking visual feature vectors $\mathcal{V}$ and knowledge vectors $\mathcal{K}$ as input, we further represent the

[1] Note that Visual Genome is a large-scale dataset containing images annotated by triples of semantic concepts (i.e., objects, attributes, and relationships), but we construct the knowledge graphs only using the triples without images and bounding box annotations.

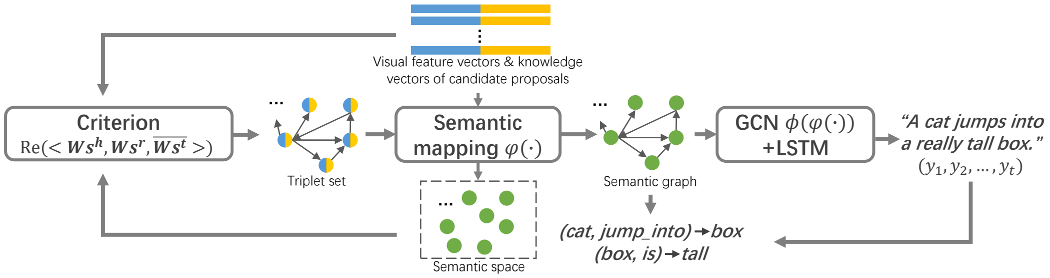

Visual feature vectors & knowledge vectors of candidate proposals
Criterion
Re(< $\boldsymbol{W}\boldsymbol{s}^h, \boldsymbol{W}\boldsymbol{s}^r, \overline{\boldsymbol{W}\boldsymbol{s}^t}$ >)
...
Triplet set
Semantic mapping $\varphi(\cdot)$
...
Semantic space
...
Semantic graph
GCN $\phi(\varphi(\cdot))$ +LSTM
"A cat jumps into a really tall box."
$(y_1, y_2, ..., y_t)$
(cat, jump_into)→box
(box, is)→tall

Figure 3: Our C-R reasoning. We design "semantic mapping" and "GCN+LSTM" modules for commonsense reasoning and relation reasoning, respectively. The two modules are alternatively updated through back propagation. Given the features of the candidate proposals, the "criterion" module selects semantic features of the graph learned by the "semantic mapping" module.

candidate proposals as semantic features $\mathcal{S}$ using a non-linear mapping function: $\boldsymbol{s}_i = \varphi(\boldsymbol{v}_i, \boldsymbol{k}_i)$, $\boldsymbol{s}_i \in \mathcal{S}$, i.e., semantic mapping. The semantic features are learned to satisfy that the correlations and constrains among objects, relationships and attributes are inferred by a commonsense reasoning criterion to generate the semantic graph of an image or a video. The semantic mapping $\varphi(\cdot)$ is updated by the back-propagation of the Commonsense and Relation Reasoning (C-R Reasoning) framework. Different from existing methods of visual relationship detection (Liang, Lee, and Xing 2017; Lu et al. 2016; Zhang et al. 2017) that utilize language prior or regularize relation embedding space, our method leverages commonsense reasoning in the semantic space to generate a relevant semantic graph for describing the image or video without requiring any explicit supervision.

Concretely, the knowledge graphs are collections of factual triplets, where each triplet represents a relationship between a head entity and a tail entity. Let $\mathcal{S}^h$, $\mathcal{S}^r$ and $\mathcal{S}^t$ represent the entity sets of head, relationship and tail. We learn a commonsense reasoning criterion to represent the semantic features via complex vectors, therefore not only the symmetric but also the antisymmetric relations among the entities can be measured. Following (Trouillon et al. 2016), we set the criterion to measure the real part of the composition of the semantic triplet $(\boldsymbol{s}^h, \boldsymbol{s}^r, \boldsymbol{s}^t)$ and represent the correlation, and thus the correlation in the triplet is given by

