NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

TRECVID 2022 Video to Text Description (VTT)

Semantic Alignment Network for Video Captioning

Tao Wang, Ping Li*

School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, P. R. China

Introduction

- In this poster, we present our solutions for the Video to Text (VTT) description task in TRECVID 2022.
- Since consecutive frames are likely to contain redundant information, previous methods often simply abandon or merge the redundant frames.
- To solve the above problems, we propose a Semantic Alignment Network (SAN), which attempts to (1) establish a mapping relationship between generated words and video frames by the attention mechanism and then (2) to decode the semantically aligned video frames for predicting the next word.

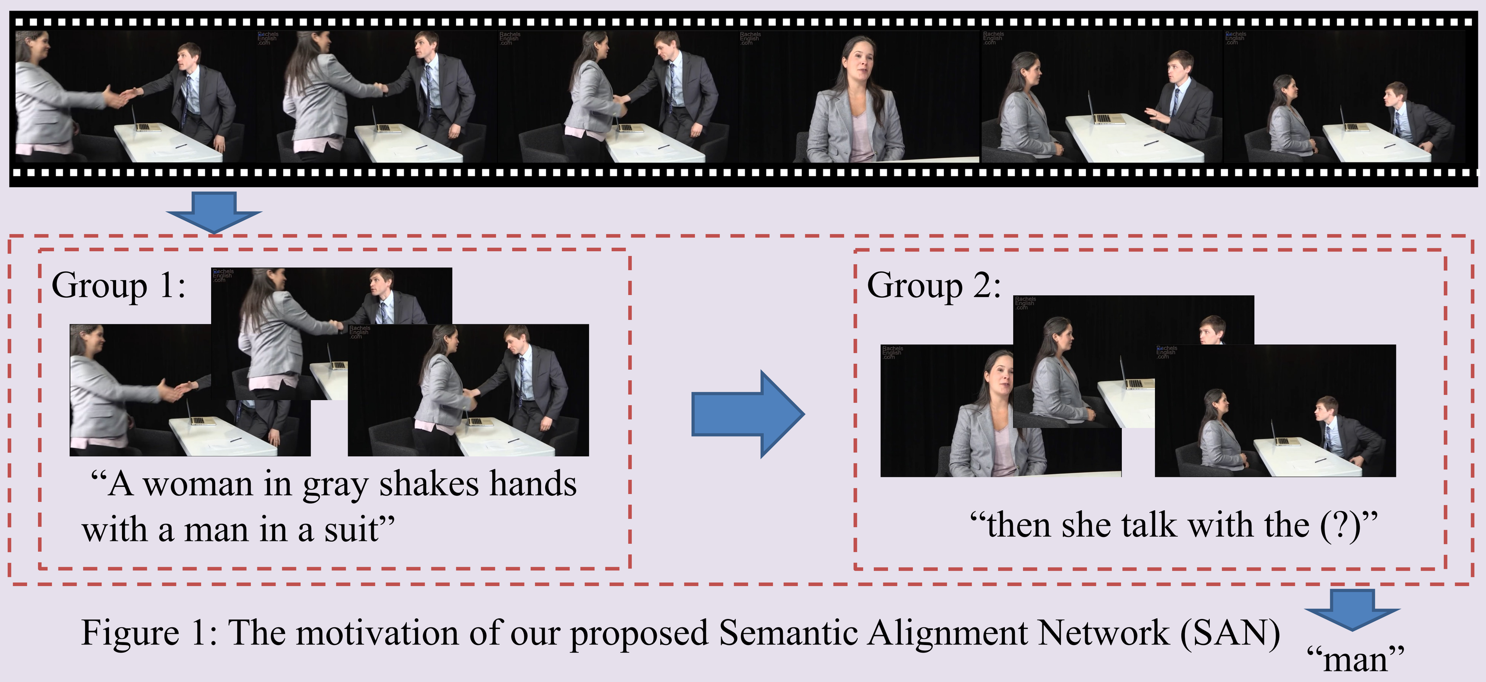

Figure 1: The motivation of our proposed Semantic Alignment Network (SAN)

