Learning to Discretely Compose Reasoning Module Networks for Video Captioning

Ganchao Tan[1] , **Daqing Liu**[1*] , **Meng Wang**[2] and **Zheng-Jun Zha**[1†]

[1]University of Science and Technology of China

[2]Hefei University of Technology

{tgc1997, liudq}@mail.ustc.edu.cn, eric.mengwang@gmail.com, zhazj@ustc.edu.cn

Abstract

Generating natural language descriptions for videos, *i.e.*, video captioning, essentially requires step-by-step reasoning along the generation process. For example, to generate the sentence "a man is shooting a basketball", we need to first locate and describe the subject "man", next reason out the man is "shooting", then describe the object "basketball" of shooting. However, existing visual reasoning methods designed for visual question answering are not appropriate to video captioning, for it requires more complex visual reasoning on videos over both space and time, and dynamic module composition along the generation process. In this paper, we propose a novel visual reasoning approach for video captioning, named Reasoning Module Networks (RMN), to equip the existing encoder-decoder framework with the above reasoning capacity. Specifically, our RMN employs 1) three sophisticated spatio-temporal reasoning modules, and 2) a dynamic and discrete module selector trained by a linguistic loss with a Gumbel approximation. Extensive experiments on MSVD and MSR-VTT datasets demonstrate the proposed RMN outperforms the state-of-the-art methods while providing an explicit and explainable generation process. Our code is available at https://github.com/tgc1997/RMN.

1 Introduction

Video captioning, the task aims to automatically generate natural language descriptions for videos, has received increasing attention in computer vision and machine learning. Even though our community achieves the significant advance in visual recognition [He *et al.*, 2016; Ren *et al.*, 2016] and natural language understanding [Bahdanau *et al.*, 2015], video captioning is still a very challenging task and far away from satisfactory for it not only requires a thorough understanding of the input videos, but also requires step-by-step visual reasoning along the generation process.

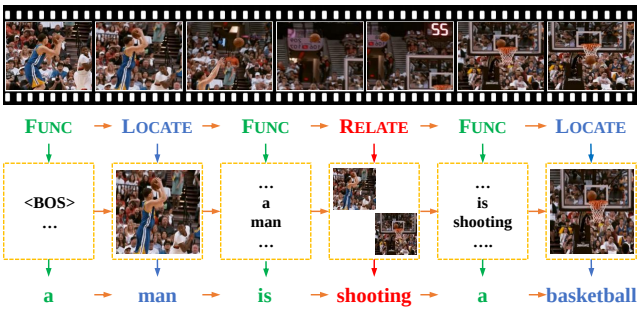

Figure 1: The caption generation process of the proposed Reasoning Module Networks (RMN). At each step, RMN first makes a dynamic and discrete selection from three fundamental reasoning modules, *i.e.*, LOCATE, RELATE, and FUNC, and then executes the corresponding reasoning module to generate the word. Specifically, LOCATE module locates one region to generate the visual words, RELATE module relates pairwise regions over both space and time to generate action words, FUNC module generates those function words according to the language context.

As illustrated in Figure 1, to describe the video, we human may have the following reasoning process: 1) identifying the subject to describe of the video, *i.e.*, "man", 2) inferring what is the man doing, *i.e.*, "shooting", 3) identifying what is the object of "shooting", *i.e.*, "basketball", 4) generating the final description "a man is shooting a basketball" by inserting several function words. In a nutshell, there are three fundamental reasoning mechanisms involved: 1) Locate one region to generate visual words, 2) Relate pairwise regions to generate action words, 3) Generate function words according to the language context.

Most existing video captioning methods [Venugopalan *et al.*, 2015; Donahue *et al.*, 2015] follow the encoder-decoder framework, where a CNN is employed as an encoder to produce the video features and an RNN is employed as a decoder to generate the captions. Those methods usually neglect the nature of the above human-level reasoning, thus hurting the explainability of the generation process. Even though there are some recent works have explored the visual reasoning in visual question answering [Andreas *et al.*, 2016; Hu *et al.*, 2017; Yang *et al.*, 2019a] and visual grounding [Cirik *et al.*, 2018; Liu *et al.*, 2019; Hong *et al.*, 2019] by decomposing the questions or referring expressions into

*Equal Contribution

†Corresponding Author

a linear or tree reasoning structure with several neural modules, the situation in the video captioning is more challenging because 1) unlike still images, videos contain richer visual content thus requiring more complex visual reasoning over both space and time, 2) unlike questions or referring expressions which are given in advance, the video descriptions are not available during the inference. Therefore, the model must dynamically compose the reasoning structure along the generation process.

