IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 38, No. 4, August, 2021, pp. 937-946
Journal homepage: http://iieta.org/journals/ts

Label Importance Ranking with Entropy Variation Complex Networks for Structured Video Captioning

Wenjia Tian, Yanzhu Hu*

Modern Postal College, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding Author Email: bupt_automation_safety_yzhu@bupt.edu.cn

https://doi.org/10.18280/ts.380403

Received: 10 April 2021
Accepted: 25 June 2021

Keywords:
video captioning, label importance, complex networks, entropy variation

ABSTRACT

Structured video captioning is a fundamental yet challenging task in both computer vision and artificial intelligence (AI). The prevalent approach is to map an input video to a variable-length output sentence with models like recurrent neural network (RNN). This paper presents a new model based on an improved scene-aware bidirectional long short-term memory network (SABi-LSTM), and names the model as label importance ranking with entropy variation complex networks of structured video captions. Structured video captioning is a three-level structured system, including a multi-feature fusion level, an SABi-LSTM level, and a label importance ranking level. The system decomposes structures of multiple levels and dimensions from different perspectives to perform video captioning. This work affirms the theoretical and practical significance of label importance ranking to video caption generation, and regards entropy as a local level metric to quantify label importance. Hence, entropy variation was proposed to define label importance, namely, the variation of the network entropy through label removal. It is assumed that the removal of an important label could cause sustainable variation to the structure. Hence, the authors defined the label importance ranking with entropy variation complex network algorithm to calculate the weight model of label nodes marked by video, and obtain the final caption of the video. Empirical results on Microsoft Video Caption (MSVD) dataset and MSR-Video to Text (MSR-VTT) dataset demonstrate the superiority of our approach for structured video captioning, especially on MSVD dataset.

1. INTRODUCTION

The development of mobile Internet enables various media to disseminate information. In the era of mobile Internet, video becomes an important carrier of information, and video analysis attracts more and more attention. With a complex structure, video usually contains a huge amount of data with rich features. Describing video in natural language is trivial for human beings, but thorny for machines. Many problems need to be solved in order to effectively process multimedia videos, and fully understand the relevant data.

Computer vision has developed rapidly thanks to the proliferation of neural networks and emergence of various open-source datasets. Against this backdrop, there are two trends in the development of video interpretation and processing technology: video classification, and video captioning. For video classification, a video clip can be classified by spatiotemporal features of image frames or action-containing videos [1-3]. For video captioning, the aim is to dividing each image into multiple regions, and label meaningful phrases or sentences. At present, video captioning and its application are still in the preliminary stage [4-7].

With the advancement of deep learning (DL) frameworks, most scholars have tried to caption videos with an encoding and decoding structure: the video features are often extracted by a convolutional neural network (CNN), the video codes are transformed into a semantic eigenvector, the statements are treated as a sequence generation process, and the words and sentences are generated iteratively by the neural network, using context information. But video captioning faces a challenging problem: The video captioning algorithm needs to detect the moving people or objects in the video based on moving speed and direction, and describe them with accurate words. However, it is difficult to automatically detect the few important people or objects in a long video, which often involves multiple interrelated events. Only these people or objects need to be described by relationships.

To improve the video captioning quality, this paper presents a novel structured video captioning model based on entropy variation complex networks. The proposed model firstly extracts multiple features from the video, including the feature of each static frame, the existence of objects in the frame, and the spatiotemporal features of the whole video, and then stitch the extracted features into a natural caption statement, according to the feature lengths of memory encoding and decoding network. Next, a new video boundary-aware bidirectional long short-term memory network (BiLSTM) was designed to identify the discontinuities in video frames, and to better encode a video with multiple actions. After that, entropy variation complex networks were introduced to generate captions, making the generated statements more certain, and the video captions more accurate. Different feature extraction methods with multiple modes and video features were adopted to obtain more information, thereby adaptively controlling the effects of different modal features on word generation, acquiring more video contents, and generating richer

descriptive texts. Then, natural language information was derived from word entropy, and natural expressions were obtained to enhance the generalization and practicality of our model. Finally, our model was proved better than other approaches on MSVD and MSR-VTT datasets.

2. LITERATURE REVIEW

The existing video captioning methods generally adopt one of the following four strategies [8]: (1) Assigning the words detected in visual contents to each sentence fragment, and generating video captions based on the predefined language template; (2) Learning the probability distribution of the joint space composed of visual contents and text sentences; (3) training the attribute detector through multi-example learning, and generating video captions by a maximum entropy language model based on the output of the detector; (4) integrating the semantic features mined from the frame sequence into video captions, using a CNN/circular CNN (CCNN) with a simple linear transfer unit. The first strategy depends heavily on the template, and generates sentences of a fixed structure. The second strategy outputs sentences with a flexible syntactic structure. Unlike these two strategies, the third and fourth strategies consider the semantic features in video captions, but fail to deeply integrate the semantic features in different domains.

When it comes to the sequence-to-sequence learning part, early studies on automatic captioning of visual contents are mostly grounded on templates [9-11]. These template-based approaches can be regarded as a bottom-up method, which generates video captions in two steps: generating visual content descriptors from local features through object detection; and filling the generated words into the predefined language template, and selecting the sentence with the highest probability of occurrence, using the probabilistic language model. The LSTM is capable of processing sequential data from the video. Venugopalan et al. [12] combined a deep CNN (DCNN) with LSTM to learn spatiotemporal features of the video: The two-dimensional (2D) network features are extracted by CNN encoder; the features of all frames are averaged to obtain the vector representation of video contents; the features are sequentially imported to LSTM to parse the dynamic video information. This hybrid approach marks a pioneering progress in video captioning. However, the averaging of image features ignores the timing features of the video [13]. The sequence-to-sequence video to text (S2VT) model further encodes the sequence of frame features with an LSTM encoder, and generates a high-quality representation vector of video contents, for the chain structure of LSTM is similar to the structure of the frame sequence. During the experiments, the order of frames in the sequence were shuffled randomly, highlighting the importance of sequence stationarity. The main defect of S2VT is the insufficient use of local information. The model performance mainly depends on the expression ability of global features. If the length of video objects varies from tens of seconds to tens of minutes, it would be difficult to obtain expressive image features with CNN. Besides, local information tends to be lost, when the expression ability of features is limited. Jin et al. [14] fused various types of features, namely, image features, video features, and species features, to represent the video, and obtained accurate results with the abundance of extracted features. Pasunuru et al. [15] proposed a multi-task learning method, including unsupervised frame prediction, synonym generation, and caption generation, for video captioning, to reshare features and parameters in the three tasks, thereby improving the model accuracy.