$$\begin{aligned} &\mathrm{Re}(< \boldsymbol{W}\boldsymbol{s}^h, \boldsymbol{W}\boldsymbol{s}^r, \overline{\boldsymbol{W}\boldsymbol{s}^t} >) \\ &= < \mathrm{Re}(\boldsymbol{W}\boldsymbol{s}^h), \mathrm{Re}(\boldsymbol{W}\boldsymbol{s}^r), \mathrm{Re}(\boldsymbol{W}\boldsymbol{s}^t) > \\ &+ < \mathrm{Re}(\boldsymbol{W}\boldsymbol{s}^h), \mathrm{Im}(\boldsymbol{W}\boldsymbol{s}^r), \mathrm{Im}(\boldsymbol{W}\boldsymbol{s}^t) > \\ &+ < \mathrm{Im}(\boldsymbol{W}\boldsymbol{s}^h), \mathrm{Re}(\boldsymbol{W}\boldsymbol{s}^r), \mathrm{Im}(\boldsymbol{W}\boldsymbol{s}^t) > \\ &- < \mathrm{Im}(\boldsymbol{W}\boldsymbol{s}^h), \mathrm{Im}(\boldsymbol{W}\boldsymbol{s}^r), \mathrm{Re}(\boldsymbol{W}\boldsymbol{s}^t) >, \end{aligned} \tag{3}$$

where $\boldsymbol{s}^h \in \mathcal{S}^h$, $\boldsymbol{s}^r \in \mathcal{S}^r$, and $\boldsymbol{s}^t \in \mathcal{S}^t$, $\boldsymbol{W} \in$ is a weight matrix that converts the semantic features into complex vectors, $\overline{\boldsymbol{W}\boldsymbol{s}^t}$ is the complex conjugate of $\boldsymbol{W}\boldsymbol{s}^t$, and $< \cdot >$ denotes the multi-linear dot product of the vectors in the triplet. $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote the real and imaginary parts of a number, respectively. Note that the form of the triplet is ordered, and attribute vertices could only be the tail entities.

We select triplets with large responses on the criterion from the candidate proposals to generate the semantic graph. In analogy to non-maximum suppression (NMS), we eliminate triplets whose scores are lower than $-1$, and suppress triplets with more than one vertex which is the same with the local maxima.

**Relation reasoning.** We use the GCN (Johnson, Gupta, and Fei-Fei 2018) to propagate information along edges of the graph and contextually encode features in the semantic graph for generating relation-aware features.

As for the captioning, we introduce an attention mechanism to aggregate the triplets in the relation-aware graph for generating captions. Specifically, we adopt the captioning model of (Anderson et al. 2019), which is composed of a top-down attention LSTM for weighting visual features and a language LSTM for generating captions. The input to the top-down attention LSTM layer at time step $t$ is the concatenation of the previous hidden state $\boldsymbol{h}^2_{t-1}$ of the language LSTM layer, the global features $\boldsymbol{g}$, and the embedding vector $\boldsymbol{u}_{t-1}$ of the previously generated word. Thus, the hidden state of the top-down attention LSTM is given by

$$\boldsymbol{h}^1_t = \mathrm{LSTM}([\boldsymbol{h}^2_{t-1}, \boldsymbol{g}, \boldsymbol{u}_{t-1}], \boldsymbol{h}^1_{t-1}), \tag{4}$$

where $[\cdot, \cdot, \cdot]$ denotes the concatenation operation. Then we use $\boldsymbol{h}^1_t$ as the query of the attention operation to weight the triplets in relation-aware graph. The $g$-th triplet $\boldsymbol{t}_g$ in the graph is represented by the concatenation of the relation-aware features of the head, relationship, and tail entities. Supposing that there are $G$ triplets in the graph, the attention weight at time step $t$ is given by $\boldsymbol{\alpha}_t = [\alpha_{1,t}, \ldots, \alpha_{G,t}]$, where each $\alpha_{g,t}$ is calculated by fusing $\boldsymbol{h}^1_t$ and $\boldsymbol{t}_g$ after a normalization operation. The input to the language LSTM layer at time step $t$ can thus be obtained by concatenating $\sum_{g=1}^{G} \alpha_{g,t}\boldsymbol{t}_g$ with $\boldsymbol{h}^1_t$, and the output is the conditional distribution over the words.