To tackle the above two challenges, we propose a novel video captioning framework named Reasoning Module Networks (RMN). Firstly, to perform visual reasoning over both space and time, our RMN employs three fundamental spatio-temporal reasoning modules: a) LOCATE module to locate one single region over the video by a spatial-temporal attention, thus generating the visual words, *e.g.*, "man" and "basketball" in Figure 1; b) RELATE module to relate pairwise regions over the video by first detecting the object of each frame, and then modeling the action by pairing two frames, thus generating the action words, *e.g.*, "shooting" in Figure 1; and c) FUNC module to generate the function words according to the language context, *e.g.*, "a" and "is" in Figure 1. Secondly, to compose the reasoning structure along the generation process, our RMN employs a dynamic and discrete module selector. To jointly train both the three modules and the selector in an end-to-end manner, we adopt the recently proposed Gumbel approximation [Jang *et al.*, 2017] to make the discrete sampling process differentiable, and then constraint the module selector with a linguistic loss of part-of-speech (POS) tag labels.

We validate the effectiveness of the proposed RMN by conducting extensive experiments on two widely-used datasets MSVD [Chen and Dolan, 2011] and MSR-VTT [Xu *et al.*, 2016]. RMN outperforms the state-of-the-art video captioning methods on most metrics. Qualitative results indicate the generation process is explicit and explainable.

Our main contributions are three-fold: 1) We propose a novel framework named reasoning module networks (RMN) for video captioning with three *spatio-temporal* visual reasoning modules; 2) We adopt a *discrete* module selector to *dynamically* compose the reasoning process with modules; 3) Our RMN achieves new state-of-the-art performance with an explicit and explainable generation process.

2 Related Work

2.1 Video Captioning

There are two main directions to solve the video captioning problem. In the early stage, template-based methods [Kojima *et al.*, 2002; Guadarrama *et al.*, 2013], which first define a sentence template with grammar rules and then aligned subject, verb and object of the sentence template with video content, were widely studied. Those methods are hard to generate flexible language due to the fixed syntactic structure of the predefined template. Benefit from the rapid development of deep neural networks, the sequence learning methods [Venugopalan *et al.*, 2015; Yao *et al.*, 2015; Pan *et al.*, 2017] are widely used to describe the video with flexible natural language, most of these methods are based on the encoder-decoder framework. [Venugopalan *et al.*, 2015] proposed S2VT model which regards the video captioning task as a machine translation task. [Yao *et al.*, 2015] introduced a temporal attention mechanism to assign weights to the features of each frame and then fused them based on the attention weights. [Li *et al.*, 2017; Chen and Jiang, 2019] further applied spatial attention mechanisms on each frame.

Recently, [Wang *et al.*, 2019] and [Hou *et al.*, 2019] proposed to leverage Part-of-Speech (POS) tags to boost video captioning. [Wang *et al.*, 2019] encodes the predicted POS sequences into hidden features, which further guides the generation process. [Hou *et al.*, 2019] mixes word probabilities of multiple components at each timestep conditioned on the inferred POS tags. However, both of them lack the reasoning capability for rich video content. On the contrary, we propose three well-designed reasoning module networks that correspond to three fundamental reasoning mechanisms.

2.2 Neural Module Networks

Neural module networks is a general framework that explicitly models the compositionality of languages by decomposing the network into neural modules. It has been widely used in visual question answering and visual grounding. [Andreas *et al.*, 2016] employs an off-the-shelf parser to parse the questions into a tree, leading to brittleness caused by parsing errors. [Hu *et al.*, 2017] trains an RNN to decode the language into the sequence, thus requiring extra human annotations. [Hu *et al.*, 2018] removes language parser and additional annotations by using a soft and continuous module layout. However, the case of video captioning is more complex since there are no fully observed captions during inference thus all the above methods are not applicable. To this end, we design a dynamic module selector to construct the reasoning procedure step-by-step during the generation process.

Recently, the pioneering works [Liu *et al.*, 2018; Zha *et al.*, 2019; Yang *et al.*, 2019b; Tian and Oh, 2019] are trying to adopt neural module networks into image captioning. However, their modules are designed to produce several types of features and their module compositions relied on a soft attention mechanism. As a contrast, our RMN employs several sophisticated spatio-temporal reasoning modules to perform more complex visual reasoning over videos, and designs a discrete and dynamic module selector to make a selection at each step, making the generation process explicit and explainable.

3 Approach

In this section, we will describe the proposed Reasoning Module Networks (RMN) in more detail. Figure 2 gives a walk-through example of our RMN at timestep t. Our RMN can be divided into four stages: Encoding (Section 3.1), Module Reasoning (Section 3.2), Module Selection (Section 3.3), and Decoding (Section 3.4). In Section 3.5, we detail the joint training strategy for both reasoning modules and the module selector.

