Not All Words are Equal: Video-specific Information Loss for Video Captioning

Jiarong Dong[12]
dongjiarong@ict.ac.cn
Ke Gao[1]
kegao@ict.ac.cn
Xiaokai Chen[12]
chenxiaokai@ict.ac.cn
Junbo Guo[1]
guojunbo@ict.ac.cn
Juan Cao[1]
caojuan@ict.ac.cn
Yongdong Zhang[1]
zhyd@ict.ac.cn

[1] Institute of Computing Technology,
Chinese Academy of Sciences,
Beijing, China

[2] University of Chinese Academy
of Sciences,
China

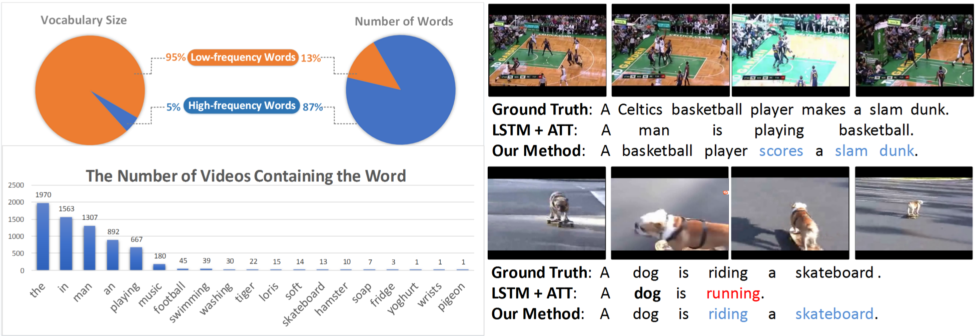
Abstract

An ideal description for a given video should fix its gaze on salient and representative content, which is capable of distinguishing this video from others. However, the distribution of different words is unbalanced in video captioning datasets, where distinctive words for describing video-specific salient objects are far less than common words such as 'a' 'the' and 'person'. The dataset bias often results in recognition error or detail deficiency of salient but unusual objects. To address this issue, we propose a novel learning strategy called Information Loss, which focuses on the relationship between the video-specific visual content and corresponding representative words. Moreover, a framework with hierarchical visual representations and an optimized hierarchical attention mechanism is established to capture the most salient spatial-temporal visual information, which fully exploits the potential strength of the proposed learning strategy. Extensive experiments demonstrate that the ingenious guidance strategy together with the optimized architecture outperforms state-of-the-art video captioning methods on MSVD with CIDEr score 87.5, and achieves superior CIDEr score 47.7 on MSR-VTT. We also show that our Information Loss is generic which improves various models by significant margins.

1 Introduction

Video captioning aims at generating both semantically and syntactically correct descriptions for a video. An ideal description for a given video should fix its gaze on salient and representative content, which is capable of distinguishing this video from others. Inspired by the recent progress of machine translation, LSTM network combined with attention mechanism

Corresponding author: Ke Gao.



(a) (b)

Figure 1: Illustration of description ambiguity in video captioning caused by the bias of dataset corpus. (a) The pie chart is about Distribution of vocabulary size and the number of words of the training corpus. The bar chart below is about the number of videos containing different words. (b) Examples of video description generation.

has dominated this task [17, 27]. In this case, the concept recognition is accomplished during the sentence generation process. Existing video captioning models are mainly optimized to maximize the probability of every ground truth word under cross entropy loss learning strategy. However, due to unbalanced distribution of word frequencies, the common words comprise the majority of the training loss and dominate the gradient. Consequently, models rarely achieve satisfactory performance to predict representative and distinctive words, which provide most information of given videos.

Figure 1(a) shows a data analysis about MSVD dataset, which is also the case in most video captioning datasets. The top 5% highest frequency words occupy almost 90% content of training corpus. That is, the captioning model spends the vast majority of effort on learning just very small subset of the vocabulary. Moreover, the most of high-frequency words are function words and common semantic words, which present limited information about the video-specific content. For example, the most frequent word "the" appears in all videos of MSVD, whereas the most of distinctive words just appear in a few specific videos. Directly training with this unbalanced data, the existing methods often suffer from description ambiguity, including detail deficency and recognition error [28]. As shown in Figure 1(b), the traditional method(LSTM + ATT for short in this paper) wrongly predicts ’riding a skateboard’ as ’running’.