**Objective.** Two losses are effectively combined to train the entire captioning model. One loss $L_c$ is a cross-entropy loss for generating sentences:

$$L_c = -\sum_{t=1}^{T} \log\big(Pr(y_t|y_{1:t-1}, \mathcal{I})\big), \tag{5}$$

where $Pr(y_t|y_{1:t-1}, \mathcal{I})$ denotes the probability that the prediction is the ground-truth word $y_t$ given the previous word sequence $y_{1:t-1}$ and all the features $\mathcal{I}$ of the input images or videos. Specifically, $\mathcal{I}$ includes the global features and candidate proposal features (visual feature vectors and knowledge vectors) of the input images or videos.

The other loss $L_s$ guides the learning of the semantic features of each vertex to capture correlation information with its adjacent vertices. $L_s$ is measured by the commonsense reasoning criteria when the semantic features are mapped into the complex domain:

$$L_s = \sum_{g=1}^{G}\sum_{t=1}^{T}(\alpha_{g,t} - \gamma)\log(1+ \\ \exp(-\mathrm{Re}(< \boldsymbol{W}\boldsymbol{s}_g^h, \boldsymbol{W}\boldsymbol{s}_g^r, \overline{\boldsymbol{W}\boldsymbol{s}_g^t} >))) + \lambda||\boldsymbol{W}||_2^2, \tag{6}$$

where the parameter $\lambda$ represents the importance of the regularization term, and $\gamma$ is a threshold that determines triplets to be punished. In the experiments, we set $\lambda = 0.01$ and $\gamma = 0.3$, empirically.

Consequently, the overall loss is defined as

$$L = L_c + \beta L_s, \tag{7}$$

where $\beta$ is a hyper-parameter. Since $L_s$ is constrained on the learning of attention weights $\{\boldsymbol{\alpha}_t | t = 1, \ldots, T\}$ guided by $L_c$, we set $\beta$ to 0 during the first few epochs of training, and 0.1 afterwards.

**Iterative algorithm.** Our C-R Reasoning method theoretically can be trained in an end-to-end manner. However, the commonsense reasoning module faces an optimization challenge: the construction of the semantic graph involves hard assignment operations, i.e., selecting triplets. To address this issue, we develop an iterative algorithm that alternates between semantic graph generation via commonsense reasoning and captioning via relation reasoning, as summarized in Algorithm 1.

# Experiments

## Datasets

We conduct experiments on a video captioning dataset, MSVD (Guadarrama et al. 2013), and an image captioning dataset, MSCOCO (Lin et al. 2014). The MSVD dataset comprises 1,970 video clips collected from Youtube, each annotated with roughly 40 captions. We follow the split in (Venugopalan et al. 2015a) which divides the videos into three parts: 1,200 training videos, 100 validation videos and 670 testing videos. The MSCOCO dataset contains above 100K images with 5 captions each. We follow the standard split by (Karpathy and Fei-Fei 2017) which takes 113,287 images for training, 5,000 for validation and 5,000 for testing.

We also conduct qualitative experiments on a Chinese video captioning dataset, Youku-VC, to further validate the effectiveness of our method on the task of video captioning in different languages. The Youku-VC dataset contains 1,430 short videos collected from Youku[2], and each video annotated with 10 Chinese sentences. We split the dataset into 1,000 training videos, 215 validation videos and 215 testing videos.

The metrics of BLEU-4 (B@4) (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014), CIDEr (Vedantam, Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016) are used for evaluations by the MSCOCO toolkit (Chen et al. 2015). For all the metrics, higher values indicate better performance.