3.1 Encoder

For the given video of N frames, we first represent it with three types of features, *i.e.*, appearance features $\boldsymbol{V_a}$ extracted

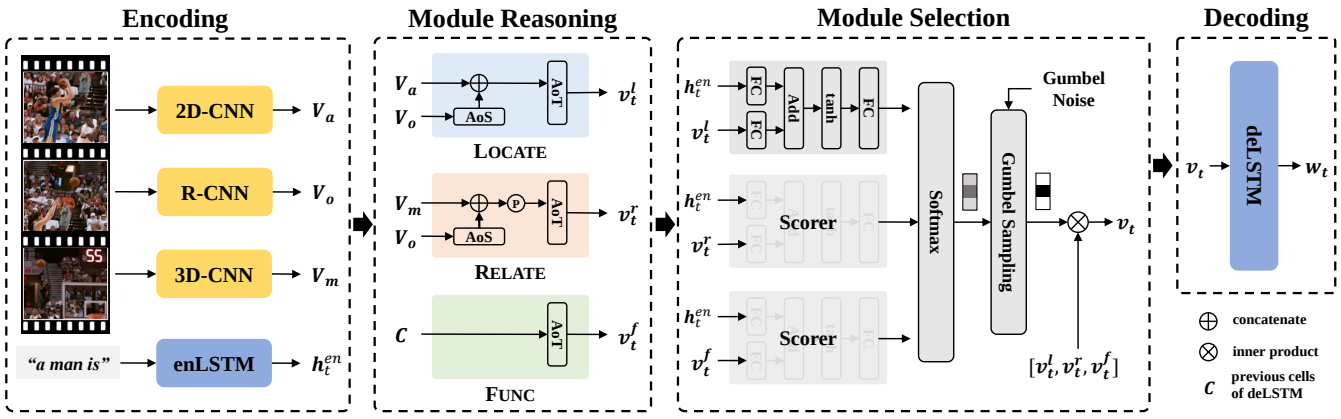

Figure 2: The overview of our proposed Reasoning Module Networks (RMN) which consists of four stages. At the encoding stage (cf. Section 3.1), we represent the given video by several visual features and encode previously generated words into the hidden state. At the module reasoning stage (cf. Section 3.2), we perform three fundamental visual reasoning over both space and time by LOCATE, RELATE, and FUNC modules. At the module selection stage (cf. Section 3.3), we dynamically and discretely select one determined reasoning module to produce the final reasoning result. At the decoding stage (cf. Section 3.4), we decode the reasoning result into word.

from a 2D-CNN, object features $\boldsymbol{V_o}$ extracted from a R-CNN on each frame, and motion features $\boldsymbol{V_m}$ extracted from a 3D-CNN. Note that to model temporal information for those visual features, we have post-processed $\boldsymbol{V_a}$ and $\boldsymbol{V_m}$ with Bi-LSTMs. And $\boldsymbol{V_o}$ contains an additional dimension on space.

For the previous generated words $\{w_1, \cdots, w_{t-1}\}$, we encode them by a LSTM (denoted as enLSTM) which takes the global visual feature $\bar{\boldsymbol{v}}$, the last generated word embedding vector $\boldsymbol{e}_{t-1}$, and the last timestep hidden state $\boldsymbol{h}^{de}_{t-1}$ of deLSTM (cf. Eqn. (9)) as input:

$$\boldsymbol{h}^{en}_t = \textbf{enLSTM}(\bar{\boldsymbol{v}}, \boldsymbol{e}_{t-1}, \boldsymbol{h}^{de}_{t-1}). \tag{1}$$

It is worth noticing that since $\boldsymbol{h}^{en}_t$ encodes rich information from history, it will be used to guide the reasoning process and the module selection.

3.2 Reasoning Modules

As mentioned in the Introduction, there are three fundamental reasoning mechanisms involved in video captioning. Therefore, we design three corresponding reasoning modules. In the following, we will first introduce the attention functions used in the modules, and then describe each module in detail.

Attention Functions

We follow the widely used additive attention formulation [Bahdanau *et al.*, 2015]:

$$\mathbf{A}(\boldsymbol{V}, \boldsymbol{q}) = \text{softmax}(\boldsymbol{w}_1^T \tanh(\boldsymbol{W}_2\boldsymbol{V} + \boldsymbol{W}_3\mathbb{1}^T\boldsymbol{q}))\boldsymbol{V}, \tag{2}$$

where $\boldsymbol{V}$ and $\boldsymbol{q}$ are values and queries of attention, $\boldsymbol{w}_1$, $\boldsymbol{W}_2$, and $\boldsymbol{W}_3$ are trainable parameters, $\mathbb{1}$ is an all-one vector. Note that, the attention can be executed on any dimensions, therefore, we further define $\textbf{AoS}(\cdot)$ as the Attention over the dimension on Space, and define $\textbf{AoT}(\cdot)$ as the Attention over the dimension on Time.