To address the above issues, we propose a novel learning strategy called Information Loss to focus on learning the representative and distinctive words. Specially, we calculate importance value of every word for the content of given videos, which takes into account both information relevance and information content. Information relevance measures the degree of relevance between a word and the gist of the video. Information content corresponds to the discriminate degree of a word. Therefore, the distinctive words about the salient visual content would obtain high importance value, whereas common words and irrelevant words have low value. Then, the loss of every word is dynamically resized based on this importance value. In this case, the model pays more attention to learn video-specific informative words in every video sample, while the commonly used words can be well learned in overall captioning corpus.

In order to further exploit the potential strength of Information Loss, we established a new framework with hierarchical visual features and an optimized hierarchical attention mechanism. The optimized hierarchical attention model is established to adaptively extract the most important visual features from inter-and-intra video frames, which is the foundation of generating informative video descriptions.

In summary, the main contributions of this paper are: 1) A novel learning strategy called Information Loss is proposed to alleviate the description ambiguity problem caused by the bias of captioning corpus. 2) Moreover, an optimized framework consists of hierarchical visual representations and corresponding hierarchical attention is established to fully exploit the potential strength of the proposed learning strategy. 3) Extensive experiments demonstrate that the optimized structure combined with ingenious guidance can outperform state-of-the-art video captioning methods.

2 Related Work

Model structure for video captioning Most existing methods can be divided into the bottom-up methods [6, 18] and the top-down methods [7, 24]. The mechanism of the bottom-up method is intuitive. Some predefined concepts are recognized only on visual signals, and then linked into a sentence with fixed language templates. Captions generated in this way often lack flexibility. The top-down methods are inspired by the recent progress of machine translation, where LSTM network combined with attention mechanism has dominated video captioning. The attention mechanism is exploited based on the observation that not all parts of a video are equally important, which has been successfully applied in video captioning to select important frame-level features [17, 27]. There are few works focusing on both object-level and frame-level features in video captioning. Tu *et al.* [20] explore spatial and temporal attention. However, they ignore the interaction between objects over time. Therefore, we establishes the connection of objects between frames to focus on the informative and discriminative visual information of the entire video.

Learning strategy for video captioning Most of traditional video captioning systems are trained with the cross-entropy loss [25, 27]. Although effective for sequence generation tasks, the cross-entropy loss cannot guarantee the semantic correctness [11] and informativeness of generated captions. Pan *et al.* [11] proposes relevance loss together with cross entropy loss to enforce the relationship between sentence semantics and visual content. The Reinforcement Learning has been introduced in image captioning [10, 15] to improve models by sentence-level rewards, which needs well pretrained model to initialization. However, the both methods suffer from the imbalance learning problem caused by the dataset bias and have not taken distinctiveness into consideration. Our Information Loss focuses on learning representative and distinctive words to overcome dataset bias, which are complementary for these methods.

3 Approach

3.1 Overall Framework

As shown in Figure 2, the overall framework consists of three important components: hierarchical visual features extraction, hierarchical attention mechanism and Information Loss

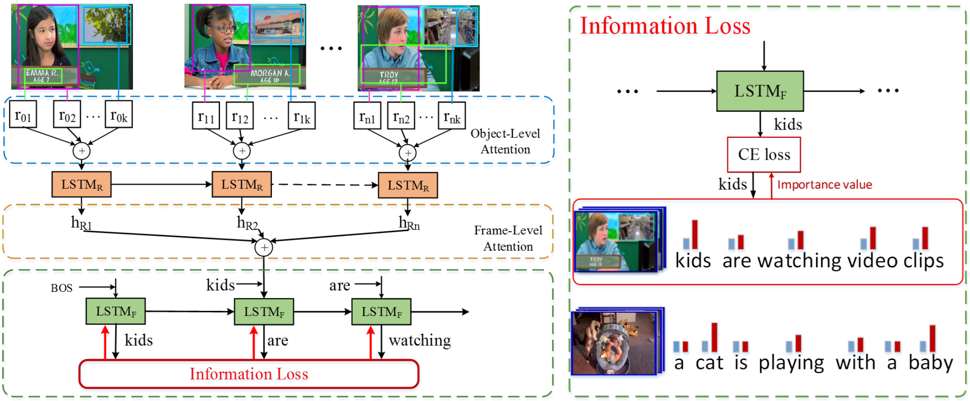


Figure 2: The framework of our Hierarchical Attention Model with Information Loss(HA_IL). The hierarchical attention mechnism established on hierarchical visual features captures the most salient visual information. Compared to traditional learning strategy(blue bar), the Information Loss focus on learning the relationship between 'kids' and video content with high importance value(red bar) of 'kids'.

guided learning strategy. Hierarchical attention mechnism is established on hierarchical visual features to capture the salient spatial-temporal visual content. Information Loss is the objective of our model which aims to address the description ambiguity problem. Details are given in section 3.2, section 3.3, section 3.4 respectively.

