Deep Learning for Video Captioning: A Review

Shaoxiang Chen[1] , **Ting Yao**[3] and **Yu-Gang Jiang**[1,2*]
[1]Shanghai Key Lab of Intelligent Info. Processing, School of Computer Science, Fudan University, China
[2]Jilian Technology Group (Video++), Shanghai, China
[3]JD AI Research, China
{sxchen13, ygj}@fudan.edu.cn, yaoting5@jd.com

Abstract

Deep learning has achieved great successes in solving specific artificial intelligence problems recently. Substantial progresses are made on Computer Vision (CV) and Natural Language Processing (NLP). As a connection between the two worlds of vision and language, video captioning is the task of producing a natural-language utterance (usually a sentence) that describes the visual content of a video. The task is naturally decomposed into two sub-tasks. One is to encode a video via a thorough understanding and learn visual representation. The other is caption generation, which decodes the learned representation into a sequential sentence, word by word. In this survey, we first formulate the problem of video captioning, then review state-of-the-art methods categorized by their emphasis on vision or language, and followed by a summary of standard datasets and representative approaches. Finally, we highlight the challenges which are not yet fully understood in this task and present future research directions.

1 Introduction

Visual perception and language expression are two key capabilities of human intelligence, and video captioning is a perfect example towards learning from human to bridge vision and language. The goal of video captioning is to automatically describe the visual content of a video with natural language. Practical applications of automatic caption generation include leveraging descriptions for video indexing or retrieval, and helping those with visual impairments by transforming visual signals into information that can be communicated via text-to-speech technology.

Video captioning has already received intensive research attention before the prevalence of deep learning. At the early stage, video captioning approaches [Kojima *et al.*, 2002; Guadarrama *et al.*, 2013] first detect visual concepts in a video with hand-crafted features and then generate the sentence based on pre-defined templates. Such methods highly depend on the templates and the generated sentences are always with fixed syntactical structures, not to mention that the design of hand-crafted features is also bounded for video understanding. Instead, current deep learning based video captioning often performs sequence to sequence learning in an encoder-decoder paradigm. In between, an encoder equipped with powerful deep neural networks is exploited to learn video representation. A decoder of sentence generation is utilized to translate the learned representation into a sentence with more flexible structures.

The learning of video representation is the basis of video understanding, and in general involves both feature extraction and aggregation. The ultimate goal is to extract features from multiple modalities, and then aggregate them spatially and temporally to produce a compact representation. The recent advances in 2D and 3D Convolutional Neural Networks (CNNs) have successfully improved the state-of-the-art of representation learning from visual [He *et al.*, 2016], audio [Hershey *et al.*, 2017] and motion [Tran *et al.*, 2015] information. Nevertheless, feature aggregation particularly for video captioning remains an open challenge. Several techniques from different perspectives, e.g., spatially [Chen and Jiang, 2019a], temporally [Venugopalan *et al.*, 2015] and modality-wise [Xu *et al.*, 2017], have been studied for exploring feature aggregation in video captioning.

The decoder of sentence generation shares the same learning objectives and evaluation metrics with the sequence generation tasks in NLP field such as text summarization and machine translation. As such, challenges, e.g., exposure bias and *objective mismatch* (more details in Sec. 4.2), also exist for the decoder in video captioning due to the recursive nature. Though there are some methods proposed in NLP area , e.g., [Ranzato *et al.*, 2016], to solve the issues, the complexity of video content and relatively small captioning corpus make it difficult if directly applying these solutions to video captioning. Furthermore, considering that videos in real life are usually long, how to recapitulate all the video content that are worthy of mention is still a valid question.

Unlike the existing survey on video captioning in [Aafaq *et al.*, 2018], we comprehensively discuss deep learning based methods in this work. Particularly, we carefully categorize and review state-of-the-art methods, summarize the benchmarks, uniquely highlight the challenges with possible solutions and present future research directions.

[*] Corresponding Author

2 Problem Formulation

Given an input video $V = \{f_1, ..., f_N\}$ (N: the length of frame sequence), the target of video captioning is to generate a sentence (i.e., word sequence) $Y = \{y_1, ..., y_T\}$ to describe the video's content. Thus, video captioning task is often tackled as a problem of sequence-to-sequence learning. Most video captioning frameworks are designed as an *encoder-decoder* structure, where the encoder learns condensed video representation from multi-modal features and the decoder produces sentence word-by-word depending on the learned representation from encoder.