**Algorithm 1:** C-R Reasoning.

```
Input: Visual feature vectors $\mathcal{V} = \cup_{n=1}^{N}\mathcal{V}_n$ and
       knowledge vectors $\mathcal{K} = \cup_{n=1}^{N}\mathcal{K}_n$ of $N$ images
       or videos.
Output: C-R Reasoning model.
Initialization: $\mathcal{H}_n = \mathcal{K}_n,\ \forall n = 1, \cdots, N$;
repeat
    • Semantic Graph Generation:
    for $n = 1, \cdots, N$ do
        Select object, relationship and attribute vertices
          from $(\mathcal{V}_n, \mathcal{K}_n)$ by using (3) on $\mathcal{H}_n$ to generate
          $(\mathcal{V}_n^S, \mathcal{K}_n^S)$;
    end
    $\mathcal{V}^S \Leftarrow \cup_{n=1}^{N}\mathcal{V}_n^S,\ \mathcal{K}^S \Leftarrow \cup_{n=1}^{N}\mathcal{K}_n^S$;
    Map $\mathcal{V}^S$ and $\mathcal{K}^S$ into semantic space $\varphi(\mathcal{V}^S, \mathcal{K}^S)$;
    • Captioning:
    Map $\varphi(\mathcal{V}^S, \mathcal{K}^S)$ into $\phi(\varphi(\mathcal{V}^S, \mathcal{K}^S))$ by using
      relation reasoning based on GCN;
    • Update:
    Update parameters of $\phi(\cdot)$, $\varphi(\cdot)$, and the
      sequence-based language model by minimizing $L$.
    $\mathcal{H}_n \Leftarrow \varphi(\mathcal{V}_n, \mathcal{K}_n),\ \forall n = 1, \cdots, N$;
until Convergence;
```

## Implementation Details

For video captioning, the visual feature vector of each sampled local region (i.e., 3D cuboid) is extracted by concatenating features after average pooling from the corresponding region in the feature map of the last convolutional layers of ResNeXt-101 (Xie et al. 2017) and IRv2 (Szegedy et al. 2017). The visual feature of each video frame is the concatenation of outputs of the pooling layer after the last convolutional layers of ResNeXt-101 and IRv2. The visual features of the entire video are derived from the res5c layer of ResNeXt-101 and the inception-c layer of IRv2. For image captioning, the visual feature vector of each sampled local region (i.e., 2D patch) is calculated after ROI pooling from the corresponding region in the feature map of the res5c layer of ResNet-101 (He et al. 2016). The visual feature of the entire image is the output of the pool5 layer of ResNet-101. For Chinese video captioning, the sentences are tokenized by jieba[3].

In visual mapping, to reduce computational cost, we employ the RPN (Ren et al. 2017) without NMS to densely sample candidate object regions with scores higher than threshold 0.7. For data augmentation, we repeatedly conduct k-means clustering operations to obtain multiple groups of candidate proposals from each image or video, and the number of clusters is set from 5 to 10. In knowledge mapping, the number of the sparse attention operations is set to 3 according to the mAP of the multi-label classification by the non-linear mapping networks on the validation set. In the

[2] https://www.youku.com

[3] https://github.com/fxsjy/jieba

| Methods | Detector | B@4 | METEOR | CIDEr |
|---|---|---|---|---|
| Gao et al. (2017) | | 50.8 | 33.3 | 74.8 |
| Gan et al. (2017) | | 51.1 | 33.5 | 77.7 |
| Wang et al. (2018b) | | 52.8 | 33.1 | - |
| Wang et al. (2018a) | | 52.3 | 34.1 | 80.3 |
| Aafaq et al. (2019) | ✓ | 47.8 | 35.0 | 78.1 |
| Zhang et al. (2019) | ✓ | 56.9 | 36.2 | 90.6 |
| Ours | | **57.0** | **36.8** | **96.8** |

Table 1: Comparison results on the MSVD dataset.