3.2 Hierarchical Visual Features Extraction

The exploitation of object-level features is a key component of our method to generate informative captions. We use Faster R-CNN [14] to generate a set of semantic region features from every frame. The Faster R-CNN model is pretrained on Visual Genome data [23] by [1] and can recognizes 1600 objects, which produces rich and diverse information. For every video frame, we select the top k most confident objects detected by Faster R-CNN model and extract features of these objects from the pool5 layer. Therefore, we get a set of semantic object-level features of a video, denoted as $VR = \{vr_1, \ldots, vr_n\}$, $vr_i = \{vr_{i1}, \ldots, vr_{ik}\}, vr_{ij} \in \mathbb{R}^{D1}$, where vr_{ij} means the j_{th} region feature of i_{th} frame. Then we apply ResNet [5] to extract global frame-level features, denoted as $VF = \{vf_1, \ldots, vf_n\}, vf_i \in \mathbb{R}^{D2}$, where n is the number of frames sampled from each video. In addition, for every video, we apply C3D [19] to extract several clip-level features, denoted as $VC = \{vc_1, \ldots, vc_n\}, vc_i \in \mathbb{R}^{D3}$. Finally, a given video is encoded as a feature set $V = \{VR, VF, VC\}$.

3.3 Hierarchical Attention Mechanism

Hierarchical attention mechanism is an effective achitecture to exploit hierarchical visual features. However, it is not easy to choose the real salient objects among hundreds of objects for the given video. As shown in Figure 3, different from the previous work [20] which directly attends to objects mostly based on syntactic context, we introduce an extra LSTM to establish the connection of objects between frames and focus on the salient objects based on

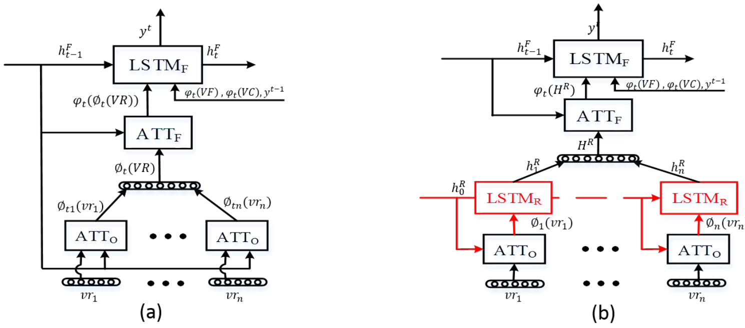


Figure 3: An illustration of hierarchical attention model of STAT(a) and our visual model(b).

the video content. Our model is composed of two LSTM layers associated with two attention layers. The first one is denoted as $LSTM_R$ which models temporal relationship of objects in a given video. The second one is denoted as $LSTM_F$ which acts as a language model to establish a syntactic context and generate natural language sentences. The two attention layers are focusing on object-level and frame-level information respectively.

Object-level attention with $LSTM_R$ The object-level attention with $LSTM_R$ is used to select salient semantic regions. Given the previous hidden state and object-level features at time step i, $LSTM_R$ encodes video information until i_{th} frame into the hidden state h_i^R:

$$h_i^R = LSTM_R(\phi_i(vr_i), h_{i-1}^R) \tag{1}$$

where the visual feature presentation of the i_{th} frame $\phi_i(vr_i)$ is the weighted sum of all the k semantic regions: $\phi_i(vr_i) = \sum_{j=1}^{k} \alpha_{ij} vr_{ij}$. The salient objects are emphasized by the attention weight α_{ij} which is obtained from the object-level attention layer:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j=1}^{k} \exp(e_{ij})} \quad \text{with} \quad e_{ij} = w_e^T \tanh(W_e h_{i-1}^R + U_e vr_{ij} + z_e) \tag{2}$$

where w_e, W_e, U_e, z_e are parameters to be learnt. It is observed from equation (2) that the attention weight of each object is not only determined by their own features but also by the video information of previous frames. Our motivation lying behind this is that the salient objects of a given video are the ones which have important relations to video context.