To model the video content, we firstly extract features from multiple modalities: $\mathcal{F} = \{F_V, F_M, F_A, F_S\}$, where F_V, F_M, F_A and F_S denote visual, motion, audio and semantic features respectively.

$$\mathcal{F} = f_{feat}(V), \tag{1}$$

where f_{feat} is an ensemble of *feature extraction* functions (usually pre-trained deep neural networks) for multiple modalities of the video. The features $\mathcal{F}$ may be further aggregated into a more condensed representation, and the process of feature aggregation is conducted depending on some changing state:

$$F_t = f_{aggr}(\mathcal{F}, s_t), \tag{2}$$

where f_{aggr} is the *feature aggregation* function, s_t is an optional state vector (e.g. the model's state when generating the t-th word) and F_t is the aggregated feature. f_{feat} and f_{aggr} constitute the encoder. The *language model* (or decoder) then takes F_t (and optionally s_t) and predicts the distribution of the word y_t:

$$p_t = f_{lang}(F_t, s_t), \tag{3}$$

where f_{lang} is the updating function in LSTM [Hochreiter and Schmidhuber, 1997] or its variants. The final prediction of Y is obtained based on the distributions $\{p_1, ..., p_T\}$.

3 Video Representation

In this section, we review representative methods for video representation learning in the existing literature. The process to obtain video representation can be divided into two major steps: feature extraction (Sec. 3.1) and feature aggregation (Sec. 3.2). These methods are also applicable to other video understanding tasks.

3.1 Multimodal Feature Extraction

This survey mainly focuses on deep feature representations. A good set of features is the foundation of a performant video captioning method. Deep learning has been successfully applied to multiple modalities where sufficient amount of data is available, and the learned representations have nice transferability so that they can be directly leveraged by other tasks.

Visual. Visual appearance is the most important feature for understanding video contents. State-of-the-art convolutional neural networks (CNNs) have surpassed human performance in recognizing images. Activation vectors from higher layers of a trained CNN can capture global visual appearance of its input image, and is now used as the default feature for video captioning. Popular choices of CNN are VGG Net [Simonyan and Zisserman, 2014], ResNet [He *et al.*, 2016] and Inception Networks [Szegedy *et al.*, 2016]. [Song *et al.*, 2017] have shown that a CNN with higher recognition accuracy can further boost video captioning performance.

Motion. Motion feature is crucial for capturing the actions and temporal interactions in video, which complements the static visual appearance. 3D CNN such as C3D [Tran *et al.*, 2015] learns spatiotemporal feature by processing a consecutive sequence of video frames with 3-dimensional convolutions, and can selectively attend to both motion and appearance. Thus, the higher-layer activation vectors of 3D CNN are commonly leveraged as motion feature for video captioning.

Audio. Audio feature is helpful for distinguishing events such as "person talking to the phone" and "person listening to the phone playing music". MFCCs (Mel Frequency Cepstral Coefficients) is a widely adopted audio feature, and video captioning works [Ramanishka *et al.*, 2016; Shen *et al.*, 2017] usually apply Bag-of-Audio-Words [Pancoast and Akbacak, 2014] to obtain a fixed-length audio feature. Recently, [Wang *et al.*, 2018c] demonstrates that sound representation learned by CNN such as VGGish [Hershey *et al.*, 2017] is more effective than MFCCs for video captioning.

Semantic. Semantic feature refers to a wide category of features that explicitly capture semantic contents in videos. MMVD [Ramanishka *et al.*, 2016] shows that the video-level category information can boost video captioning. Simply incorporating category information into the encoder can yield better captioning performance. M&M TGM [Chen *et al.*, 2017] further predicts latent topics from multimodal features (except semantic feature), then integrates the predicted topics into the designed topic-aware decoder. LSTM-TSA [Pan *et al.*, 2017] adopts the weakly-supervised attribute detection method of [Fang *et al.*, 2015] to detect frame- and video-level fine-grained attributes. Next a transfer unit is utilized to dynamically incorporate attribute information into LSTM-based decoder. In this sense, semantic features of any granularity can improve video captioning, which is because they provide the decoder (language model) with more prior knowledge about the video content.