As for the natural language processing part, HALogen representation [15, 16], head-driven phrase structure grammar (HPSG) [17], and document planner [18] define the description rules for the structure of language expressions, ensuring the grammatical correctness of the generated sentences. Following production rules, the syntax can generate lots of different configurations from a relatively small vocabulary. With the aid of DL, Rohrbach et al. [19] proposed an LSTM encoder and decoder with conditional random field (CRF), which integrates probability distribution in language processing to generate statements. Huang et al. [20, 21] added attention mechanism to image captioning. Chen et al. [22, 23] processed natural languages in video captioning, focusing on image objects. Most strategies in the last two years, such as dual-stream recurrent neural network, object relational graph (ORG) with teacher-recommended learning (TRL), and spatio-temporal graph with knowledge distillation (STG-KD) [24-26], are optimized with features of video images. Few of them take the natural captioning of sentences into full consideration. Thus, the importance of natural language processing in video captioning is seriously underestimated.

The importance of nodes is often ranked by the information on network structure [27-29]. One of the basic tools to capture the structure information of complex networks is entropy [30, 31]. This paper proposes a comprehensive evaluation method of node importance based on relative entropy, which optimizes the sentence captions using different centrality indices through linear programming. Rather than directly indicate node importance ranking, relative entropy is a hybrid metric that integrates the importance ranking of existing nodes. The expected force is a node attribute derived from local network topology, and could be approximated as the entropy of forwarding connectivity of each transmission cluster. On this basis, a video captioning model was designed based on multiple modes and information entropy. Apart from integrating multi-mode features, the statement effect generated by video captioning was optimized according to the principle of information entropy, such that the captions are smoother and more consistent with video expression.

3. METHODOLOGY

The research roadmap is shown in Figure 1 below.

(a) Label importance ranking with entropy variation (label importance is defined as the variation of network entropy through the removal of the label, assuming that the removal of an important node could cause substantial variation in network structure)

(b) SABiLSTM encoding unit

(c) Multi-feature processing (a semantic network is constructed using entropy variation complex networks, and feature words are sorted in descending order of hybrid eigenvalues to generate descriptive statements)

(d)Whole image processing (the three-level structured video captioning system decomposes the structures of multiple levels and dimensions from different perspectives to establish the topology of video captioning network)

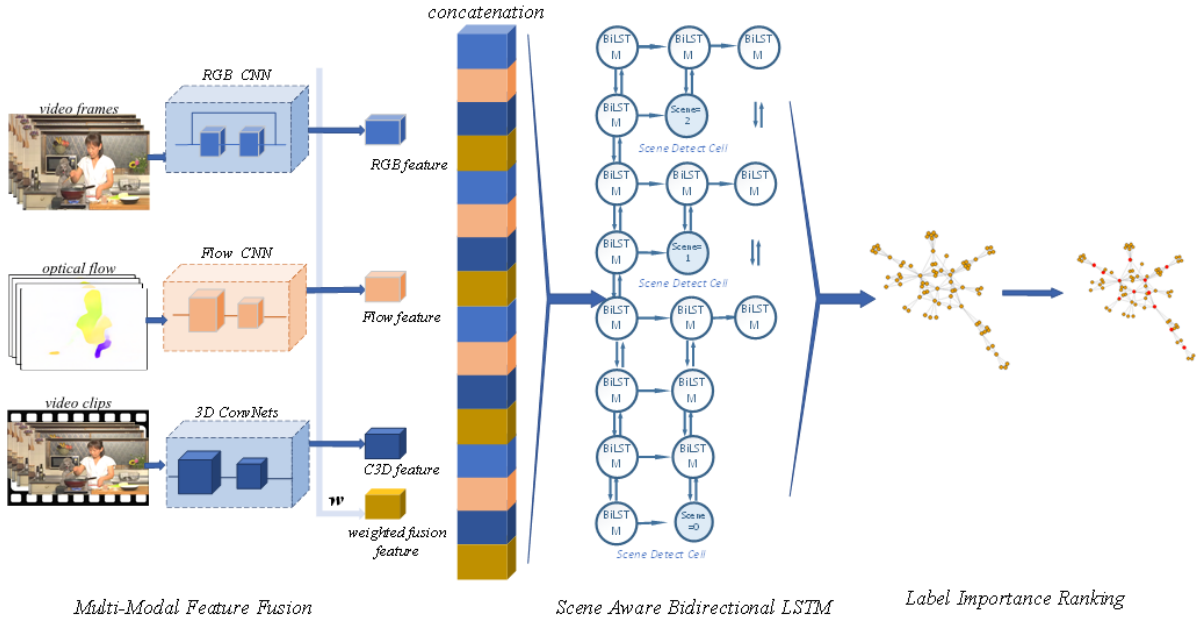

Figure 1. Research roadmap

3.1 Feature selection of entropy-based complex networks

3.1.1 Construction of complex networks

Each complex network is composed of nodes and edges. The smallest unit that represents the complete semantic information in a text is a sentence. Hence, this paper treats sentences as nodes, and analyzes the structural features of a text with sentences as the unit. Edges are defined as the connections between two sentences with a common noun. The two sentences linked up by an edge may elaborate on the same topic, or convey supplementary information on the same topic. Although they might contain redundant information, the two sentences are highly similar in contents. Based on the common noun relationship of sentence pairs, it is possible to build a complex network for the target text.

