A Dataset for Telling the Stories of Social Media Videos

Spandana Gella[1]* **Mike Lewis**[2] **Marcus Rohrbach**[2]
[1]University of Edinburgh, [2]Facebook AI Research
spandana.gella@ed.ac.uk, {mikelewis,mrf}@fb.com

Abstract

Video content on social media platforms constitutes a major part of the communication between people, as it allows everyone to share their stories. However, if someone is unable to consume video, either due to a disability or network bandwidth, this severely limits their participation and communication. Automatically telling the stories using multi-sentence descriptions of videos would allow bridging this gap. To learn and evaluate such models, we introduce VideoStory, a new large-scale dataset for video description as a new challenge for multi-sentence video description. Our VideoStory captions dataset is complementary to prior work and contains 20k videos posted publicly on a social media platform amounting to 396 hours of video with 123k sentences, temporally aligned to the video.

1 Introduction

Telling stories about what we experience is a central part of human communication (Mateas and Sengers, 2003). Increasingly, stories about our experiences are captured in the form of videos and then shared on social media platforms. One goal of automatically understanding and describing such videos with natural language is to generate multi-sentence descriptions which convey the story, making them accessible to situationally (e.g. bandwidth) or physically ("blind") disabled people. However, it is still a challenge for vision and language models to automatically encode and describe temporal content in videos with multi-sentence descriptions (Rohrbach et al., 2014; Zhou et al., 2018b). To better understand the stories shared on social media we collect and annotate a novel dataset consisting of videos from a social media platform. Importantly, we collect descriptions containing multiple sentences, as single sentences would typically not be able to capture the narration and plot of the video.

We introduce a large-scale multi-sentence description dataset for videos. To build a dataset of high quality, diverse and narratively interesting videos, we choose videos that had high engagement on a social media platform. Existing video captioning datasets, such as ActivityNet Captions (Krishna et al., 2017) or cooking video datasets (Regneri et al., 2013; Zhou et al., 2018a), have focused on sets of pre-selected human activities, whereas social media videos contain a great diversity of topics. Videos with high engagement tend to be narratively interesting, because humans find very predictable videos less enjoyable, meaning that captioning of the videos accurately requires integrating information from the entire video to describe a sequence of events (see Figure 1). Together, this creates a diverse and challenging new benchmark for video and language understanding.

We present a thorough analysis of the new benchmark, demonstrating that linguistic and video context is crucial to accurate captioning and that the captions have a temporal consistency. We also show baseline results using state-of-the-art models.

2 Multi-Sentence VideoStory Dataset

In Table 1 we summarize existing video description datasets; most provide only single-sentence descriptions or are restricted to narrow domains. Other multi-sentence description datasets are proposed for story narration of sets of images taken from a Flickr album (Huang et al., 2016; Krause et al., 2017). Other related work includes visual summarization of Flickr photo albums (Sigurdsson et al., 2016a) or videos (De Avila et al., 2011; Zhang et al., 2016) where the idea is to pick the key images or frames that summarize the visual content.

*Work done while SG was intern at Facebook AI Research.