| Methods | Detector | B@4 | METEOR | CIDEr | SPICE |
|---|---|---|---|---|---|
| Gan et al. (2017) | | 33.0 | 25.7 | 101.2 | - |
| Anderson et al. (2019) | ✓ | 36.2 | 27.0 | 113.5 | 20.3 |
| Yao et al. (2018) | ✓ | 37.1 | 28.1 | 117.1 | 21.1 |
| Zhou et al. (2019) | ✓ | 29.9 | 25.6 | 107.2 | - |
| Li et al. (2019) | ✓ | 33.8 | 26.2 | 110.3 | 19.8 |
| Yang et al. (2019) | ✓ | 36.9 | 27.7 | 116.7 | 20.9 |
| Ours | | 36.7 | 28.1 | 117.3 | 20.1 |
| Ours with detector | ✓ | **37.7** | **28.2** | **120.1** | **21.6** |

Table 2: Comparison resuls on the MSCOCO dataset.

sequence-based language model, both the number of hidden units in each LSTM and the size of the input word embedding is set to 512. During training, the convergence criterion is considered as that the CIDEr score on the validation set stops increasing in 10 consecutive epochs. During inference, the sizes of beam search are set to 3 and 5 to generate sentences in image and video captioning, respectively.

## Comparison with the State-of-the-Art Methods

Table 1 shows the comparison results on the MSVD dataset. From the results, it is interesting to observe that: (1) Comparing with (Gao et al. 2017; Wang et al. 2018a; 2018b; Gan et al. 2017) which are simple sequence-to-sequence captioning models without exploiting their relations, our method achieves better performances, which proves the advantage of our joint commonsense and relation reasoning. (2) Our method outperforms (Aafaq et al. 2019; Zhang and Peng 2019) which detect objects from videos using detectors pre-trained on image dataset. It validates that using prior knowledge to identify objects in our method is more general than pre-training object detectors on images, since there exists a domain gap between the image and video datasets.

Table 2 shows the comparison results between our method and several recent methods that are closely related to our method on the MSCOCO dataset. All the compared methods (Gan et al., 2017; Anderson et al., 2019; Yao et al., 2018; Zhou et al.,2019; Li et al.,2019) use explicit high-level semantic concepts of objects and relationships for image captioning. We can have the following observations: (1) The fact that our method achieves better results than (Gan et al. 2017) where the semantic information is not exploited validates that C-R Reasoning can benefit learning semantic relationships for image captioning. (2) Compared with (Anderson et al., 2019; Yao et al., 2018; Zhou et al.,2019; Li et al.,2019) which use pre-trained detectors to explore visual relationships for captioning, our method still achieves comparable performances without any detectors, demonstrating that exploiting relationships actually benefits from prior knowledge and does not necessarily rely on pre-trained detectors. (3) For fair comparison, we also show the results of our method using the pre-trained Faster R-CNN detector to extract the initial regions from images. As shown in the bottom row of Table 2, our method with a detector outperforms all the compared methods.

| Methods | B@4 | METEOR | CIDEr |
|---|---|---|---|
| Anderson et al. (2019) | 48.7 | 33.2 | 83.4 |
| Ours w/o CR | 48.6 | 32.9 | 79.5 |
| Ours w/o RR | 54.9 | 36.7 | 92.4 |
| Ours | **57.0** | **36.8** | **96.8** |

Table 3: Results of ablation studies on the MSVD dataset.

## Ablation Study

To analyze our method in depth, ablation studies are conducted to evaluate the effect of each individual component and the results on the MSVD dataset are reported in Table 3.

**Effect of C-R Reasoning.** To analyze the effect of C-R Reasoning, we compare our method with the Up-Down method by (Anderson et al. 2019) that uses Faster R-CNN (Ren et al. 2015) to detect spatial regions and extracts visual features from the regions as input to a bottom-up attention model. For fair comparison, the visual features used in the Up-Down model are the same with ours. As can be seen from Table 3, our method achieves better results than Up-Down for all the metrics, which clearly validates that C-R Reasoning can significantly boost the performance.

**Effect of commonsense reasoning.** To analyze the effect of commonsense reasoning, we remove the commonsense reasoning, and instead, apply the Faster R-CNN to generate the semantic graph (i.e., "Ours w/o CR"). As shown in Table 3, the great improvement of our method over "Ours w/o CR" validates the importance of commonsense reasoning on learning the most relevant semantic concepts and relationships for captioning.