Framework

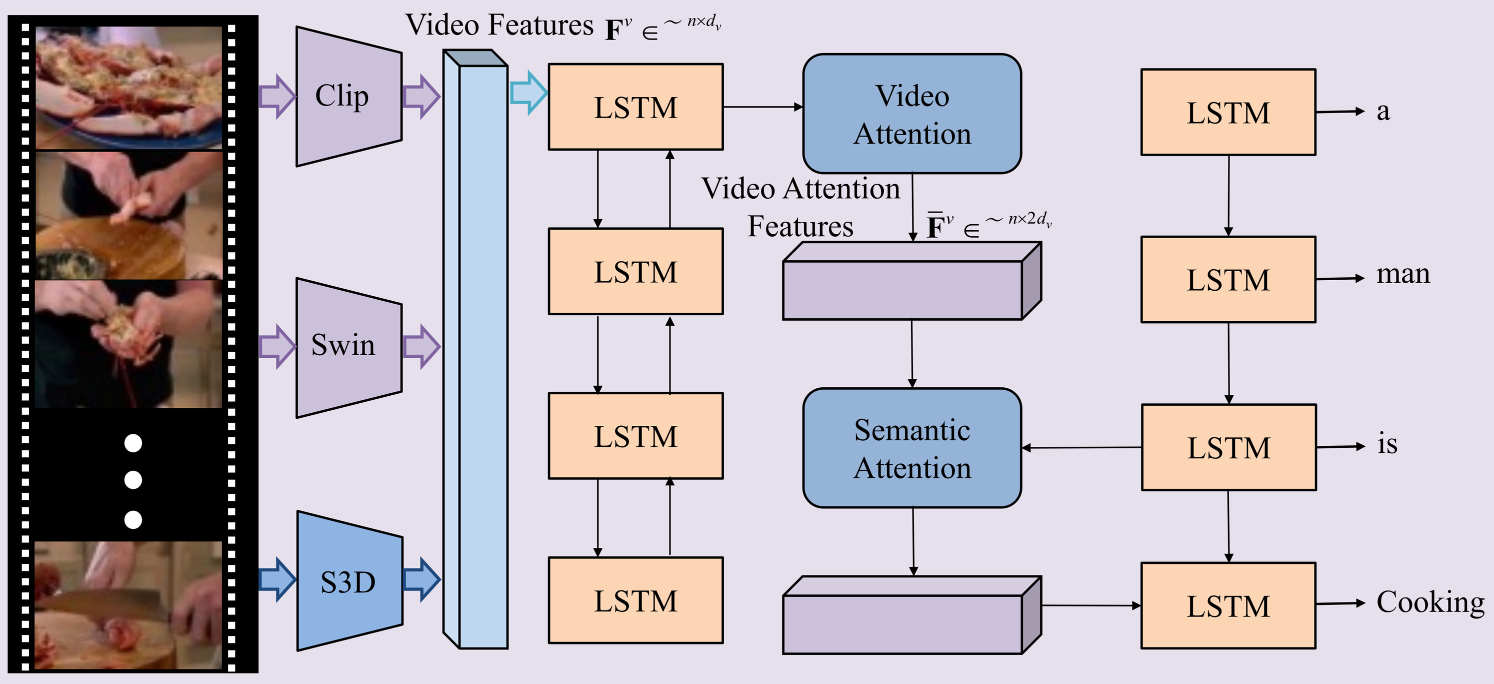

Figure 2: The framework of SAN

SAN consists of three components: (a) Visual Encoder, (b) Semantic Aligner, and (c) Sentences Decoder. Video Encoder uses multiple different models to generate video embedding; Semantic Aligner includes Bi-LSTM, video attention module, and semantic attention module, and builds a mapping between words and video frames.