3.2 Feature Aggregation Is Important

Video features are often extracted from multiple modalities and are usually sequences of variable length as the video. Feature aggregation is a common way to aggregate them into a fixed-length representation because large amount of video features lead to (1) high computational cost which is beyond current GPU's capability and (2) more parameters that result in overfitting easier.

Temporal Attention. The simplest way to aggregate a feature sequence, as in [Venugopalan *et al.*, 2015], is using a LSTM/GRU to encode the sequence and take the final encoding state as the aggregated feature for decoding. However, treating video features as a flat sequence is not effective, because (1) the length of gradient flow to the earliest frame is as long as the sequence, which leads to gradient vanishing; (2) each feature in the sequence contributes the same to the decoder, which makes the model also pay attention to background noises. Temporal attention (also known as *dynamic attention*) [Yao *et al.*, 2015], is a mechanism which learns to dynamically assign weights to each feature in the sequence such that the decoder can pay more attention to relevant features when generating certain words. Thus the computation of attention weights involves both visual feature sequence and the decoder state. Another effect is that the decoder and each feature is directly connected by a weighted path, which shortens the length of gradient flow and leads to more effective learning. hLSTMat [Song *et al.*, 2017] is an improved temporal attention mechanism which makes the decoder depend less on visual features when generating non-visual words, but instead rely on language model's state.

Spatial Attention. Different regions of the video frames also contribute differently to the final word prediction, e.g. objects are clearly more important than background. Spatial attention methods aim to learn spatial attention maps, which indicate the importance of different regions. Dynamic attention can also be applied spatially if regions are treated sequentially. Thus, MAM-RNN [Li *et al.*, 2017] adopts two-level spatial and temporal dynamic attention for video captioning. When computing spatial attention weights for a certain frame, MAM-RNN additionally incorporates the attention weights from previous frame. In this way, the spatial attention maps are linked across time. SAM [Wang *et al.*, 2018a] tries to learn a model that distinguishes foreground from background in videos with out explicit supervision. Saliency scores are computed from the spatial feature map to separate foreground and background according to a threshold, which results in two maps representing foreground and background scores. The two maps are then aggregated as foreground and background context input to the language model. Learning spatial attention without any supervision or guidance is hard. MGSA (Motion Guided Spatial Attention) [Chen and Jiang, 2019a] uses optical flow to capture motion information in videos and computes spatial attention map based on optical flow images. The motivation is the fact that human attention is more likely to be drawn to the rapidly changing areas.

Multimodal Feature Fusion. Using multimodal features is ubiquitous in video captioning methods, in contrast, multimodal feature fusion strategy is rarely explored. MMVD [Ramanishka *et al.*, 2016] simply concatenates features from multiple modalities as the input to decoder. It is obvious that the importance of each modality is different for various types of videos. Therefore, Attention Fusion [Hori *et al.*, 2017] and MA-LSTM [Xu *et al.*, 2017] independently proposed similar multimodal attention mechanisms. They apply dynamic attention to different streams (visual, motion and audio) of features after they are individually aggregated by temporal attention, allowing each modality to contribute differently to caption generation.

4 Caption Generation

Given the generated word probabilities at each time step $\{p_1, ..., p_T\}$ and ground truth caption $\hat{Y} = \{\hat{y}_1, ..., \hat{y}_T\}$, the most common learning objective for captioning is to maximize the log-likelihood of all the ground truth words:

$$\max_{\theta} \sum\nolimits_{t=1}^{T} p_t(\hat{y}_t), \tag{4}$$

where θ is all the learnable parameters of the captioning model. This objective is widely adopted for sequence generation tasks such as machine translation and captioning. However, there are two major problems with it. First, there is a discrepancy between this objective function and the automatic evaluation metrics such as BLEU [Papineni *et al.*, 2002]. This is often referred to as *objective mismatch*. And there is also a gap between these metrics and human judgment. Second, this objective alone maybe insufficient to train a good language model since video captioning datasets have a much smaller corpus compared to pure NLP datasets.

In the rest of this section we review methods which emphasize caption generation (the decoder) and aim at addressing the above issues, including: (1) Semantic supervision, which designs auxiliary objectives to exploit visual semantic concepts to improve captioning quality; (2) Approaches to mitigate the objective mismatch problem; (3) Dense captioning, which requires jointly localizing and describing multiple events in a video. Note that some methods are for image captioning, but are applicable to video captioning as well.