LOCATE Module

It is designed to generate visual words, *e.g.*, "man" and "basketball". Generating this type of words requires the model to attend on one specific region over both space and time. Therefore, we first apply an Attention over Space (AoS) for object features $\boldsymbol{V_o}$ and then apply an Attention over Time (AoT) together with $\boldsymbol{V_a}$, formally:

$$\boldsymbol{v}^l_t = \textbf{AoT}(\textbf{AoS}(\boldsymbol{V_o}, \boldsymbol{h}^{en}_t) \oplus \boldsymbol{V_a}, \boldsymbol{h}^{en}_t), \tag{3}$$

where $\oplus$ denotes concatenate operation.

RELATE Module

It is designed to generate action words, *e.g.*, "riding" and "shooting". Modeling action requires reasoning over both space and time. Take Figure 1 as an example, to generate the word "shooting", we must be aware that the man is holding a basketball, and then notice that the basketball is flying to the basket, finally we can inference that the man is "shooting". It means we must connect two scenes at different time. To achieve this reasoning mechanism, we insert a pairwise interaction (*i.e.*, concatenate every possible pair of two tensors) between the AoS and AoT, formally:

$$\begin{aligned} \boldsymbol{v}^r_t &= \textbf{AoT}(\mathbf{P}(\boldsymbol{M}, \boldsymbol{M}), \boldsymbol{h}^{en}_t), \\ \boldsymbol{M} &= \textbf{AoS}(\boldsymbol{V_o}, \boldsymbol{h}^{en}_t) \oplus \boldsymbol{V_m}, \end{aligned} \tag{4}$$

where $\mathbf{P}(\cdot, \cdot)$ denotes the pairwise interaction function that $\mathbf{P}_{ij}(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{A}_i \oplus \boldsymbol{B}_j$.

FUNC Module

It is designed to generate function words to complete the whole sentences. Since the function words only require language information, thus we propose to recall the history decoder cell states of deLSTM (cf. Eqn. (9)) to generate the current words by an AoT, formally:

$$\begin{aligned} \boldsymbol{v}^f_t &= \textbf{AoT}(\boldsymbol{C}, \boldsymbol{h}^{en}_t), \\ \boldsymbol{C} &= [\boldsymbol{c}^{de}_1, \cdots, \boldsymbol{c}^{de}_{t-1}]. \end{aligned} \tag{5}$$

So far, we have detailed the three proposed reasoning modules. Next, we will describe how to make a selection of those modules at each timestep.

3.3 Module Selector

As illustrated in Figure 2, the module selection consists of two steps. First, we calculate a score for each module to measure the probability of each module could be selected. Second, we sampling one determined module based on those scores with Gumbel Softmax [Jang *et al.*, 2017] strategy.

In detail, we formulate the scoring function $\mathbf{S}(\cdot,\cdot)$ as:

$$\mathbf{S}(\boldsymbol{h},\boldsymbol{v}) = \mathrm{fc}(\tanh(\mathrm{fc}(\boldsymbol{h}) + \mathrm{fc}(\boldsymbol{v}))). \tag{6}$$

Thus we can get $s_t^l = \mathbf{S}(\boldsymbol{h}_t^{en}, \boldsymbol{v}_t^l)$, $s_t^r = \mathbf{S}(\boldsymbol{h}_t^{en}, \boldsymbol{v}_t^r)$, and $s_t^f = \mathbf{S}(\boldsymbol{h}_t^{en}, \boldsymbol{v}_t^f)$ for the three LOCATE, RELATE, and FUNC modules, respectively.

After that, the straightforward selection method is directly choosing the module with the maximum score. However, it will block the gradients off the module selector for the function $\arg\max$ is non-differentiable. To this end, we apply a relaxation between forward pass and backward pass, formally:

$$\begin{aligned} &\text{Forward}: \boldsymbol{z}_t = \arg\max(\log([s_t^l, s_t^r, s_t^f]) + G), \\ &\text{Backward}: \tilde{\boldsymbol{z}}_t = \mathrm{softmax}((\log([s_t^l, s_t^r, s_t^f]) + G)/\tau), \end{aligned} \tag{7}$$

where we take an approximation that relaxes the discrete one-hot decision vector $\boldsymbol{z}_t$ into a continuous $\tilde{\boldsymbol{z}}_t$. G is the Gumbel noise drawn from i.i.d. Gumbel(0, 1). τ is a temperature parameter to control the strength of softmax. Please refer to [Jang *et al.*, 2017] for more mathematical details.