After preprocessing, the nouns in each sentence are mapped to the network. Then, two matrices A and W could be defined, representing adjacency matrix and n-order matrix weight (N is the number of nodes), respectively. In matrix A, if there is an edge between nodes i and j, then A_{ij} equals 1, and any other item equals zero. In matrix W, w_{ij} is the number of occurrences of common words between i and j.

The original text should be preprocessed before constructing the weighted complex network. The proprocessing technqiues include word segmentation, word extraction, and the filtering of meaningless workds (e.g., pause words). If the text is in English, the feature words need to be restored to the prototype form, that is, the set of feature words should be extracted from the text.

The edges are assigned between nodes by Cancho and Solé's method, that is, an edge is added between the keywords in a sentence that spans no greater than 2 keys, and the weight of the edge spans no greater than 2 co-occurrences. In this way, a text can be transformed into a weighted complex network. For a given text T, the internal keys are treated as nodes in a network of $V=(v1, v2,...,vn)$, where V_i is a keyword, according to the edges between key nodes:

$$E = \{(Vi, vj) \mid Vi, vj \in V\},$$
$$W = \{wij \mid (vi, vj) \in V\} \quad (1)$$

In this way, a weighted complex network can be constructed as $G=(V, E, W)$.

3.1.2 Weighting of entropy variation in complex networks

Based on entropy-weighted complex networks, the text captioning algorithm needs to model the text as an entropy-weighted complex network, and then analyze each network node that represents a feature word. In addition, the feature words with a large composite eigenvalue will be extracted as the keywords of the text. According to complex network theories, the nodes with a large composite eigenvalue tend to cluster in local modules, and play a key role in linking up the nodes within the entire network, increasing the density and intensity of network edges. To calculate the composite eigenvalue of each node, our algorithm fully considers the weighted clustering coefficient and the number of intermediate nodes, and extracts the nodes with a large composite eigenvalue from the weighted complex network. The corresponding feature words are the keywords of the text. The algorithm flow is detailed below.

Step 1. Input the original text T for keyword extraction.

Step 2. Preprocess the original text, and extract the feature words.

Step 3. Take each feature word as a network node, connect the words in the same sentence with a span no greater than 2 with edges, and weigh each edge with the co-occurrence of words, creating a weighted complex network.

Step 4. Pair the nodes in the weighted complex network, compute the weighted clustering coefficient of nodes, the number of intermediate nodes, and the composite eigenvalue CP of network nodes:

$$CP_i = \alpha \frac{WC_i}{2\sum_{j=1}^{N} WC_j} + \frac{(1-\alpha)P_i}{2(N-1)(N-2)} \quad (2)$$

where, α is an adjustable parameter; N, is the number of nodes in the network; $W = \{wij \mid (vi, vj) \in V\}$; C is the distance between node i and the other nodesm $C_i = \frac{1}{\sum_{j=1}^{N} d_{ij}}$; P is the

importance of a node, $P_i = \frac{1}{\sum_{j=1}^{N} d_i}$.

Step 5. Sort the feature nodes in descending order of eigenvalue, extract the first k feature words with a large composite eigenvalue, and take them as the k keywords of the text.

Step 6. Output the k keywords of the text.

3.1.3 Application of entropy variation complex networks

In our model, the semantics obtained by the SABi-LSTM are built into a semantic network, with each feature word as a node, and an edge between each pair of words spanning no greater than 2. The number of word co-occurrences is the weight of each edge. In the resulting weighted complex network, the weighted clustering coefficient of node v can be calculated by formula (1), and the intermediate nodes of node v can be computed by formula (2). After that, the CP value of the node is calculated, and the feature words are sorted in descending order by that value. Eventually, these keywords are constructed into description statements.

3.2 Multi-feature fusion

The flow of multi-feature fusion is depicted in Figure 2. Linear and nonlinear mappings have little difference in the case of high-dimensional data. But linear classification or regression usually achieves better results at a faster speed. After the high-dimensional features are extracted from the video, linear mapping is performed to reduce the dimensions of the features, and realize bidirectional labeling of the video and sentences.