4.1 Auxiliary Semantic Supervision

One straightforward way to exploit visual concepts and improve captioning quality is to make sentence semantics consistent with visual contents by enforcing such consistency constraint. Representative methods [Pan *et al.*, 2016; Gao *et al.*, 2017b] project the encoded visual feature vector and the averaged sentence embedding vector onto a common space and then add an optimization term to minimize their distance, which is jointly optimized with the captioning objective. However, there is a tradeoff on the strength of consistency constraint in these methods, which needs to be carefully tuned by human. Instead, Semantic Attention (SA) [You *et al.*, 2016] exploits semantic concepts by making the model's attention cover all the semantic concepts in an image. SA first detects semantic concepts in image and then applies dynamic attention upon the concepts at each word generation step. A regularization term is added to enforce the completeness of attention paid to all concepts and the sparsity of attention at any particular time step. Intuitively, SA learns to fully exploit fine-grained visual information as well as focus on a specific one at each step. SA uses nearest neighbor search in a large dataset to retrieve visual concepts from similar images, so its capability may be limited by the dataset, especially when applied to videos. In contrast, M&M TGM [Chen *et al.*, 2017] uses predicted semantic topics to guide the learning of video captioning model. A topic mining module first mines topics from the training descriptions by clustering, then it is used as the teacher to train a topic predictor. For caption generation, the predicted topics are fed to an extended LSTM decoder with a set of topic-dependent weight matrices, which works as an ensemble of several topic-aware decoders. The captioning and topic prediction objectives are jointly optimized by multitask training.

Dataset	Domain	No of Videos	No of Clips	Duration (hrs)	No of Sent.	Avg Word	Vocab. Size	Temp. Anno.
MSVD [Chen and Dolan, 2011]	Open	1,970	1,970	5.3	70,028	8.7	13,010	✓
TACoS [Regneri *et al.*, 2013]	Cooking	127	3,290	10.1	18,818	9.0	1,413	✓
YouCook [Das *et al.*, 2013]	Cooking	88	-	2.3	3,502	12.6	2,329	×
TACoS-multilevel [Rohrbach *et al.*, 2014]	Cooking	185	14,105	15.7	52,593	8.3	2,864	×
MPII-MD [Rohrbach *et al.*, 2015]	Movie	94	68,337	73.6	68,375	9.6	24,549	✓
M-VAD [Torabi *et al.*, 2015]	Movie	92	48,986	84.6	55,904	9.3	18,269	✓
MSR-VTT [Xu *et al.*, 2016]	Open	7,180	10,000	41.2	200,000	9.3	29,316	×
TGIF [Li *et al.*, 2016]	Open	100,000	-	86.1	125,781	10.6	11,806	×
ActivityNet Captions [Krishna *et al.*, 2017]	Open	20,000	73,000	849.0	73,000	13.5	10,646	✓
DiDeMo [Hendricks *et al.*, 2017]	Open	10,464	26,892	144.2	41,206	7.5	7,587	✓

Table 1: Standard datasets for evaluating video captioning methods. **Avg Word** means average number of words per sentence. **Temp. Anno.** stands for temporal annotation.

4.2 Addressing Objective Mismatch

The cause of the objective mismatch problem is that the computation of sequence-level evaluation metrics such as BLEU [Papineni *et al.*, 2002], is not differentiable. Thus, they can't be directly optimized by back-propagation and gradient descent methods like normal objective functions such as Eq. 4. This inconsistency between objective functions and evaluation metrics is a common issue of language generation tasks, such as machine translation [Ranzato *et al.*, 2016].

Current solutions are based on REINFORCE [Williams, 1992], which is a class of reinforcement learning algorithms that can optimize any metric of interest through maximizing the expected *reward* of model samples and trains on sampled sequences by using policy gradients. Self-Critical Sequence Training (SCST) [Rennie *et al.*, 2017] is a form of REINFORCE that chooses CIDEr [Vedantam *et al.*, 2015] as the reward signal and utilizes the greedy output of the model as the baseline to reduce variance. SCST achieved significant improvement in terms of CIDEr score, at the cost of having to evaluate baseline sequences at every step. Other than directly optimizing for the evaluation metrics, CIDEnt [Pasunuru and Bansal, 2017] adopts a novel entailment-corrected reward, which combines a learned entailment score function to correct the phrase-matching metrics like CIDEr. CIDEnt improves the logical correctness of the generated captions and performs better than just optimizing CIDEr.