Visual Encoder

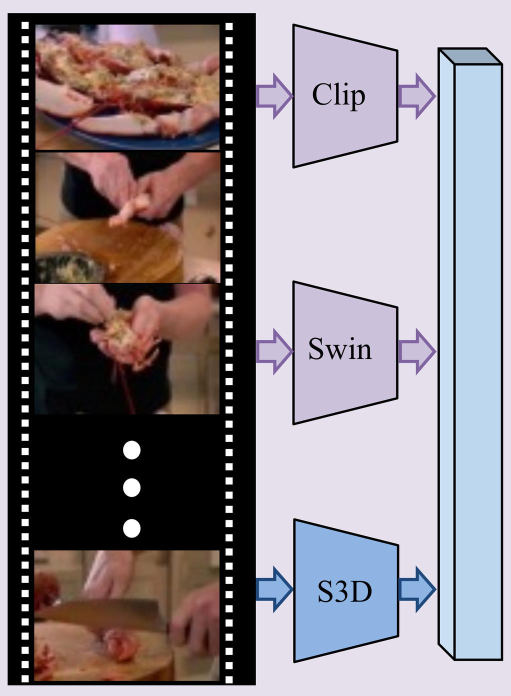

Figure 3: The structure of visual encoder

- Given an input video V including N frames $\{f_i\}_{i=1}^N$, the clips $\{c_i\}_{i=1}^N$ are uniformly sampled and each clip c_i consists of consecutive frames around each sampled frame f_i.
- We adopt the Clip[1], Swin-Transformer[2] and S3D[3] as video encoders to generate video features $F^v \in \mathbb{R}^{n \times d_v}$.

Semantic Aligner

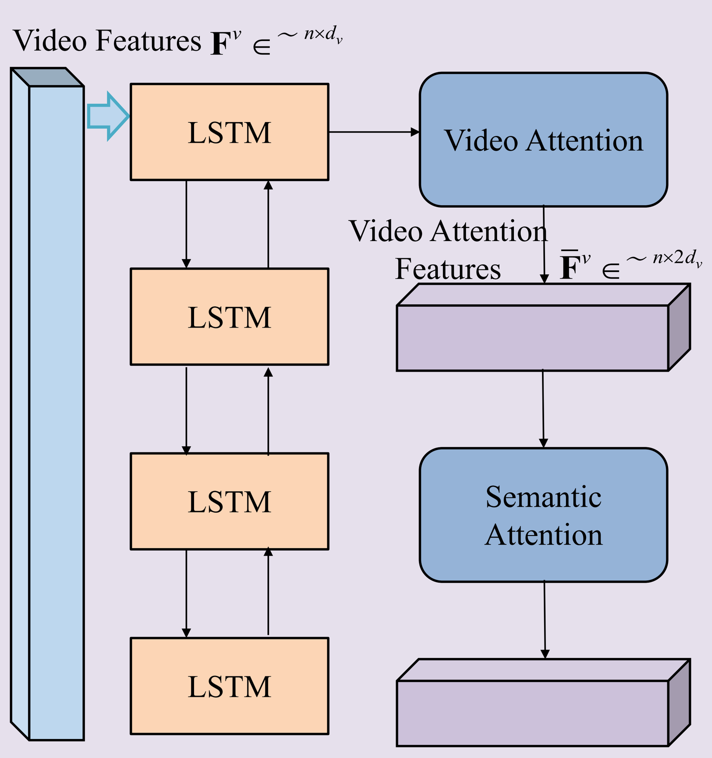

Figure 4: The structure of semantic aligner

- Input the video features $F^v \in \mathbb{R}^{n \times d_v}$ into a Bi-LSTM to get bidirectional video features $\hat{F}^v \in \mathbb{R}^{n \times 2d_v}$.
- Feed the bidirectional video features $\hat{F}^v$ into the video attention module to get video attention features $\bar{F}^v \in \mathbb{R}^{n \times 2d_v}$.