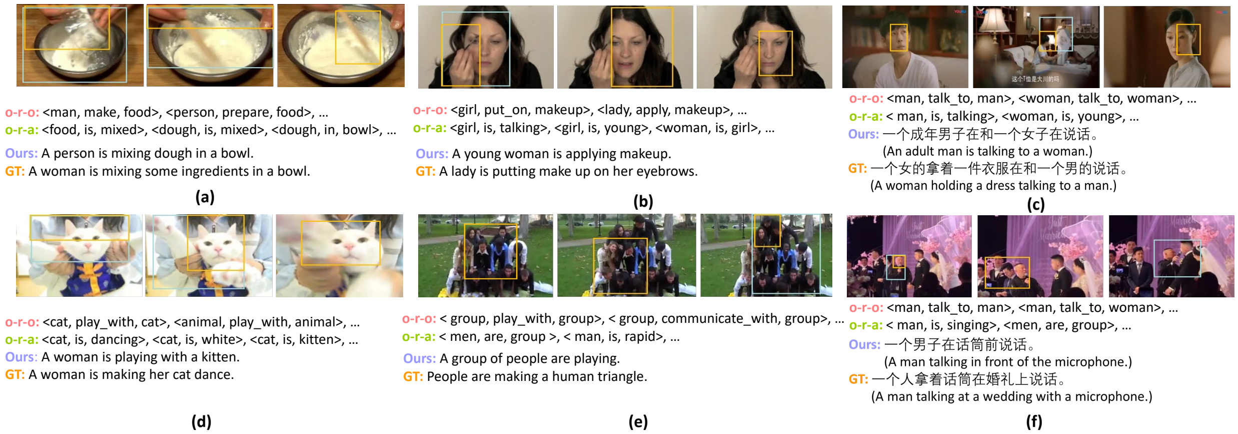
**Effect of relation reasoning.** To analyze the effect of relation reasoning, we remove the GCN (i.e., Ours w/o RR). As shown in Table 3, our method outperforms "ours w/o RR", verifying that learning relation-aware features based on semantic graphs is beneficial for improving the performance.

## Convergence Performance

For a more intuitive view of our iterative algorithm, we plot the learning curves of the CIDEr and B@4 scores on the test set of MSVD in Figure 5. Iteration 0 means that our model is trained without the loss $L_S$ at the beginning. As illustrated in Figure 5, our model converges after three iterations, and CIDEr drops afterwards because of overfitting.

## Qualitative Analysis

Figure 4 shows several exemplars of video captioning results on the MSVD dataset. For each exemplar, the top three images represent randomly sampled frames from the video.


o-r-o: <man, make, food>, <person, prepare, food>, ...
o-r-a: <food, is, mixed>, <dough, is, mixed>, <dough, in, bowl>, ...
Ours: A person is mixing dough in a bowl.
GT: A woman is mixing some ingredients in a bowl.

(a)

o-r-o: <girl, put_on, makeup>, <lady, apply, makeup>, ...
o-r-a: <girl, is, talking>, <girl, is, young>, <woman, is, girl>, ...
Ours: A young woman is applying makeup.
GT: A lady is putting make up on her eyebrows.

(b)

o-r-o: <man, talk_to, man>, <woman, talk_to, woman>, ...
o-r-a: < man, is, talking>, <woman, is, young>, ...
Ours: 一个成年男子在和一个女子在说话。
(An adult man is talking to a woman.)
GT: 一个女的拿着一件衣服在和一个男的说话。
(A woman holding a dress talking to a man.)

(c)

o-r-o: <cat, play_with, cat>, <animal, play_with, animal>, ...
o-r-a: <cat, is, dancing>, <cat, is, white>, <cat, is, kitten>, ...
Ours: A woman is playing with a kitten.
GT: A woman is making her cat dance.

(d)

o-r-o: < group, play_with, group>, < group, communicate_with, group>, ...
o-r-a: < men, are, group >, < man, is, rapid>, ...
Ours: A group of people are playing.
GT: People are making a human triangle.