Efforts are also made to close the gap between evaluation metrics and human judgment. CIDEr measures the similarity of a generated caption against a small set of ground truth (reference) sentences written by humans, and it is shown to capture human judgment of consensus better than previous metrics. Most recently, the SPICE metric [Anderson *et al.*, 2016] is proposed to compare generated sentence and reference sentences from a semantic similarity perspective by parsing them into a scene graph representation. While SPICE ignores the syntactic quality of sentences, it outperforms CIDEr in terms of capturing human judgments, and they can be combined to get better captioning quality [Liu *et al.*, 2017].

4.3 Dense Captioning

A long video might contain multiple events, a practical video captioning system should jointly localize and describe each event. This task is known as dense captioning and is clearly more challenging. h-RNN [Yu *et al.*, 2016] is an early attempt that generates paragraph to describe a video by using hierarchical RNN. The sentence generator RNN takes video feature and paragraph state as input, and applies temporal attention to generate sentences. The paragraph state is produced by another RNN, which recurrently takes previous sentence embedding to update current paragraph state. However, the generated sentences are not localized. Another attempt [Shen *et al.*, 2017] is to generate dense video captions that are spatially localized. In the absence of spatial annotation, Shen *et al.* first adopted multiple instance learning to detect semantic concepts in video frames, and then selected spatial region sequences using submodular maximization with the objective to maximize informativeness, coherence and diversity within each sequence. The region sequences are then individually described by a LSTM-based language model.

Krishna *et al.* constructed the ActivityNet Captions [Krishna *et al.*, 2017] dataset and annotated multiple temporally localized sentences per video. They first proposed the dense event captioning task with a benchmark and a baseline method. The evaluation for dense captioning is based on proposals, i.e. temporally localized video segments. In the baseline, proposals that might contain events are first generated by a variant of action proposal method (DAP [Escorcia *et al.*, 2016]). Representations of the proposals are fed to a LSTM for caption generation, and attention is used to incorporate contexts for the proposals. Following works generally used variants of action proposal methods for event proposal. [Li *et al.*, 2018] proposed a dense captioning framework that consists of two parts: a temporal event proposal (TEP) module and a sentence generation (SG) module. For the TEP module, a convoultional architecture like [Lin *et al.*, 2017] is adopted to perform event/background classification, temporal boundaries refinement and descriptiveness regression for each proposal. The refined proposals and their visual attributes are fed to the SG module, which contains a LSTM network. Reinforcement learning is used to train the SG module to maximize METEOR scores. Bi-SST [Wang *et al.*, 2018b] adapted SST [Buch *et al.*, 2017] to generate event proposals. Bi-SST contains a bidirectional event proposal module which exploits both past and future context for proposal prediction. The contexts are obtained by encoding visual features with LSTM in both directions, and are then combined with visual features as input to the caption generation module. Furthermore, a context gating mechanism is designed to balance the contributions of past and future contexts. WS-DEC [Duan *et al.*, 2018] tackles dense captioning without temporal annotations, in which event proposal generation can't be trained under strong supervision. The problem is decomposed into a cycle of dual problems: caption generation and sentence localization. WS-DEC tries to reconstruct the ground truth caption by first localizing it and then generating caption based on the localized segment. The whole model is trained by minimizing the reconstruction error. Surprisingly, the results of WS-DEC are comparable to supervised methods [Krishna *et al.*, 2017]. One key part of this work, sentence localization in video, has also attracted great research attention from both the computer vision [Gao *et al.*, 2017a; Chen and Jiang, 2019b] and natural language processing [Chen *et al.*, 2018b] communities recently. The success of WS-DEC will shed some light upon utilizing sentence localization in dense captioning.

5 Datasets and Evaluation

In this section, we summarize standard video captioning datasets and evaluate representative methods.