Frame-level attention with $LSTM_F$ The frame-level attention is proposed to allow $LSTM_F$ selecting relevant subset of frames while generating words step by step. For $LSTM_F$, the input are the results of frame-level attention $\varphi_t(H^R)$, $\varphi_t(VF)$, $\varphi_t(VC)$, the previous hidden state h_{t-1}^F, and the previous word y_{t-1}:

$$h_t^F = LSTM_F(W_E y_{t-1}, \varphi_t(H^R), \varphi_t(VF), \varphi_t(VC), h_{t-1}^F) \tag{3}$$

h_t^F absorbs information from visual content and syntactic context and is used to predict word. W_E is the word embedding matrix. H^R is expanded as $\{h_1^R, ..., h_n^R\}$, where h_i^R is the output of $LSTM_R$. Note that the time step i of $LSTM_R$ and the time step t of $LSTM_F$ are irrelevant which reduces the time for training and testing. $\varphi_t(H^R)$, $\varphi_t(VF)$ and $\varphi_t(VC)$ have the same function expression with different parameters. For simplicity, we use notation X to represent the input of $\varphi_t(\cdot)$. $\varphi_t(X)$ is dynamically computed by the weighted sum of the element x_i :

$$\varphi_t(X) = \sum_{i=1}^{n} \beta_{ti} x_i \quad \text{with} \quad \beta_{ti} = \frac{\exp(m_{ti})}{\sum_{i=1}^{n} \exp(m_{ti})}, \quad m_{ti} = w_m^T \tanh(W_m h_{t-1}^F + U_m x_i + z_m) \quad (4)$$

where w_m, W_m, U_m, z_m are parameters to be learnt. The attention weight β_{ti} reflects the importance of the i_{th} frame features given h_{t-1}^F. The predicted word y_t is obtained through a single hidden layer: $y_t = softmax(U_y h_t^F)$

3.4 Information Loss

The Information Loss is designed to alleviate description ambiguity problem. For clarification, we first briefly describe the traditional objective of video captioning. Given the previous groud truth word y_{t-1} and visual features V, the cross entropy loss(CE) is used to maximize the log-likelihood of the next groud truth word:

$$L_C(s,V) = -\sum_{t=1}^{T} \log p(y_t|(V, y_{t-1})) \quad (5)$$

Where $s = \{y_1, ..., y_T\}$ is the given caption. The CE loss is equal for all words. However, the common words can be learned well with total training data while the representative and distinctive words just can be learned from specific video samples. Different words have different importance for expressing video content. The learning strategy should focus more on the video-specific informative words about salient visual content.

Our Information Loss reshapes the standard cross entropy loss to up-weight the representative and distinctive words for every video sample, which encourages the video captioning model to capture the gist of a video and generate informative descriptions. Specifically, we introduce an importance value to measure both the information relevance and information content of every word for a given video. Then, we add the importance value as modulating factor to the cross entropy loss.

In existing video captioning datasets, each video clip is annotated with scores of descriptions by different workers. Although each video clip has various content, most persons reach consensus about the salient objects and activities. That is, the representative words corresponding to the main content of a video are more likely to appear in multiple ground truth captions, whereas the irrelevant words occur less frequently. Therefore, we regard word frequency among all ground truth captions of a video as the information relevance:

$$R(y_t|V) = \frac{N_{y_t,V}}{N_V} \quad (6)$$

where N_V is the number of captions owned by the given video V, $N_{y_t,V}$ is number of captions which contain word y_t. The information relevance term encourages the captioning model to capture the main content of a video. However, many commonly used words in the whole captions are also been encouraged. Therefore, we introduce an information content term to suppress the weight of high-frequency words and elevate the weight of informative words.