$$\alpha_{i,j} = \sigma\left(\hat{\mathbf{F}}_i^v \hat{\mathbf{F}}_j^v\right)$$

$$\bar{\mathbf{F}}_j^v = \sum_{j=1}^n \alpha_{i,j} \hat{\mathbf{F}}_j^v$$

Sentence Decoder

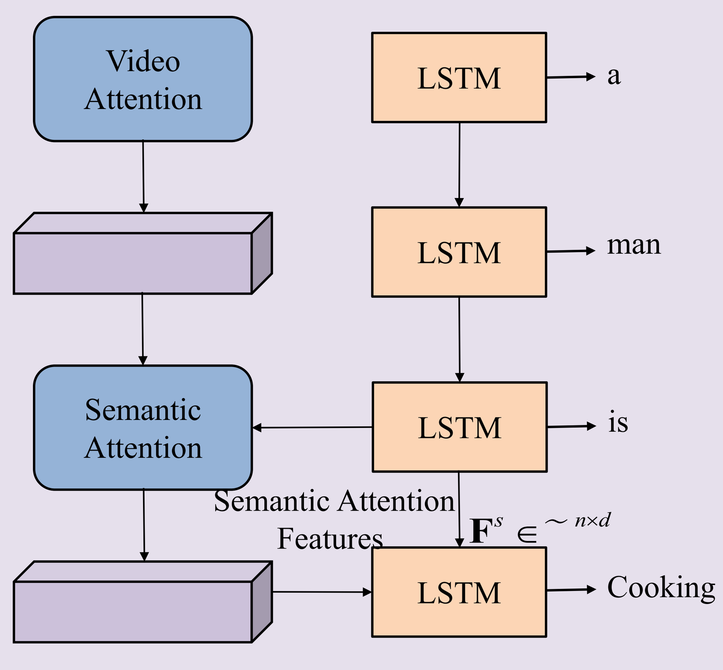

Figure 5: The structure of sentences decoder

- Send the last generated word $w_{t-1} \in \mathbb{R}^{d_s}$ and video attention features $\bar{F}^v$ into the semantic attention module to get semantic attention features $F^s \in \mathbb{R}^{n \times d}$, which builds a mapping between generated words and video frames.

$$\beta_{t-1,j} = \mathbf{u}\sigma\left(Uw_{t-1} + H\bar{\mathbf{F}}_j^v + b\right)$$

$$\mathbf{F}_t^s = \sum_{j=1}^n \beta_{t-1,j} \bar{\mathbf{F}}_j^v$$

- Finally, we get the current word w_t.

Competition Results

Organization(s)	SPICE
Renmin University of China, Tencent	0.184
Ours	0.107
Elyadata	0.102
Waseda University, Meisei University, SoftBank Corporation	0.100
Nagaoka University of Technology	0.097
Carnegie Mellon University	0.077

Organization(s)	BLEU
Renmin University of China, Tencent	0.135
Nagaoka University of Technology	0.081
Ours	0.071
Elyadata	0.069
Waseda University, Meisei University, SoftBank Corporation	0.037
Carnegie Mellon University	0.030

Organization(s)	METEOR
Renmin University of China, Tencent	0.414
Ours	0.290
Waseda University, Meisei University, SoftBank Corporation	0.287
Nagaoka University of Technology	0.281
Elyadata	0.248
Carnegie Mellon University	0.222

- Our model ranks the **second** in terms of **two metrics** (SPICE and METER) and the third in terms of BLEU on TRECVID 2022 VTT tasks.

Conclusion

In summary, we propose a Semantic Alignment Network (SAN) for video captioning in VTT task. It is able to well model the context of captions, by encoding a video into semantic groups. These semantic group consists of the phrases that partially decode the captions and the related frames. More importantly, our method enables building an inherent relationship between generated words and video frames by the attention mechanism. This contributes to generating the satisfying descriptions which are more consistent with video content.

Member Introduction

Tao Wang is currently pursuing the master degree in the School of Computer Science and Technogloy from Hangzhou Dianzi University. His research interests include computer vision.

Ping Li is currently an Associate Professor in computer science at Hangzhou Dianzi University. His research interests include machine learning, computer vision, and data mining.

Contact

Tao Wang
212050308@hdu.edu.cn

Ping Li
lpcs@hdu.edu.cn
*Corresponding author

References: [1] A. Radford, J. W. Kim, C. Hallacy, and et al. Learning transferable visual models from natural language supervision[C]. In Proceedings of the International Conference on Machine Learning (ICML), 2021: 8748-8763
[2] Z. Liu, Y. Lin, Y. Cao, and et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021: 10012-10022.
[3] S. Xie, C. Sun, J. Huang, and et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European conference on computer vision (ECCV). 2018: 305-321.