With this decision vector $\boldsymbol{z}_t$, we can get the final visual reasoning result as:

$$\boldsymbol{v}_t = \boldsymbol{z}_t \otimes [\boldsymbol{v}_t^l, \boldsymbol{v}_t^r, \boldsymbol{v}_t^f], \tag{8}$$

where $\otimes$ denotes inner product. It is worth noticing that $\boldsymbol{z}_t$ is an one-hot vector, thus only one determined module will be selected in the forward pass.

3.4 Decoder

For each timestep t, we decode the final visual reasoning result by a LSTM (denotes as deLSTM):

$$\boldsymbol{h}_t^{de}, \boldsymbol{c}_t^{de} = \mathbf{deLSTM}(\boldsymbol{v}_t, \boldsymbol{h}_t^{en}). \tag{9}$$

Recall that the hidden and cell states were used in Eqn. (1) and Eqn. (5). Further, we decode the word probability $\boldsymbol{p}_t$ as:

$$\boldsymbol{p}_t = \mathrm{softmax}(\mathbf{MLP}(\boldsymbol{v}_t \oplus \boldsymbol{h}_t^{en} \oplus \boldsymbol{h}_t^{de})), \tag{10}$$

where **MLP** is two-layers with $\tanh$ as activation function.

3.5 End-to-End Training

Thanks to the Gumbel-Softmax strategy, our RMN model can be trained in an end-to-end manner. Given a video and the corresponding ground-truth caption $\{w_t^*\}$ for $t \in [1, T]$, where T is the sentence length, the objective is to minimize the cross-entropy loss function given by:

$$\mathcal{L}_{cap} = -\sum_{t=1}^{T} \log \boldsymbol{P}_t(w_t^*). \tag{11}$$

Further, to ensure the module selector sticking on the linguistic structure, we propose to supervise the module selection by extra Part-of-Speech (POS) tag labels. Accordingly, we apply a Kullback-Leibler Divergence (KLD) loss to make the predicted module decision vector $\tilde{\boldsymbol{z}}$ and the ground-truth tag distribution $S^* = \{s_t^*\}$, $t \in [1, T]$ as close as possible, formally:

$$\mathcal{L}_{pos} = -\sum_{t=1}^{T} \mathrm{KLD}(\tilde{\boldsymbol{z}}_t \| \texttt{one\_hot}(s_t^*)). \tag{12}$$

To collect the ground-truth S^*, we first detected POS tag labels by Spacy Tagging Tool[1], and then assigned [NN*] and [JJ*] to the LOCATE module, [VB*] to the RELATE module, and the other to the FUNC module.

Therefore, the overall loss function of our RMN model is given by:

$$\mathcal{L} = \mathcal{L}_{cap} + \lambda \mathcal{L}_{pos}, \tag{13}$$

where λ is a trade-off weight.

4 Experiments

4.1 Datasets and Metrics

We conduct experiments on two widely used video captioning datasets with several standard evaluation metrics to verify the effectiveness of our proposed method.

Datasets

MSVD. The MSVD dataset [Chen and Dolan, 2011] consists of 1,970 short video clips selected from Youtube, where each one depicts a single activity in the open domain, and each video clip is annotated with multi-lingual captions. Since we only consider the English captions in this paper, each video clip has roughly 41 descriptions. To be consistent with previous works, we split the dataset to 3 subsets, 1,200 clips for training, 100 clips for validation, and the remaining 670 clips for testing.

MSR-VTT. The MSR-VTT [Xu *et al.*, 2016] is a large-scale dataset for the open domain video captioning, which consists of 10,000 video clips from 20 categories, and each video clip is annotated with 20 English sentences by Amazon Mechanical Turks. There are about 29,000 unique words in all captions. Following the existing works, we use the standard splits, namely 6,513 clips for training, 497 clips for validation, and 2,990 clips for testing.

Evaluation Metrics

We use several widely used automatic evaluation metrics to evaluate the quality of the generated captions, *i.e.*, BLEU [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005], CIDEr [Vedantam *et al.*, 2015], ROUGE-L [Lin, 2004]. Most of these metrics are originally proposed for machine translation or image captioning, the higher score indicates better quality of the captions.

4.2 Implementation Details

Dataset Preprocessing

We first convert all captions to lower case and remove punctuations, then we truncate the captions with more than 26 words and zero pad the captions with less than 26 words. The vocabulary size is set to 7,351 for MSVD and 9,732 for MSR-VTT with removing the words appear less than twice and five times respectively.