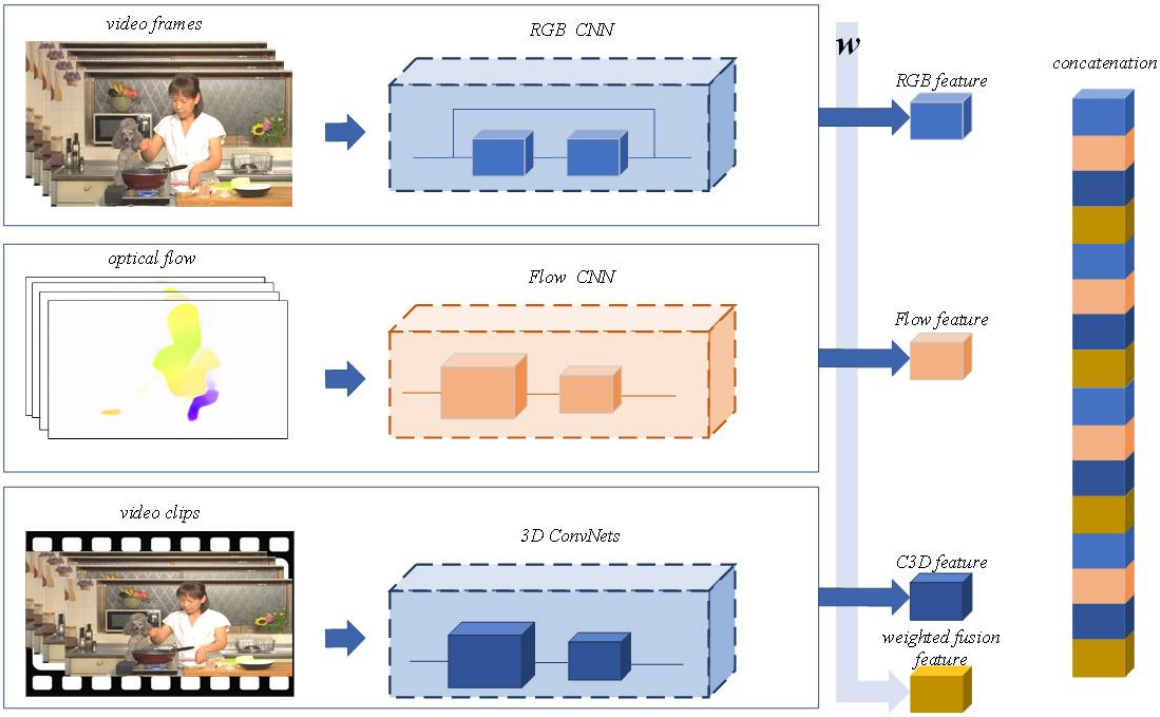

Figure 2. Multi-feature fusion

This paper fuses the multiple features of the video, which belongs to the intermediate level of information fusion. The basic theory of feature fusion is information fusion theory. Information fusion refers to the comprehensive processing of multi-source heterogeneous data, laying the basis for joint decision-making.

Taking video recognition as a pattern classification problem, the video features can be fused under two empirical assumptions: (1) Multi-feature fusion tends to improve the classification performance from that based on a single feature; (2) Multi-feature fusion is the starting point of pattern classification, while single feature is the guide for the selection of multiple image features.

Feature fusion directly employs the current feature extraction algorithm to mine multiple features from the video than a single feature. This approach is much less costly than redesigning the features or the feature extraction algorithm. Different feature descriptors, such as red-green-blue (RGB) feature, and light flow motion, are utilized comprehensively to illustrate different aspects of the video, breaking the limitation of single feature-based content captioning. The multi-feature fusion is realized in two steps.

Step 1. Feature stitching

During feature extraction, each model uses a vector F to represent the whole video, and to splice the features extracted from various models. Then, F_{fusion} is directly assembled by selecting the combination of these features. The video feature imported to the natural language captioning model.

Step 2. Weighted summation

The features extracted from different models are aligned by length, and trainable weight vectors are set for weighted summation of the features. Then, the fused features are imported as video features into the natural language captioning model:

$$\mathrm{F}_{fusion} = WF^T = (w_1F_1 + w_2F_2 + \cdots + w_mF_m) \quad (3)$$

where, F_{fusion} is the final fused feature; m is the number of features being used, $\sum_{i=0}^{m} w_i = 1$.

3.3 Video boundary-sensing BiLSTM encoding unit

This paper proposes a new LSTM unit to identify the discontinuities in video frames, i.e., the discontinuities between actions, such as to better encode a video with multiple actions. For a given input video, the video encoder can output a sequence $(s_1,s_2,\ldots,s_m)$ for the entire video based on the input $(x_1,x_2,\ldots,x_n)$. In the encoder, the connection between layers varies with the current input and hidden state.

Then, a scene- aware loop unit is defined to modify layer connectivity over time. Whenever the action changes, the hidden state and cell memory of LSTM are reinitialized, and the hidden state of the output layer is exported at the end of the segment, i.e., the feature output of the current image. Hence, the input data following time boundaries are not affected by the contents before the boundaries, and a hierarchical representation of the video is generated, where each block encompasses similar frames.

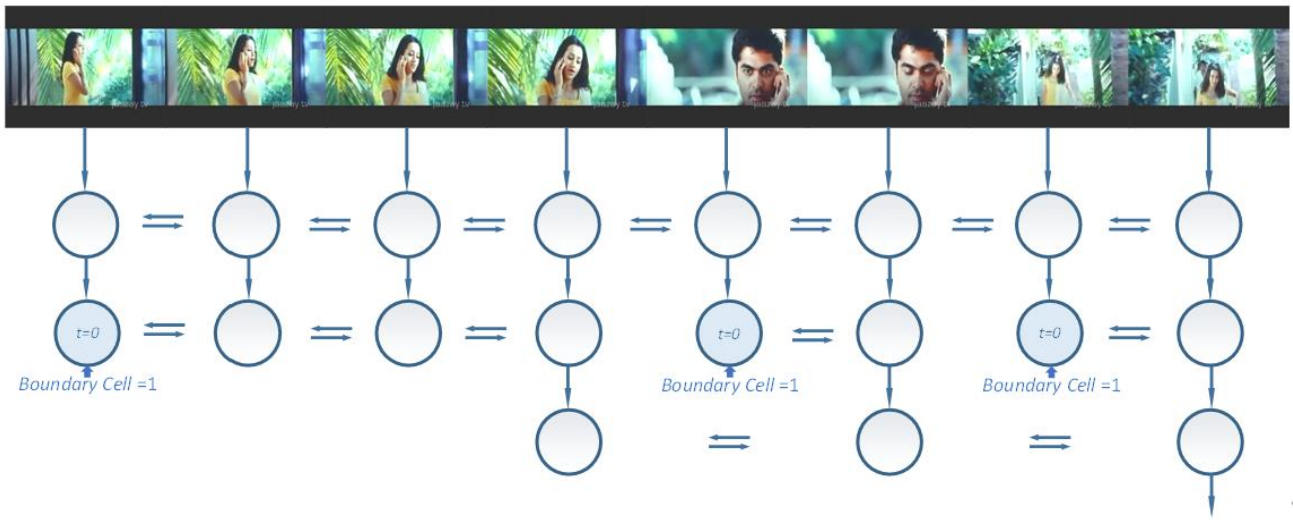

Figure 3. Time connections determined by the scene-aware encoder and the common LSTM encoder

Figure 3 compares the time connections determined by the scene-aware encoder and the common LSTM encoder. Our encoder is based on LSTM unit, which can learn patterns with wide time dependence. At the core of the encoder lies a storage unit, capable of preserving the observed inputs in a time step.

The memory is updated under the control of three gates, all of which are combinations of the current input and the previous hidden state, followed by the sigmoid activation function. The input gate controls the addition of input to the memory; the forget gate controls what the unit forgets; the output gate controls whether the current memory should be outputted.