Information content is deeply discussed in information theory, which is a synonym for the surprise when a signal is received. We follow the way in information retrieval to calculate

Models	MSVD			MSR-VTT		
	BLEU-4	METEOR	CIDEr	BLEU-4	METEOR	CIDEr
NA(R+C+F)	49.7	33.6	76.1	40.8	27.0	45.2
TA-B1(R+C)	48.5	33.5	74.5	39.9	26.9	44.9
TA-B2(R+C+F)	50.2	34.6	77.0	41.4	27.3	45.3
HA(R+C+F)	**52.3**	**35.4**	**82.9**	**42.1**	**27.4**	**46.0**

Table 1: Performance comparison between baseline models and HA on the MSVD and MSR-VTT datasets. Here, R, C, F are shorted for ResNet, C3D, Faster R-CNN features.

information content of every word, which is defined as:

$$I(y_t) = \log(\frac{1}{p(y_t)}) \quad \text{with} \quad p(y_t) = \frac{|\{V_{y_t}\}|}{|\{V\}|} \tag{7}$$

where $|\{V_{y_t}\}|$ is the number of videos in which word y_t appears, $|\{V\}|$ is the overall number of videos, and $p(y_t)$ is the frequency that word y_t appears in different videos. The smaller the frequency of word y_t, the larger quantity of information carried by y_t. Combining the information content term and information relevance term induces the importance value of a word: $f(y_t,V) = R(y_t|V)^{\gamma} I(y_t)$. $\gamma \geq 0$ is a tunable parameter which smoothly adjusts the interreation of this two terms. When $\gamma = 0$, the importance value is equal to information content. While when γ is larger, the information relevance has more influence on the importance value, and the importance value of low frequency words of a given video will be rapidly diminished. We add the importance value as modulating factor to the cross entropy loss and define the Information Loss as:

$$L_I(s,V) = -\sum_{t=1}^{T}(1+\lambda f(y_t,V))\log p(w_t|V,y_{t-1}) \tag{8}$$

λ is the trade-off parameter. This loss focus more on important words than common words based on the importance value. Intuitively, the captioning model needs to matter the visual properties and generate more percision words to reduce the information loss.

4 Experiments

4.1 Dataset and Evaluation Metrics

We conduct extensive experiments on two widely used video captioning benchmarks: MSVD [4] and MSR-VTT [26]. MSVD contains 1970 Youtube video clips with around 40 human annotated descriptions per clip. Following the standard setting provided by [27], we takes 1200 videos for training, 100 videos for validation and the remainder for testing. MSR-VTT contains 10000 video clips with 20 human annotated descriptions per clip. We use the public splits: 6513 for training, 497 for validation and 2990 for testing. To evaluate the performance, we employ three commonly used evaluation metrics, including BLEU [13], METEOR [2], and CIDEr [21]. CIDEr gives higher score for video-specific n-grams than generic n-grams [22], which is used to measure the effectiveness of our method.

Models	BLEU-4	METEOR	CIDEr
S2VT	48.8	32.5	69.7
S2VT_IL	49.1	33.3	73.2
HA	52.3	35.4	82.9
HA_IL	**54.1**	**36.6**	**87.5**

Table 2: Performance comparison between different models trained with CE and IL

γ λ	1	2	3
0.3	83.6	84.4	83.3
0.5	85.3	**87.5**	84.3
0.7	83.7	85.6	83.2
1.0	83.1	84.6	82.4

Table 3: The effect of parameter

Models	B@4	M	C
TA	41.9	29.6	51.7
MAM-RNN	41.3	32.2	53.9
STAT	51.1	32.7	67.5
LSTM-TSA	52.8	33.5	74.0
hLSTMat	53.0	33.6	73.8
HA_IL	**54.1**	**36.6**	**87.5**

Table 4: Performance on MSVD

Models	B@4	M	C
TA	35.2	25.2	-
STAT	37.4	26.6	41.5
SA-LSTM	38.7	26.9	45.9
v2t_navigator	40.8	28.2	44.8
dense caption	41.4	28.3	48.9
HA_IL	**41.9**	27.9	47.7