[1] https://spacy.io

Settings						MSVD				MSR-VTT			
Model	LOCATE	RELATE	FUNC	Discrete	$\mathcal{L}_{pos}$	B@4	R	M	C	B@4	R	M	C
RMN (LOCATE)	✓					52.5	73.1	35.8	90.0	40.7	60.5	27.7	46.9
RMN (RELATE)		✓				52.8	73.1	36.1	90.0	40.1	60.6	28.1	47.1
RMN (S)	✓	✓	✓			53.2	73.2	35.8	90.8	41.0	60.6	28.0	47.2
RMN (H)	✓	✓	✓	✓		51.5	72.0	35.1	88.4	41.9	60.9	28.1	48.1
RMN (S+L)	✓	✓	✓		✓	52.5	72.7	36.1	92.8	42.1	61.3	28.3	49.1
RMN (H+L)	✓	✓	✓	✓	✓	**54.6**	**73.4**	**36.5**	**94.4**	**42.5**	**61.6**	**28.4**	**49.6**

Table 1: The performance of ablated models with various settings on MSVD and MSR-VTT datasets. B@4, R, M, C denote BLEU-4, ROUGE_L, METEOR, CIDEr, respectively.

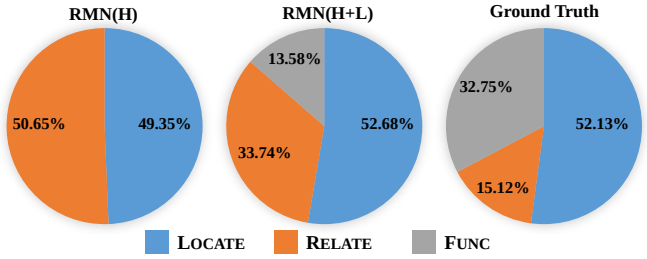

Figure 3: The proportion of each module occupied in RMN (H), RMN (H+L), and the Groud-Truth on test set of MSVD.

Figure 4: Word cloud visualizations of words generated by each module. We can find that the LOCATE module mainly generate visual words, the RELATE module mainly generate action words, and the FUNC module is likely to generate function words. Experiments conducted on MSVD of RMN (H+L).

Feature Extraction

In our experiments, we use InceptionResNetV2 (IRV2) [Szegedy *et al.*, 2017] as 2D CNN and I3D [Carreira and Zisserman, 2017] as 3D CNN to extract appearance features and motion features respectively, then we equally-spaced 26 features for each video. The IRV2 is trained on ILSVRC-2012-CLS image classification dataset [Russakovsky *et al.*, 2015] and the I3D is trained on Kinetics action classification dataset [Kay *et al.*, 2017]. We adopt Faster-RCNN [Ren *et al.*, 2016] which is trained by [Anderson *et al.*, 2018] to extract 36 region features for each frame (26 equally-spaced frames for each video).

Training Details

Our model is optimized by Adam Optimizer [Kingma and Ba, 2015], the initial learning rate is set to 1e-4. For the MSVD dataset, the hidden size of the LSTM is 512 and the learning rate is divided by 10 every 10 epochs. For the MSR-VTT dataset, the hidden size of the LSTM is 1,300 and the learning rate is divided by 3 every 5 epochs. During testing, we use beam search with size 2 for the final caption generation.

4.3 Ablation Study

As shown in Table 1, we compare our RMN against a set of other ablated models with various settings: (1) **RMN (LOCATE)**: the model that only deployed with Locate module; (2) **RMN (RELATE)**: the model that only deployed with Related module; (3) **RMN (S)**: the model that softly fuses three modules, *i.e.*, z_t in Eqn. (8) is computed as $\mathrm{softmax}(\log([s_t^l, s_t^r, s_t^f])$; (4) **RMN (H)**: the model that hard selects one of three modules; (5) **RMN (S+L)**: the model that softly fuses three modules and trained by linguistic loss; (6) **RMN (H+L)**: the model that hard selects one of three modules and trained by linguistic loss. According to the results shown in Table 1, we have the following observations.

Effect of Reasoning Modules

RMN (RELATE) consistently outperforms RMN (LOCATE) on both MSVD and MSR-VTT. This is because the RELATE module models much richer visual action information than LOCATE, leading to performance improvement over RMN (LOCATE).

RMN (S) outperforms RMN (LOCATE) and RMN (RELATE) under most metrics. This indicates that by fusing different types of reasoning mechanisms, the model can generate better video descriptions.

Effect of Gumbel Strategy

Comparing RMN (H) with RMN (LOCATE) and RMN (RELATE), we observe opposite results on MSVD and MSR-VTT. To figure out the reasons for this observation, we count the proportion of the three modules occupied in Figure 3. We can find that without the linguistic loss, RMN (H) only employs two modules, *i.e.*, LOCATE module and RELATE module, indicating the failure of training.