In each time step, the hidden state and storage location are transferred to the next time step, or reinitialized according to the detected state. Hence, the seamless update and processing of the input sequence are interrupted. The boundaries of each block are derived by a learnable function, which varies with the inputs. Finally, the scene-aware encoder is established as a linear combination the current input and the hidden state, followed by the activation function, a combination of sigmoid function and step function:

$$s_t = \tau\left(v_s^T \cdot \left(W_{si}x_t + W_{sh}h_{t-1} + b_s\right)\right) \tag{4}$$

$$\tau(x) = \begin{cases} 1, if\ \sigma(x) > 0.5 \\ 0, otherwise \end{cases} \tag{5}$$

where, v^T_s is a learnable row vector; W_{sh} and b_s are learning weight and bias, respectively.

Let s_t be the state before unit update. According to the state, the hidden state and memory unit at the start of a new segment will be transferred or reinitialized:

$$\mathrm{i}_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \tag{6}$$

$$\mathrm{f}_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \tag{7}$$

$$g_t = \sigma(W_{gx}x_t + W_{gh}h_{t-1} + b_g) \tag{8}$$

$$c_t = \mathrm{f}_t \odot c_{t-1} + \mathrm{i}_t \odot g_t \tag{9}$$

$$o_t = \phi(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \tag{10}$$

$$h_t = o_t \odot \phi(c_t) \tag{11}$$

$$h_{t-1} \longleftarrow h_{t-1} \cdot (1 - s_t) \tag{12}$$

$$c_{t-1} \longleftarrow c_{t-1} \cdot (1 - s_t) \tag{13}$$

where, $\odot$ is the Hadamard product; σ is an S-shaped function; ϕ is the hyperbolic tangent function tanh; W_* is the learning weight matrix; b_* is the learning bias vector. The internal state h and the memory unit C are initialized as zero.

Then, the gates are recalculated from the resulting state and memory unit, and adopted for the next time step. The encoder generates output only at the end of the segment. If S_t=1, the hidden state of time step t-1 will be passed onto the next layer.

Figure 4 presents the structure of our scene-aware encoder. The above formulas are executed layer by layer to produce a variable-length output set $(s_1, s_2, \cdots, s_m)$, where m is the number of segments. Each output conceptually summarizes the contents of the segments detected in the video. The output

set is passed onto another layer to build a hierarchical representation of the video. Hence, the output of the scene-aware encoder is imported to another LSTM layer, and the final hidden state is taken as the eigenvector of the entire video.

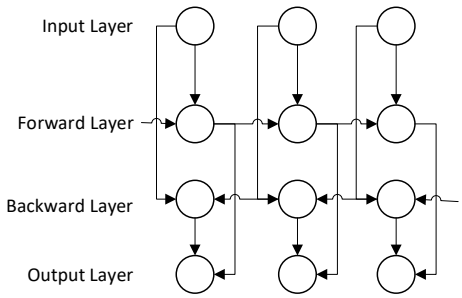

Figure 4. BiLSTM unit

The previous video encoding methods simply stack many layers together, adding to the nonlinearity of LSTM structure, or build a layered architecture in which the lower level encodes fixed-length blocks, while the higher level merges the blocks into the final video representation. By contrast, our encoder can generate variable length blocks according to the input features, and encode them in a hierarchical structure, without damaging the structure of the neural network.

4. EXPERIMENTAL VERIFICATION

4.1 Datasets

Two popular benchmark datasets about video captioning were chosen to evaluate our technique, namely, Microsoft Video Caption (MSVD) [32] dataset and MSR-Video to Text (MSR-VTT) dataset [33].

The MSVD is composed of 1,970 open domain YouTube videos, which predominantly show only a single activity each. Each clip spans between 10 and 25 seconds. The dataset provides multilingual human-annotated sentences as captions for the videos. The English captions were selected for our experiments. On average, 41 ground truth captions can be associated with a single video. For benchmarking, the common rule was adopted to design a training set of 1,200 samples, a validation set of 100 samples, and a testing set of 670 samples [34-36].

The MSR-VTT is a relatively new dataset of various open domain videos for captioning. A total of 7,180 videos are transformed into 10,000 clips. These clips are grouped into 20 classes. Following the common rule, the 10,000 clips were divided into a training set of 6,513 samples, a validation set of 497 samples, and a testing set of 2,990 samples. Each video has 20 annotations per sentence, which were made by Amazon Mechanical Turk (AMT) workers. This is one of the largest clip-sentence pair datasets available for the video captioning. That is why this dataset was chosen for benchmarking.

4.2 Evaluation metrics and data preprocessing

4.2.1 Evaluation metrics

Our technique was compared with the existing methods against three popular metrics: bilingual evaluation understudy (BLEU) [37], metric for evaluation of translation with explicit ordering (METEOR) [38], and consensus-based image caption evaluation (CIDER) [39]. The original definitions of these metrics were taken. A variable threshold (0.3, 0.5, 0.7, and 0.9) was designed according to the summary of the mean values of the three metrics. METEOR was adopted as the main comparative index, because it is the closest index to human judgment, when only a few reference captions are available [36].

4.2.2 Data preprocessing

The captions in both datasets were converted to lowercase, and all punctuations were removed, before marking each sentence. The vocabulary of MSVD was set to 5,497, and that of MSR-VTT to 23,500. Then, vectors were initialized randomly to generate embedded vectors, and the resulting vectors were phased to produce the final vector. After that, a dataset-specific network adjustment was carried out for the pertained word embedding. To train our model, a start tag and an end tag were included in the caption to cope with the variable length of sentences. The maximum sentence length was set to 40 words in the MSVD dataset, and 50 words in the MSRVTT dataset. These length limits are based on the available captions in the dataset. If the length of a sentence exceeds the limit, the sentence will be cut off; if the length is too short, the sentence will be truncated to zero length.