Table 5: Performance on MSR-VTT

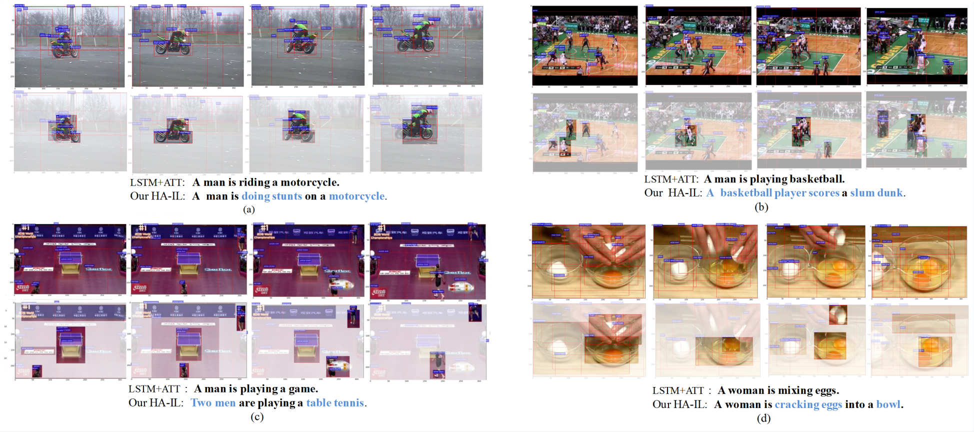
4.2 Implementation Details

We tokenize video captions and preserve the words which appear at least 5 times in training dataset while other words are replaced by word UNK. As result, we obtain a vocabulary size of 3656 for MSVD dataset and 8760 for MSR-VTT dataset. The object-level features is extracted by Faster R-CNN with dimension 2048. We set the number of objects from each frame to be 16. For frame-level features, we adopt the 2048-dimension pool5 feature from ResNet101 and the 4096-dimension fc6 feature from C3D. We uniformly sample 40 frames for each video with interval 8. We use a wider interval for long videos and pad zero frames for short videos. We empirically set the dimension of all the hidden units to 512. Dropout is used in both the input and output of two LSTM layers. We apply ADAM algorithm [8] with learning rate 10^{-4} to optimize our model under both cross entropy loss and Information Loss. We anneal the learning rate by a factor of 0.8 every 30 epochs. λ and γ are set to be 0.5 and 2 respectively. The sensitivity of λ and γ will be discussed in Section 4.3.

4.3 Experiment Analysis

Evaluation of hierarchical attention model Table 1 compares the performance of Hierarchical Attention Model (HA) and baseline models. NA applies none attention with mean-pooled features fed to every step of LSTM. TA-B1 applies temporal attention mechnism and uses none object-level features. TA-B2 applies temporal attention with hierarchical features. All of these models are trained under cross entropy loss. Compared to TA-B1, TA-B2 achieves better performance because of object-level features. Compared to NA, TA-B2 achieves very limited improvements which shows that temporal attention cannot exploit the hierarchical visual features effectively. On the contrary, our HA method outperforms NA model by a large margin on all metrics, such as over 8.9% relative improvement on CIDEr.

Evaluation of information loss To study the generalization ability of proposed Informa-



LSTM+ATT: A man is riding a motorcycle.
Our HA-IL: A man is doing stunts on a motorcycle.
(a)

LSTM+ATT: A man is playing basketball.
Our HA-IL: A basketball player scores a slum dunk.
(b)

LSTM+ATT : A man is playing a game.
Our HA-IL: Two men are playing a table tennis.
(c)

LSTM+ATT : A woman is mixing eggs.
Our HA-IL: A woman is cracking eggs into a bowl.
(d)

Figure 4: Visualization of HA_IL. For each frame, we visualize the attention weights of image regions with the degree of transparency. The less important region is covered with white color more heavily whereas the most salient region remains unchanged. It's clear that our proposed method is capable of focusing on the video-level salient objects.

tion Loss (IL) method, we have trained S2VT [25] model and our HA model under Information Loss(IL) learning strategy, namely, S2VT_IL and HA_IL. S2VT is our implemetion of [25] under cross entropy loss. Results are shown in Table 2. Compared to models learned by cross entropy loss, both performance of S2VT_IL and HA_IL are improved in all metrics when learned by our Information Loss. Especially, HA_IL improved by larger margin than S2VT_IL, which indicates that our HA model provides stronger visual representation to further reach the potentials of Information Loss. Overall, these results confirms the generalization and effectiveness of our Information Loss.

Study on parameters γ and λ Our Information Loss introduces two new hyperparameter, the tunable parameter γ and tradeoff parameter λ. Results using different γ and λ are shown in Table 3. When $\lambda = 0$, our loss is equivalent to cross entropy loss which is the result of HA model. As λ increases, the loss pays more attention in words with high importance value. With $\lambda = 0.5$, the IL yields the highest CIDEr score. γ is used to adjust the interation of information relevance and information content. When $\gamma = 2$, the two terms have harmonious collaboration. When $\gamma = 2, \lambda = 0.5$, the model achieves the best score.