Therefore, the reason for the opposite results is that RMN (H) failed in training on MSVD while succeeded on MSR-VTT. Actually, even though Gumbel strategy can relax the discrete decision to continuous decision, it still requires large training examples. Therefore, RMN (H) underperforms RMN (LOCATE) and RMN (RELATE) on MSVD which only contains 1,200 training samples, while outperforming them on MSR-VTT which contains 6,513 training samples.

Similarly, even though RMN(S) surpasses RMN(H) under some metrics, but if we employ the linguistic loss to make

GT: there is a man riding a vehicle into the forest
RMN: a man is riding a vehicle in a forest

(a) MSR-VTT: 7393

GT: a group of boys are playing basketball with a coach
RMN: a group of people are playing basketball

(b) MSR-VTT: 7458

GT: there is a man in blue is swimming in the sea
RMN: a man is swimming in the sea and talking about the sea

(c) MSR-VTT: 7505

GT: a man is playing a flute
RMN: a man is playing a flute

(d) MSVD: z0zb--BOhDY_16_22

GT: a man and a woman are speaking on the phone
RMN: two people are talking on the phone

(e) MSVD: q3I3R_gqy8M_34_37

GT: someone is pouring tomato sauce into a pot of meat
RMN: a man is pouring sauce into a pot

(f) MSVD: hJFBXHtxKIc_204_209

Figure 5: Visualization of some video captioning examples on MSVD and MSR-VTT (better view in color). The first line in each example is one of the ground truth captions and the second line is generated by our RMN. Word in blue, red, green color denotes it is generated by LOCATE, RELATE, FUNC, respectively.

	MSVD			
Models	B@4	R	M	C
MAM-RNN [Li *et al.*, 2017]	41.3	68.8	32.2	53.9
RecNet [Wang *et al.*, 2018]	52.3	69.8	34.1	80.3
MARN [Pei *et al.*, 2019]	48.4	71.9	35.1	92.2
OA-BTG [Zhang and Peng, 2019]	**56.9**	-	36.2	90.6
POS-CG [Wang *et al.*, 2019]	52.5	71.3	34.1	92.0
Mixture [Hou *et al.*, 2019]	52.8	71.8	36.1	87.8
RMN (S+L)	52.5	72.7	36.1	92.8
RMN (H+L)	54.6	**73.4**	**36.5**	**94.4**

	MSR-VTT			
Models	B@4	R	M	C
RecNet [Wang *et al.*, 2018]	39.1	59.3	26.6	42.7
MARN [Pei *et al.*, 2019]	40.4	60.7	28.1	47.1
OA-BTG [Zhang and Peng, 2019]	41.4	-	28.2	46.9
POS-CG [Wang *et al.*, 2019]	42.0	61.6	28.2	48.7
Mixture [Hou *et al.*, 2019]	42.3	**62.8**	**29.7**	49.1
RMN (S+L)	42.1	61.3	28.3	49.1
RMN (H+L)	**42.5**	61.6	28.4	**49.6**

Table 2: Comparing with the state-of-the-art on MSVD and MSR-VTT datasets. B@4, R, M, C denote BLEU-4, ROUGE_L, METEOR, CIDEr, respectively. The highest score is highlighted in bold and the second highest is underlined.

sure both soft fusion and hard module selector works well, RMN (H+L) outperforms RMN (S+L) with a margin.

Effect of Linguistic Loss

Comparing RMN (H+L) with RMN (H), as well as comparing RMN (S+L) with RMN (S), we find that the linguistic loss consistently improves the performance, indicating the importance of linguistic information for video captioning.

We can also observe the consistent results in Figure 3. By imposing the linguistic loss, we can enforce the module selector sticking with the linguistic structure, and the proportion of each module is approaching to the ground-truth.

Further, we carried out the word cloud statistics as shown in Figure 4. We can find that the three modules have a strong pattern: the LOCATE module mainly generates visual words, *e.g.*, "man" and "person"; the RELATE module mainly generates action words, *e.g.*, "playing" and "talking"; the FUNC module mainly generates function words, *e.g.*, "about" and "to". It indicates that the model has learned linguistic knowledge correctly.

4.4 Comparison with State-of-the-Art

We compare our proposed RMN with the most recent state-of-the-art methods on MSVD and MSR-VTT datasets. According to whether they leverage the POS labels, we categorize them into two groups: 1) traditional encoder-decoder based models MAM-RNN [Li *et al.*, 2017], RecNet [Wang *et al.*, 2018], MARN [Pei *et al.*, 2019], and OA-BTG [Zhang and Peng, 2019], and 2) POS strengthened model POS-CG [Wang *et al.*, 2019] and Mixture [Hou *et al.*, 2019].