4.2.3 Experiments

The video or frames can be represented well by the features extracted from well-trained, high-quality models. The following features were extracted for our experiments:

(1) Spatial features

This paper adopts the pretrained model to extract the spatial features from each frame sequence. In recent years, great breakthroughs have been made in CNN-based image classification, target detection, image semantics segmentation, etc. The features extracted by CNN can express the original image well. Hence, this paper selects residual network (ResNet), a popular CNN, to extract high-quality data from each dataset, i.e., the features of all images in the preprocessed frame sequence, and compute the mean of the features of the frame sequence. In this way, each video was represented by a 2,048-dimensional eigenvector.

(2) Motion features

Each video consists of many continuous frames. The motion changes between these frames makes it imperative to analyze the motion features in video analysis. This paper relies on Vedantam et al.'s method [40] to extract the optical flow of adjacent frames, and normalize the extracted data to [0, 255]. The normalized data were stored as image files. The number of optical flow images is 1 fewer than the number N of frames in the video. The pretrained model was applied to extract the top fully-connected layer (Fc7) features from each optical flow image as motion features, and compute the mean of the optical flow sequence features. Hence, a 2,048-dimensional eigenvector was obtained to represent the motion features of each frame.

(3) Temporal features

Unlike single-image captioning problem, video captioning involves the temporal correlation between frames. Thus, it is necessary to extract temporal features of each video. Hence, convolution three-dimensional (C3D) video features were extracted, and a model pretrained on the C3D features of Sports-1M Dataset was adopted. During the preprocessing, the frame size was adjusted to match the input dimension of the network. For 3D CNN, 16 frame clips of the extracted key frames were adopted as input.

Next, the words that appear at least three times were retained, yielding a vocabulary of 10,298 words. During network training, a start of sentence <BOS> tag and an end of sentence <EOS> tag were added to the beginning and the end of the caption, respectively, allowing the model to handle variable-length captions. During the test, SABi-LSTM was given a <BOS> tag, denoting the caption input of the first scene. Then, the probability of the section word was sampled according to the predicted distribution, and taken as the input of the next step until the prediction of an <EOS> tag. The generated sentences and phrases were imported to the entropy variable complex network, and the sentences were optimized according to node weights. The top-k value of the complex network was set to 5.

Then, the hyperparameters of our model was adjusted on the verification set, using the root mean square propagation (RMSProp) algorithm. The model was trained at the learning rate of 2×10^{-4}. During our experiments, the batch size was set to 100 for training, i.e., the whole model was iterated for 100 rounds. The sparse cross-entropy loss was adopted to train our model on NVIDIA Tesla P100 GPU, under the PyTorch framework.

4.3 Results

The results of our technique and the cutting-edge video captioning methods are compared in Table 1, where the columns are the metrics BLEU-4, METEOR, and CIDER. The results of the contrastive methods were directly drawn from the existing literature, which use the same evaluation scheme.

4.3.1 Results on MSVD

The merit of our technique, denoted as Entropy-SABi-LSTM-(R+C+F), was compared with several traditional LSTM-based methods, i.e., a CNN encoder plus a LSTM decoder model (MP-LSTM), and an LSTM-based encoder-decoder model (S2VT), as well as state-of-the-art methods like LSTM-E, gated recurrent unit (GRU)-routing convolutional network (RCN), physical process wrapped recurrent neural network (p-RNN), PickNet-VL [27], TDConvED [28], DenseLSTM [41], and BALSTM-(R+C) [42]. Table 1 compares the performance of different methods on MSVD. It can be observed that our technique is superior to all the other methods.

The BLEU-4, METEOR, and CIDER of our technique were also contrasted with those of the most advanced approaches that can achieve the best performance with various visual features. The comparison shows that our technique performed better than those approaches, such as MP-LSTM, S2VT, and BALSTM-(R+C). Owing to the entropy variable decoder, our technique outperformed BALSTM-(R+C) by 8.4%, 1% and 13.6% in BLEU-4, METEOR, and CIDER, respectively.

Table 1. Performance of different methods on MSVD

Model	BLEU-4	METEOR	CIDER
MP-LSTM(V)	37.0	29.2	53.3
MP-LSTM(V+C)	39.4	29.7	55.1
MP-LSTM(R)	50.4	32.5	71.0
S2VT(V+O)[43]		29.8	
S2VT(V+C)	42.1	30.0	58.8
LSTM-E (V+C)	45.3	31.0	
GRU-RCN	47.9	31.1	67.8
p-RNN (V+C)[43]	49.9	32.6	65.8
PickNet-VL[27]	46.1	33.1	76
TDConvED (V+C)[28]	49.8	32.7	67.2
DenseLSTM(V+C)[41]	50.4	32.9	72.6
BALSTM-(R+C)[42]	42.5	32.4	63.5
SABi-LSTM -(F)	38.9	27.5	55.1
SABi-LSTM -(C)	40	29.5	61.2
SABi-LSTM -(R)	43.2	30.2	61.6
SABi-LSTM-(R+C)	47.71	32.8	70.9
SABi-LSTM-(R+C+F)	48.2	33	73.4
Entropy-SABi-LSTM-(R+C+F)	50.9	33.4	77.1

Note: All values are reported as percentage (%). The short name in the brackets indicates the features, where G, V, C, O, R, and M denote GoogleNet, VGGNet, C3D, Optical flow, ResNet, and motion feature learned by 3D CNN on manual descriptors, respectively. The same below.

The above results reflect the effectiveness of our technique, i.e., label importance ranking with entropy variable complex networks, in video captioning, and its superiority over other techniques. Hence, label importance ranking is a promising direction for video captioning [44-46].

4.3.2 Results on MSR-VTT

The merit of out technique was also tested on MSR-VTT dataset against traditional LSTM-based methods like MP-LSTM and S2VT, as well as state-of-the-art techniques, namely: TA, LSTM-E, Hierarchical LSTM with Adjusted Temporal Attention (hLSTMat), Manhaĵan LSTM (MA-LSTM), M3, MM, Multi-Column Convolutional Neural Network (MCNN) +MCF, PickNet-VL [27], TDConvED [28], DenseLSTM [41], and BALSTM-(R+C) [42].