5.1 Datasets

The early datasets mainly come from specific domains like cooking and movie, since these videos were easier to obtain. For TACoS [Regneri *et al.*, 2013], YouCook [Das *et al.*, 2013] and TACoS-multilevel [Rohrbach *et al.*, 2014], the sentences are descriptions about a person's cooking procedures. Their vocabularies as well as amount of data are very limited. MPII-MD [Rohrbach *et al.*, 2015] and M-VAD [Torabi *et al.*, 2015] are created from audio descriptions for movies. Both the number of clips and vocabulary size are larger compared to cooking datasets. The first open domain dataset, MSVD [Chen and Dolan, 2011] contains web videos from different categories but has limited size (1,970 clips).

MSR-VTT [Xu *et al.*, 2016] is the first large-scale open domain dataset. It contains 10,000 clips from 20 categories including music, sports and movie. Each clip is associated with 20 human annotations. MSR-VTT has the most descriptions and largest vocabulary among all the datasets. TGIF [Li *et al.*, 2016] has 100K animated GIFs from Tumblr and 120K sentence descriptions. The duration of each GIF is around 3.1 seconds and each GIF is described by one sentence. The visual contents of TGIF are more diverse.

Recently, the temporal localization of sentences are further emphasized in several datasets. For instance, ActivityNet Captions [Krishna *et al.*, 2017] is constructed specifically for dense event captioning task. This dataset contains 20,000 videos in total. Each video lasts 150 seconds and contains 3.65 temporally localized sentences in average. DiDeMo [Hendricks *et al.*, 2017] aims at localizing sentence in video. It consists of 10,464 long videos (25-30 seconds per video) and 41,206 localized sentence descriptions. Each long video is broken into 5-second segments, so a 30-second video contains 21 possible intervals.

Method	T	B@4	M	C
MMVD [Ramanishka *et al.*, 2016]	M	40.7	28.6	46.5
Attention Fusion [Hori *et al.*, 2017]	M	39.7	25.5	40.0
MA-LSTM [Xu *et al.*, 2017]	M	36.5	26.5	41.0
HACA [Wang *et al.*, 2018c]	M	43.4	**29.5**	49.7
Temporal Att. [Yao *et al.*, 2015]	A	34.8	25.1	36.7
hLSTMat [Song *et al.*, 2017]	A	38.3	26.3	-
MGSA [Chen and Jiang, 2019a]	A	**45.4**	28.6	50.1
LSTM-E [Pan *et al.*, 2016]	S	36.1	25.8	38.5
M&M TGM [Chen *et al.*, 2017]	S	44.3	29.4	49.3
RL Ent [Pasunuru and Bansal, 2017]	R	40.5	28.4	**51.7**

Table 2: Performance of video captioning methods on MSR-VTT. The **T** column means their emphasis, where **M**, **A**, **S** and **R** stand for multimodal features, feature aggregation, semantic supervision and reinforcement learning, respectively.

Method	Prop.	B@4	M	C
DCE [Krishna *et al.*, 2017]	DAP	2.20	4.82	17.29
JLDE [Li *et al.*, 2018]	SSAD	0.73	6.93	12.61
Bi-SST [Wang *et al.*, 2018b]	SST	**2.30**	**9.60**	12.68
WS-DEC [Duan *et al.*, 2018]	N/A	1.27	6.30	**18.77**

Table 3: Performance of dense event captioning methods on ActivityNet Captions. Prop. stands for event proposal method.

5.2 Evaluation

We summarize the evaluation results of representative video captioning methods on the MSR-VTT dataset in Table 2. The BLEU@4 [Papineni *et al.*, 2002], METEOR [Lavie and Agarwal, 2007] and CIDEr scores are reported. We group the methods based on their emphasis. HACA, MGSA, M&M TGM and RL Ent are the top performing methods from each category and are orthogonal to each other. In terms of CIDEr score, RL Ent outperforms others by a clear margin since it is directly optimized for this metric. The performances of the dense video captioning methods are depicted in Table 3. We additionally indicate the adopted event proposal methods in the dense video captioning systems. Overall, Bi-SST exhibits the best METEOR score on ActivityNet Captions. Surprisingly, WS-DEC has achieved comparable results with the others despite that it is weakly supervised. The detailed analysis of these results can be referred in the original papers.

6 Conclusion and Future Directions

This paper reviews recent deep learning based video captioning methods, and discusses several important topics related to both computer vision and natural language processing. We also summarize the benchmarks and provide performance comparisons between the representative approaches. Though extensive efforts have been made on video captioning with deep learning, there are still several open challenges.