As shown in Table 2, we can find that: 1) the methods that leverage POS labels outperform the methods without POS information, and 2) our proposed reasoning module network outperforms the methods with POS labels on most metrics and achieves new state-of-the-art. It is worth noting that CIDEr is proposed for captioning task specifically, and is considered more consistent with human judgment. Our model achieves the best CIDEr score on both datasets, which demonstrates our RMN model can generate captions that more in line with human descriptions.

4.5 Qualitative Analysis

In this section, we would like to investigate the generation process of our model by qualitative results. Here we provide some video captioning examples in Figure 5. As expected, our module selection is explicit and reasonable, for example, we compose "man", "flute", and "phone" with the LOCATE module, "playing", "riding", and "talking" with RELATE module, and "a", "of", and "the" with FUNC module.

The above observations suggest that the generation process of the proposed RMN is totally explicit, since at each timestep the module selects one determined module to produce the current word. In addition, each module is well-designed to perform spatio-temporal visual reasoning, *e.g.*, the spatio-temporal attention for LOCATE module and pairwise interaction reasoning over both space and time for RELATE module, indicating that our model is explainable.

5 Conclusion

In this paper, we proposed a novel reasoning neural module networks (RMN) for video captioning that performs visual reasoning on each step along the generation process. Specifically, we designed three sophisticated reasoning modules for spatio-temporal visual reasoning. To dynamically compose the reasoning modules, we proposed a discrete module selector which is trained by a linguistic loss with a Gumbel approximation. Extensive experiments verified the effectiveness of the proposed RMN, and the qualitative results indicated the caption generation process is explicit and explainable.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2017YFB1300201, the National Natural Science Foundation of China (NSFC) under Grants U19B2038, 61620106009 and 61725203 as well as the Fundamental Research Funds for the Central Universities under Grant WK2100100030.

References

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[Andreas *et al.*, 2016] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, 2016.

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[Carreira and Zisserman, 2017] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[Chen and Dolan, 2011] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011.

[Chen and Jiang, 2019] Shaoxiang Chen and Yu-Gang Jiang. Motion guided spatial attention for video captioning. In *AAAI*, 2019.

[Cirik *et al.*, 2018] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Using syntax to ground referring expressions in natural images. In *AAAI*, 2018.

[Donahue *et al.*, 2015] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[Guadarrama *et al.*, 2013] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Hong *et al.*, 2019] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *T-PAMI*, 2019.

[Hou *et al.*, 2019] Jingyi Hou, Xinxiao Wu, Wentian Zhao, Jiebo Luo, and Yunde Jia. Joint syntax representation learning and visual cue translation for video captioning. In *ICCV*, 2019.

[Hu *et al.*, 2017] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, 2017.

[Hu *et al.*, 2018] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *ECCV*, 2018.

[Jang *et al.*, 2017] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

[Kay *et al.*, 2017] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv:1705.06950*, 2017.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Kojima *et al.*, 2002] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *IJCV*, 50(2):171–184, 2002.

[Li *et al.*, 2017] Xuelong Li, Bin Zhao, Xiaoqiang Lu, et al. Mam-rnn: Multi-level attention model based rnn for video captioning. In *IJCAI*, 2017.

[Lin, 2004] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, July 2004.

[Liu *et al.*, 2018] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for sequence-level image captioning. In *ACM MM*, 2018.

[Liu *et al.*, 2019] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *ICCV*, 2019.

[Pan *et al.*, 2017] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, 2017.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[Pei *et al.*, 2019] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *CVPR*, 2019.

[Ren *et al.*, 2016] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2016.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.

[Szegedy *et al.*, 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

[Tian and Oh, 2019] Junjiao Tian and Jean Oh. Image captioning with compositional neural module networks. In *AAAI*, 2019.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[Venugopalan *et al.*, 2015] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *ICCV*, 2015.

[Wang *et al.*, 2018] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *CVPR*, 2018.

[Wang *et al.*, 2019] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. Controllable video captioning with pos sequence guidance based on gated fusion network. In *CVPR*, 2019.

[Xu *et al.*, 2016] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[Yang *et al.*, 2019a] T. Yang, Z. Zha, and H. Zhang. Making history matter: History-advantage sequence training for visual dialog. In *ICCV*, 2019.

[Yang *et al.*, 2019b] Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. In *ICCV*, 2019.

[Yao *et al.*, 2015] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015.

[Zha *et al.*, 2019] Z. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu. Context-aware visual policy network for fine-grained image captioning. *T-PAMI*, 2019.

[Zhang and Peng, 2019] Junchao Zhang and Yuxin Peng. Object-aware aggregation with bidirectional temporal graph for video captioning. In *CVPR*, 2019.