(e)

o-r-o: <man, talk_to, man>, <man, talk_to, woman>, ...
o-r-a: < man, is, singing>, <men, are, group>, ...
Ours: 一个男子在话筒前说话。
(A man talking in front of the microphone.)
GT: 一个人拿着话筒在婚礼上说话。
(A man talking at a wedding with a microphone.)

(f)

Figure 4: Qualitative results on MSVD and Youku-VC. Yellow and blue bounding boxes represent the candidate proposals of objects and relations (if any), respectively. The "o-r-o" and "o-r-a" denote the typical triplets of "object-relation-object" and "object-relation-attribute" in semantic graphs, respectively. "Ours" represents the captions generated by our method, and "GT" represents one of the ground-truth sentences. Sentences in parentheses are the translations of the Chinese captions.

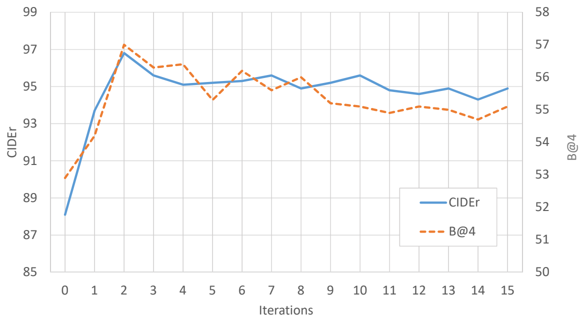


Figure 5: Learning curves of the CIDEr and B@4 scores.

The bounding boxes in images indicate the inferred candidate proposals. Below the images, we show some typical triplets of "object-relationship-object (o-r-o)" and "object-relationship-attribute (o-r-a)" that are generated from the knowledge graphs. Our captioning results and the ground-truth (GT) sentences are shown at the bottom. It is interesting to observe that our method can detect some "difficult" objects for generating accurate captions. For example, as shown in Figure 4(b) and (d), the tiny-size object of "makeup" and the heavily occluded "person" are successfully inferred by referring to the prior knowledge of <woman, put_on, makeup> and <woman, play_with, cat>, respectively. We also show examples of the results on the Youku-VC dataset in Figure 4(c) and (f).

## Conclusion

We have presented a novel joint commonsense and relation reasoning approach to image and video captioning by exploiting prior knowledge, which alternates between commonsense reasoning to build a semantic graph and relation reasoning to generate textual descriptions. It can learn semantic relationships between objects to comprehensively understand the visual cues, and generate sentences that accurately describe the image content, without requiring any predefined object or relationship detectors. Thanks to the joint learning strategy, our captioning model is able to achieve the global semantic coherency within an image or a video, thus further improves the captioning performance. Experiments on both image and video captioning benchmarks demonstrate that our method outperforms the state-of-the-art methods. In the future, we will exploit more prior knowledge for commonsense reasoning and incorporate motion information into relation reasoning for video captioning.

**Acknowledgments** This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant No.61673062 and Alibaba Group through Alibaba Innovative Research Program.

## References

Aafaq, N.; Akhtar, N.; Liu, W.; Gilani, S. Z.; and Mian, A. 2019. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 12487–12496.

Aditya, S.; Yang, Y.; Baral, C.; Aloimonos, Y.; and Fermüller, C. 2018. Image understanding using vision and reasoning through scene description graph. *Comput. Vision and Image Understanding* 173:33–45.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: semantic propositional image caption evaluation. In *Proc. Eur. Conf. Comput. Vision*, 382–398.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2019. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 6077–6086.

Berant, J.; Chou, A.; Frostig, R.; and Liang, P. 2013. Semantic parsing on freebase from question-answer pairs. In *Empirical Methods in Natural Language Processing*, 1533–1544.

Chen, X.; Fang, H.; Lin, T.; Vedantam, R.; Gupta, S.; Dollár, P.; and

Zitnick, C. L. 2015. Microsoft COCO captions: Data collection and evaluation server. *Arxiv*.