As shown in Table 4, our technique did not achieve the optimal values on BLEU-4 and METEOR, but realized the best performance on CIDER. In an entropy-complicated network, the optimization of sentences will be optimized in a caption more suitable for human language. CIDER, which mainly evaluates the similarity between candidate and reference sentences, is more capable of generating fluent sentences [47-50].

4.3.3 Ablation analysis

Several ablations were executed to analyze different modules with different combinations on the MSVD dataset. Table 3 compares the results of ablation experiments with different input modes. Compare with the first three models, when a certain mode was used alone, the RGB feature could

achieve better results. With the increase of modes and introduction of modal fusion, however, better results were achieved on MSVD.

Table 2. Performance of different methods on MSR-VTT

Model	METEOR	CIDER	Bleu_4
MP-LSTM(R)	25.4	35.8	34.1
MP-LSTM (G+C+A)	25.6	38.1	35.7
S2VT(R)	25.8	36.7	34.4
S2VT (G+C+A)	26.0	39.1	36.0
TA (R)	24.9	34.5	33.2
TA(G+C+A)	25.1	36.7	34.8
LSTM-E(R)	25.7	36.1	34.5
LSTM-E(G+C+A)	25.8	38.5	36.1
hLSTMat(R)	26.3	-	38.3
MA-LSTM(G+C+A)	26.5	41	36.5
M3(V+C)	26.6	-	38.1
MM(R+C)	27	41.8	38.3
MCNN+MCF(R)	27.2	42.1	38.1
PickNet-VL(R)	27.2	42.1	38.9
TDConvED1(R)	26.8	40.7	37.1
TDConvED2(R)	27.2	41.9	39
TDConvED (R)	27.5	42.8	39.5
DenseLSTM(V+C)	32.9	72.6	50.4
Entropy-SABi-LSTM-(R+C+F)	33.4	77.1	50.9

Table 3. Results of feature ablations

Model	RGB feature	Flow feature	Temporal feature	CIDER
Entropy-SABi-LSTM	√			61.6
Entropy-SABi-LSTM		√		55.1
Entropy-SABi-LSTM			√	61.2
Entropy-SABi-LSTM	√		√	70.9
Entropy-SABi-LSTM	√	√	√	73.4

Table 4 reports the results of ablation experiments on different modules. Compare with the traditional LSTM network, adding scene-awareness unit and BiLSTM could greatly improve the network accuracy. Meanwhile, adding entropy variation complex network alone also led to better results. This means the modules being added could effectively improve the performance of video captioning.

Table 4. Results of module ablations

Model	Scene-aware	Bidirectional	Entropy	CIDER
SALSTM	√			63.5
BiLSTM		√		70.2
Entropy-LSTM			√	73.6
SABi-LSTM	√	√		73.4
Entropy-SABi-LSTM	√	√	√	77.1

5. DISCUSSION

It can be also inferred from Table 1 that our technique clearly outshined the average pooling methods and the common LSTM-based method (S2VT). Besides, METEOR was compared with the state-of-the-art methods that rely on rich visual features for optimal performance. In this regard, our approach surpassed the closest competitor by a wide margin (1%).

Concerning the performance on MSR-VTT (Table 2), our technique outperformed other baselines in CIDER. Similar to the observations on MSVD, our technique achieved better performance than the basic CNN+RNN, MP-LSTM, by replacing RNN with convolutional layers in the decoder.

Further, different aspects of our technique were evaluated empirically. A few highlights are mentioned in the following text. If necessary, the readers can request for supplementary materials that support the discussion.

Whereas all the components of the proposed technique contribute to the overall performance, the biggest innovation of our work is the application of entropy variation complex networks to calculate the node weight of video captions. Unlike to the “nearly standard” averaging pooling in the existing captioning pipeline, the proposed use of entropy variation complex networks promises a significant performance gain for any method. Hence, the authors recommended replacing the mean pooling operation with our entropy variable complex network for the future techniques.

6. CONCLUSIONS

The present study was designed to determine the effect of label importance ranking in video captioning. Compared with the previous studies, our paper presents a completely new idea: the video captioning network is improved in the natural language processing part, instead of the convolutional part.

Specifically, our technique relies on label importance ranking with entropy variation complex network to caption videos. Firstly, multiple features were extracted from each static frame, including the feature of each static frame, the existence of objects in the frame, and the spatiotemporal features of the whole video. Then, the extracted features were fused into a natural caption statement through the encoding and decoding by an LSTM. During model training, the multi-modal video features extracted by different methods were fused with the information entropy of words in natural language. The fusion helps to adaptively control the influence of different modal features on the generated words, and to obtain the caption text containing more details of the video and more natural expression, making our technique more generalizable and practical. In addition, information entropy was applied to improve the caption generation process, and the uncertainty of idioms was introduced to improve the accuracy of video captioning.

Caption language generation offers an intriguing perspective to video captioning. This issue will be further explored in future research.

ACKNOWLEDGMENTS

This work was supported in part by Subproject of Beijing Municipal Science and Technology Commission (Grant No.: Z181100000618006).

REFERENCES

[1] Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M. (2014). C3D: Generic features for video analysis. CoRR, abs/1412.0767, 2(7): 8.