State-of-the-art models On MSVD dataset, we compare HA_IL with five methods: TA [27], MAM-RNN [9], STAT [20], LSTM-TSA [12], hLSTMat [17]. TA is the first work to employ temporal attention in video captioning. MAM-RNN exploits region level attention. STAT explores object level features with spatial-temporal attention. LSTM-TSA incorporates semantic attributes with RNN frameworks. hLSTMat proposes an adjusted temporal attention. On MSR-VTT dataset, we conduct comparision with five methods : TA, STAT, SA-LSTM [3], v2t_navigator [7], dense caption [16]. SA-LSTM is the baseline method of MSR-VTT. v2t_navigator is the champion of the MM16 VTT challenge. dense caption focuses on dense video captioning and achieves the best result of MSR-VTT.

Overall results As shown in Table 4, our HA_IL model outperforms previous methods in all metrics by a large margin on MSVD, such as 18% improvement in CIDEr. The performance of our base model HA is better than STAT, suggesting that our hierarchial attention model associated with two LSTM layers is easier to capture the gist of a video. These results confirm the effectiveness of our HA_IL. As shown in Table 5, our HA_IL model achieves competitive results on MSR-VTT dataset, such as 41.9 on BLEU-4. Notably, v2t_navigator and dense caption use more kinds of features and execute a lot of engineering efforts such as sentence re-ranking. Hence, we consider that our method achieves comparable performance with less features. It's worth mentioning that our Information Loss can be used for different base models without any modification. Figure 4 visualizes four video examples with detected semantic regions to show how our HA_IL works. Sentences generated by baseline model and HA_IL are also provided. It's distinct to see that our model successfully focuses on the video-level salient objects and is able to capture salient details('doing stunts', 'slum dunk', 'cracking'). Especially, although the two men in the example (c) are very small, our visual model attends to them across the entire video. These examples further confirm that our optimized framework and proposed training strategy can jointly capture the gist of the video and generate informative descriptions about given videos.

5 Conclusion

This paper observes that the unbalanced distribution of words is a significant cause of recognition error and detail deficiency in video captioning. To address this issue, we propose a novel strategy called HA_IL. HA_IL uses Information Loss to alleviate the imbalance learning problem and hierarchical attention mechnism to generate video representation. Our HA_IL outperforms state-of-the-art video captioning methods with relative improvements 18.2% in CIDEr on MSVD dataset. What' s more , comparative studies show that our Information Loss is generic for various models.

Acknowledgement

This work was supported by the National key research and development program (2017YFC0 820601), National Nature Science Foundation of China (61571424, U1703261), Beijing science and technology project (Z171100000117010), Beijing Municipal Natural Science Foundation Cooperation Beijing Education Committee: No. KZ 201810005002.

References

[1] 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 VQA. *CoRR*, 2017.

[2] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 2014.

[3] Jianfeng Dong, Xirong Li, Weiyu Lan, Yujia Huo, and Cees GM Snoek. Early embedding and late reranking for video captioning. In *MM*. ACM, 2016.

[4] 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.

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

[6] Haiqi Huang, Yueming Lu, Fangwei Zhang, and Songlin Sun. A multi-modal clustering method for web videos. In *International Conference on Trustworthy Computing and Services*, 2012.

[7] Qin Jin, Jia Chen, Shizhe Chen, Yifan Xiong, and Alexander Hauptmann. Describing videos using multi-modal fusion. In *MM*. ACM, 2016.

[8] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] Xuelong Li, Bin Zhao, and Xiaoqiang Lu. Mam-rnn: Multi-level attention model based rnn for video captionings. *IJCAI*, 2017.

[10] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, 2017.

[11] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

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

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

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[15] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*, 2016.

[16] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. Weakly supervised dense video captioning. In *CVPR*, 2017.

[17] Jingkuan Song, Lianli Gao, Zhao Guo, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. Hierarchical LSTM with adjusted temporal attention for video captioning. In *IJCAI*, 2017.

[18] Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond J. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, 2014.

[19] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3D: generic features for video analysis. *CoRR*, abs/1412.0767, 2014.

[20] Yunbin Tu, Xishan Zhang, Bingtao Liu, and Chenggang Yan. Video description with spatial-temporal attention. In *MM*. ACM, 2017.

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

[22] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *CVPR*, 2017.

[23] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

[24] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

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

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

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

[28] Xishan Zhang, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li, and Qi Tian. Task-driven dynamic fusion: Reducing ambiguity in video description. *CVPR*, 2017.