Denkowski, M. J., and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Association for Computational Linguistics*, 376–380.

Donahue, J.; Hendricks, L. A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Darrell, T.; and Saenko, K. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2625–2634.

Fader, A.; Zettlemoyer, L.; and Etzioni, O. 2014. Open question answering over curated and extracted knowledge bases. In *Knowledge Discovery and Data Mining*, 1156–1165.

Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; and Deng, L. 2017. Semantic compositional networks for visual captioning. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 1141–1150.

Gao, L.; Guo, Z.; Zhang, H.; Xu, X.; and Shen, H. T. 2017. Video captioning with attention-based LSTM and semantic consistency. *IEEE Trans. Multimedia* 19(9):2045–2055.

Guadarrama, S.; Krishnamoorthy, N.; Malkarnenkar, G.; Venugopalan, S.; Mooney, R. J.; Darrell, T.; and Saenko, K. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proc. IEEE Conf. Comput. Vision,*, 2712–2719.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 770–778.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 1219–1228.

Karpathy, A., and Fei-Fei, L. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(4):664–676.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision* 123(1):32–73.

Li, X., and Jiang, S. 2019. Know more say less: Image captioning based on scene graphs. *IEEE Trans. Multimedia* 21(8):2117–2130.

Liang, X.; Lee, L.; and Xing, E. P. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 4408–4417.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comput. Vision*, 740–755.

Lu, C.; Krishna, R.; Bernstein, M. S.; and Li, F. 2016. Visual relationship detection with language priors. In *Proc. Eur. Conf. Comput. Vision*, 852–869.

Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Proc. Int.Conf. Learn. Represent.*

Martins, A. F. T., and Astudillo, R. F. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proc. Int. Conf. Mach. Learn.*, 1614–1623.

Pan, P.; Xu, Z.; Yang, Y.; Wu, F.; and Zhuang, Y. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 1029–1038.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics*, 311–318.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proc. Adv. Neural Inf. Process*, 91–99.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(6):1137–1149.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object hallucination in image captioning. In *Empirical Methods in Natural Language Processing*, 4035–4045.

Su, Z.; Zhu, C.; Dong, Y.; Cai, D.; Chen, Y.; and Li, J. 2018. Learning visual knowledge memory networks for visual question answering. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 7736–7745.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proc. AAAI Conf. Artif. Intell.*, 4278–4284.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; and Bouchard, G. 2016. Complex embeddings for simple link prediction. In *Proc. Int. Conf. Mach. Learn.*, 2071–2080.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 4566–4575.

Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R. J.; Darrell, T.; and Saenko, K. 2015a. Sequence to sequence - video to text. In *Proc. IEEE Conf. Comput. Vision,*, 4534–4542.

Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R. J.; and Saenko, K. 2015b. Translating videos to natural language using deep recurrent neural networks. In *NAACL-HLT*, 1494–1504.

Wang, B.; Ma, L.; Zhang, W.; and Liu, W. 2018a. Reconstruction network for video captioning. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 7622–7631.

Wang, J.; Wang, W.; Huang, Y.; Wang, L.; and Tan, T. 2018b. M3: multimodal memory modelling or video captioning. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 7512–7520.

Xie, S.; Girshick, R. B.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 5987–5995.

Xue, N. 2011. Natural language processing with python. *Natural Language Engineering* 17(3):419–424.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 10685–10694.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *Proc. Eur. Conf. Comput. Vision*, 684–699.

Zhang, J., and Peng, Y. 2019. Object-aware aggregation with bidirectional temporal graph for video captioning. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 8327–8336.

Zhang, H.; Kyaw, Z.; Chang, S.; and Chua, T. 2017. Visual translation embedding network for visual relation detection. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 3107–3115.

Zhou, Y.; Sun, Y.; and Honavar, V. G. 2019. Improving image captioning by leveraging knowledge graphs. In *Winter Conference on Applications of Computer Vision*, 283–293.