Modeling object interaction. Currently, the interpretability of video captioning methods is mainly derived from the spatial and temporal attention mechanism, which explains the importance of each spatial/temporal region with respect to the generated word. However, in complex videos, there are multiple interactions and visual relationships between the objects, which are hard to be fully captured by spatial and temporal attention. Similar to the observations in [Yao *et al.*, 2018] that modeling object relations is helpful for image captioning, it is also crucial to exploit both spatial and temporal object interaction for video captioning.

Improving event proposal. Existing dense captioning approaches utilize variants of action proposal algorithms for event proposal generation. Nevertheless, events in this task are significantly more complex than actions. Hence how to leverage finer-grained information (such as visual concepts) for producing event proposals is vital. Encouraging results have been obtained by doing so for the closely related sentence localization problem [Chen and Jiang, 2019b].

Novel decoder structures. LSTMs have been the most common choice to build the decoder for video captioning. However, due to its complex gating mechanisms, it is hard to determine how much visual information has contributed to the generation of a certain word. This inevitably impedes the exploration of visual representation learning methods. Recently, non-recurrent models [Gehring *et al.*, 2017; Vaswani *et al.*, 2017] have demonstrated their potential for sequence modeling in NLP field. As such, using non-recurrent decoders (e.g., convolutional decoder [Chen *et al.*, 2019]) may be helpful for addressing this issue.

We haven't reached the performance upper bound yet, especially for dense video captioning [Ghanem *et al.*, 2017; Ghanem *et al.*, 2018]. The next big leap of performance improvement can come from video representation, event localization and language generation, and above we have suggested one concrete problem for each topic. Other than improving performance, there are several interesting directions that are also crucial to the application of video captioning: 1) The ability to describe unseen objects, which may be extended by incorporating external knowledge [Yao *et al.*, 2017]. 2) More interpretable models, which would require the support of datasets annotated in greater detail [Zhou *et al.*, 2018]. 3) Robustness, captioning models are also vulnerable to adversarial examples [Chen *et al.*, 2018a] like other deep learning based visual recognition models.

References

[Aafaq *et al.*, 2018] Nayyer Aafaq, Syed Zulqarnain Gilani, Wei Liu, and Ajmal Mian. Video description: A survey of methods, datasets and evaluation metrics. *arXiv preprint arXiv:1806.00186*, 2018.

[Anderson *et al.*, 2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In *ECCV*, pages 382–398, 2016.

[Buch *et al.*, 2017] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. SST: single-stream temporal action proposals. In *CVPR*, pages 6373–6382, 2017.

[Chen and Dolan, 2011] David Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL-HLT*, pages 190–200, 2011.

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

[Chen and Jiang, 2019b] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, 2019.

[Chen *et al.*, 2017] Shizhe Chen, Jia Chen, Qin Jin, and Alexander G. Hauptmann. Video captioning with guidance of multimodal latent topics. In *ACM MM*, pages 1838–1846, 2017.

[Chen *et al.*, 2018a] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In *ACL*, pages 2587–2597, 2018.

[Chen *et al.*, 2018b] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, pages 162–171, 2018.

[Chen *et al.*, 2019] Jingwen Chen, Yingwei Pan, Yehao Li, Ting Yao, Hongyang Chao, and Tao Mei. Temporal deformable convolutional encoder-decoder networks for video captioning. In *AAAI*, 2019.

[Das *et al.*, 2013] Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, pages 2634–2641, 2013.

[Duan *et al.*, 2018] Xuguang Duan, Wen-bing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *NeurIPS*, pages 3063–3073, 2018.

[Escorcia *et al.*, 2016] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *ECCV*, pages 768–784, 2016.

[Fang *et al.*, 2015] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015.

[Gao *et al.*, 2017a] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *ICCV*, pages 5277–5285, 2017.

[Gao *et al.*, 2017b] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based LSTM and semantic consistency. *IEEE Trans. Multimedia*, 19(9):2045–2055, 2017.

[Gehring *et al.*, 2017] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In *ICML*, pages 1243–1252, 2017.