[2] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In

European Conference on Computer Vision, pp. 20-36. https://doi.org/10.1007/978-3-319-46484-8_2
[3] Sevilla-Lara, L., Liao, Y., Güney, F., Jampani, V., Geiger, A., Black, M.J. (2018). On the integration of optical flow and action recognition. In German Conference on Pattern Recognition, pp. 281-297. https://doi.org/10.1007/978-3-030-12939-2_20
[4] Chen, X., Zitnick, C.L. (2014). Learning a recurrent visual representation for image caption generation. arXiv preprint arXiv:1411.5654.
[5] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625-2634.
[6] Vinyals, O., Toshev, A., Bengio, S., Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156-3164.
[7] Lu, J., Yang, J., Batra, D., Parikh, D. (2018). Neural baby talk. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7219-7228.
[8] Sun, L., Jia, K., Yeung, D.Y., Shi, B.E. (2015). Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, pp. 4597-4605.
[9] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K. (2015). Sequence to sequence-video to text. In Proceedings of the IEEE International Conference on Computer Vision, pp. 4534-4542.
[10] Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., Saenko, K. (2013). Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2712-2719.
[11] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, pp. 4489-4497.
[12] Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K. (2014). Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729.
[13] Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., Mooney, R. (2014). Integrating language and vision to generate natural language descriptions of videos in the wild. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 1218-1227.
[14] Jin, Q., Chen, J., Chen, S., Xiong, Y., Hauptmann, A. (2016). Describing videos using multi-modal fusion. In Proceedings of the 24th ACM International Conference on Multimedia, pp. 1087-1091. https://doi.org/10.1145/2964284.2984065
[15] Pasunuru, R., Bansal, M. (2017). Multi-task video captioning with video and entailment generation. arXiv preprint arXiv:1704.07489.
[16] Langkilde-Geary, I., Knight, K. Halogen input representation.
[17] Levine, R.D., Meurers, W.D. (2006). Head-Driven Phrase Structure Grammar. Encyclopedia of Language & Linguistics, 5(4): 237-252.
[18] Reiter, E. (1996). Building natural-language generation systems. Computational Linguistics, 27(2): 298-300.
[19] Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., Schiele, B. (2013). Translating video content to natural language descriptions. In Proceedings of the IEEE International Conference on Computer Vision, pp. 433-440.
[20] Huang, L., Wang, W., Chen, J., Wei, X.Y. (2019). Attention on attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4634-4643.
[21] Li, X., Yuan, A., Lu, X. (2019). Vision-to-language tasks based on attributes and attention mechanism. IEEE Transactions on Cybernetics. https://doi.org/10.1109/TCYB.2019.2914351
[22] Chen, S., Jin, Q., Wang, P., Wu, Q. (2020). Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9962-9971.
[23] Li, Y., Yao, T., Pan, Y., Chao, H., Mei, T. (2019). Pointing novel objects in image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12497-12506.
[24] Xu, N., Liu, A.A., Wong, Y., Zhang, Y., Nie, W., Su, Y., Kankanhalli, M. (2018). Dual-stream recurrent neural network for video captioning. IEEE Transactions on Circuits and Systems for Video Technology, 29(8): 2482-2493. https://doi.org/10.1109/TCSVT.2018.2867286
[25] Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., Zha, Z.J. (2020). Object relational graph with teacher-recommended learning for video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13278-13288.
[26] Pan, B., Cai, H., Huang, D.A., Lee, K.H., Gaidon, A., Adeli, E., Niebles, J.C. (2020). Spatio-temporal graph for video captioning with knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10870-10879.
[27] Chen, Y., Wang, S., Zhang, W., Huang, Q. (2018). Less is more: Picking informative frames for video captioning. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 358-373.
[28] Chen, J., Pan, Y., Li, Y., Yao, T., Chao, H., Mei, T. (2019). Temporal deformable convolutional encoder-decoder networks for video captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, 33(1): 8167-8174. https://doi.org/10.1609/aaai.v33i01.33018167
[29] Liu, J.G., Ren, Z.M., Guo, Q. (2013). Ranking the spreading influence in complex networks. Physica A: Statistical Mechanics and its Applications, 392(18): 4154-4159. https://doi.org/10.1016/j.physa.2013.04.037
[30] Ai, X. (2017). Node importance ranking of complex networks with entropy variation. Entropy, 19(7): 303. https://doi.org/10.3390/e19070303
[31] Liu, J.G., Ren, Z.M., Guo, Q. (2013). Ranking the spreading influence in complex networks. Physica A: Statistical Mechanics and its Applications, 392(18):

4154-4159. https://doi.org/10.1016/j.physa.2013.04.037
[32] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2462-2470.
[33] Chen, D., Dolan, W.B. (2011). Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 190-200.
[34] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A. (2015). Describing videos by exploiting temporal structure. In Proceedings of the IEEE International Conference on Computer Vision, pp. 4507-4515.
[35] Wang, J., Wang, W., Huang, Y., Wang, L., Tan, T. (2018). M3: Multimodal memory modelling for video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7512-7520.
[36] Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Deng, L. (2017). Semantic compositional networks for visual captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5630-5639.
[37] Papineni, K., Roukos, S., Ward, T., Zhu, W.J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318.
[38] Banerjee, S., Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65-72.
[39] Vedantam, R., Lawrence Zitnick, C., Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566-4575.
[40] Vedantam, R., Lawrence Zitnick, C., Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566-4575.
[41] Zhu, Y., Jiang, S. (2019). Attention-based densely connected LSTM for video captioning. In Proceedings of the 27th ACM International Conference on Multimedia, pp. 802-810. https://doi.org/10.1145/3343031.3350932
[42] Baraldi, L., Grana, C., Cucchiara, R. (2017). Hierarchical boundary-aware neural encoder for video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1657-1666.
[43] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K. (2015). Sequence to sequence-video to text. In Proceedings of the IEEE International Conference on Computer Vision, pp. 4534-4542.
[44] Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K. (2014). Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729.
[45] Ballas, N., Yao, L., Pal, C., Courville, A. (2015). Delving deeper into convolutional networks for learning video representations. arXiv preprint arXiv:1511.06432.
[46] Trusina, A., Maslov, S., Minnhagen, P., Sneppen, K. (2004). Hierarchy measures in complex networks. Physical Review Letters, 92(17): 178702. https://doi.org/10.1103/PhysRevLett.92.178702
[47] Zhang, J., Jia, L., Niu, S., Zhang, F., Tong, L., Zhou, X. (2015). A space-time network-based modeling framework for dynamic unmanned aerial vehicle routing in traffic incident monitoring applications. Sensors, 15(6): 13874-13898. https://doi.org/10.3390/s150613874
[48] Zhao, L., Zhang, H., Wu, W. (2019). Cooperative knowledge creation in an uncertain network environment based on a dynamic knowledge supernetwork. Scientometrics, 119(2): 657-685. https://doi.org/10.1007/s11192-019-03049-4
[49] Xu, J., Mei, T., Yao, T., Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5288-5296.
[50] Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.