[Ghanem *et al.*, 2017] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam

Alwassel, Ranjay Krishna, Victor Escorcia, Kenji Hata, and Shyamal Buch. Activitynet challenge 2017 summary. *arXiv preprint arXiv:1710.08011*, 2017.

[Ghanem *et al.*, 2018] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, Victor Escorcia, Ranjay Krishna, Shyamal Buch, and Cuong Duc Dao. The activitynet large-scale activity recognition challenge 2018 summary. *arXiv preprint arXiv:1808.03766*, 2018.

[Guadarrama *et al.*, 2013] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, pages 2712–2719, 2013.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hendricks *et al.*, 2017] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5804–5813, 2017.

[Hershey *et al.*, 2017] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. CNN architectures for large-scale audio classification. In *ICASSP*, pages 131–135, 2017.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Hori *et al.*, 2017] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *ICCV*, pages 4203–4212, 2017.

[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. *International Journal of Computer Vision*, 50(2):171–184, 2002.

[Krishna *et al.*, 2017] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017.

[Lavie and Agarwal, 2007] Alon Lavie and Abhaya Agarwal. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *WMT@ACL*, pages 228–231, 2007.

[Li *et al.*, 2016] Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. TGIF: A new dataset and benchmark on animated GIF description. In *CVPR*, pages 4641–4650, 2016.

[Li *et al.*, 2017] Xuelong Li, Bin Zhao, and Xiaoqiang Lu. MAM-RNN: multi-level attention model based RNN for video captioning. In *IJCAI*, pages 2208–2214, 2017.

[Li *et al.*, 2018] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, pages 7492–7500, 2018.

[Lin *et al.*, 2017] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *ACM MM*, pages 988–996, 2017.

[Liu *et al.*, 2017] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, pages 873–881, 2017.

[Pan *et al.*, 2016] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, pages 4594–4602, 2016.

[Pan *et al.*, 2017] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, pages 984–992, 2017.

[Pancoast and Akbacak, 2014] Stephanie Pancoast and Murat Akbacak. Softening quantization in bag-of-audio-words. In *ICASSP*, pages 1370–1374, 2014.

[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*, pages 311–318, 2002.

[Pasunuru and Bansal, 2017] Ramakanth Pasunuru and Mohit Bansal. Reinforced video captioning with entailment rewards. In *EMNLP*, pages 979–985, 2017.

[Ramanishka *et al.*, 2016] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. Multimodal video description. In *ACM MM*, pages 1092–1096, 2016.

[Ranzato *et al.*, 2016] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016.

[Regneri *et al.*, 2013] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

[Rennie *et al.*, 2017] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 1179–1195, 2017.

[Rohrbach *et al.*, 2014] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *GCPR*, pages 184–195, 2014.

[Rohrbach *et al.*, 2015] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, pages 3202–3212, 2015.

[Shen *et al.*, 2017] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. Weakly supervised dense video captioning. In *CVPR*, pages 5159–5167, 2017.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Song *et al.*, 2017] 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*, pages 2737–2743, 2017.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[Torabi *et al.*, 2015] Atousa Torabi, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015.

[Tran *et al.*, 2015] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 6000–6010, 2017.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.

[Venugopalan *et al.*, 2015] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence - video to text. In *ICCV*, pages 4534–4542, 2015.

[Wang *et al.*, 2018a] Huiyun Wang, Youjiang Xu, and Yahong Han. Spotting and aggregating salient regions for video captioning. In *ACM MM*, pages 1519–1526, 2018.

[Wang *et al.*, 2018b] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, pages 7190–7198, 2018.

[Wang *et al.*, 2018c] Xin Wang, Yuan-Fang Wang, and William Yang Wang. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. In *NAACL-HLT*, pages 795–801, 2018.

[Williams, 1992] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

[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*, pages 5288–5296, 2016.

[Xu *et al.*, 2017] Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei. Learning multimodal attention LSTM networks for video captioning. In *ACM MM*, pages 537–545, 2017.

[Yao *et al.*, 2015] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015.

[Yao *et al.*, 2017] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*, pages 5263–5271, 2017.

[Yao *et al.*, 2018] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, pages 684–699, 2018.

[You *et al.*, 2016] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016.

[Yu *et al.*, 2016] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, pages 4584–4593, 2016.

[Zhou *et al.*, 2018] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. Grounded video description. *arXiv preprint arXiv:1812.06587*, 2018.