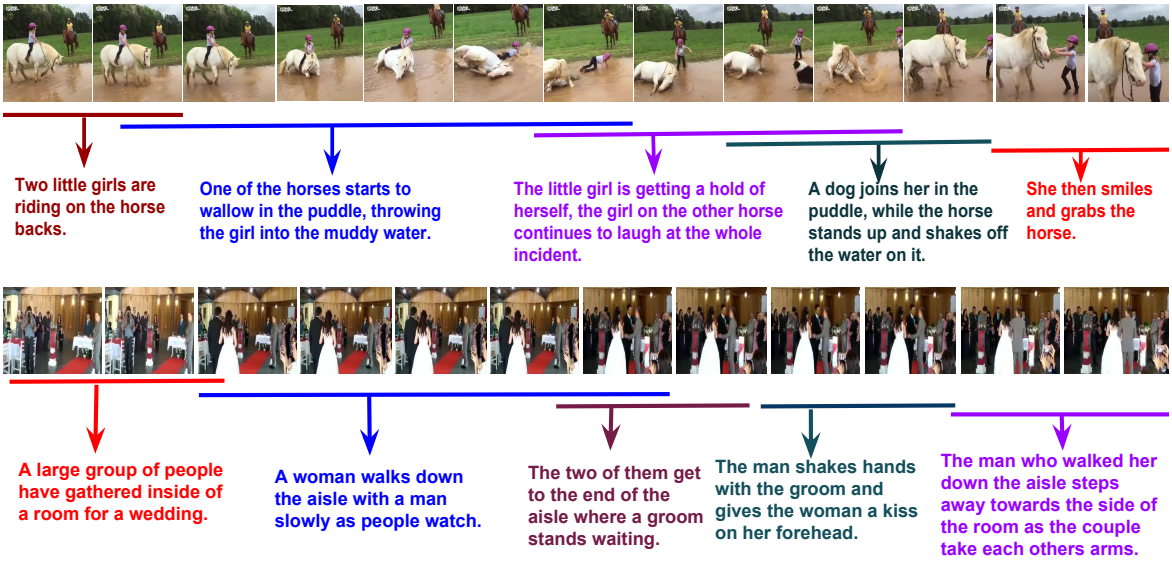

Figure 1: Example videos and multi-sentence description in our VideoStory Dataset showing temporally, overlapping time alignments. Each segment has time boundaries annotated and is described by a sentence.

Dataset	Domain	# videos:clips	Avg.D	#ActL	#sent	Loc	multi-sent.	overlap
MSVD (Chen and Dolan, 2011)	human	:2k	10s	-	70k	✓	-	-
MSR-VTT (Xu et al., 2016)	open	7k:10k	20s	-	200k	✓	-	-
Charades (Sigurdsson et al., 2016b)	human	10k:	30s	157	16.1k	-	✓	-
YouCook (Das et al., 2013)	cooking	88:-	95s		2.7k	✓	-	-
VTW (Zeng et al., 2016)	open	18k:-	90s	-	45k	-	✓	-
TGIF (Li et al., 2016)	open	:100k	3s	-	128k	✓	-	-
MPII MD (Rohrbach et al., 2015)	movie	94:68k	4s	-	68.3k	✓	(✓)	(✓)
M-VAD (Torabi et al., 2015)	movie	92:46k	6s	-	55.9k	✓	(✓)	(✓)
LSMDC (Rohrbach et al., 2017)	movie	200:128k	4s	-	128.1k	✓	(✓)	(✓)
TACoS (Regneri et al., 2013)	cooking	127:3.5k	286s:		11.8k	✓	✓	-
TACos multi-level (Rohrbach et al., 2014)	cooking	185:25k	307s:	67	75k	✓	✓	-
Youcook II (Zhou et al., 2018a)	cooking	2k:15.4	316:19.6s	-	15.4k	✓	✓	-
ActivityNet Captions (Krishna et al., 2017)	human activity	20k:100k	180:36s	203	100k	✓	✓	✓
VideoStory (Ours)	social media	20k:123k	70:18s	-	123k	✓	✓	✓

Table 1: Comparison of our dataset with other video description datasets. Avg.D: Average duration of the video/clip. #ActL: No. of action labels. Loc: temporally localized language descriptions; multi-sent: multi-sentence descriptions; overlap: allows overlap among segments. (✓) indicates datasets with multiple sentences, however they are mainly used to generate individual clip descriptions.

We select videos posted on a social media platform to create our dataset because of the variability in topics, length, viewpoints, and quality. They also tend to represent a good distribution of stories communicated by humans. We select videos from social media that are public and popular with a large number of comments and shares that triggered interactions between people. In total, our dataset consists of 20k videos with duration ranging from 20s-180s and spanning across diverse topics that are observed on social media platforms. We follow Krishna et al. (2017) to create temporally annotated sentences where each task is divided into two steps: (i) describing the video in multiple sentences, covering objects, situations and important details of the video; (ii) aligning each sentence in the paragraph with the corresponding timestamps in the video. We refer to these as video segments. In Figure 1, we present two example annotated videos describing (i) a scene where two girls are playing with horses; (ii) a wedding with a bride walking down the aisle.

We summarize the statistics of our dataset in Table 2 and compare it to prior work in Table 1. Each of the 20k videos in our VideoStory dataset is annotated with a paragraph which has on average 4.67 temporally localized sentences. As we have three paragraphs per video for validation and test set, we have a total of 26,245 paragraphs with a total of 123k sentences. Each sentence in the dataset

Split	#Videos	#Clips	#Para	#W/P	Original	shuffled
train	17,098	80,598	17,098	61.76	-	-
val	999	13,796	2,997	59.88	20.95	24.82
test	1,011	14,093	3,033	59.77	21.12	24.95
test_blind	1,039	14,139	3,117	69.45	23.81	27.99
total	20,147	122,626	26,245	62.23	-	-

Table 2: VideoStory dataset: Dataset statistics (#V: No.of unique videos. #Para: No.of unique paragraphs. #W/P: Average number of words per paragraph.) and perplexity scores for original and shuffled sentences.

has an average length of 13.32 words, and each video has the average paragraph length of 62.23 words. Each sentence is aligned to a clip of on average 18.33 seconds which covers on average 26.04% of the full video. However, the entire paragraph for each video on average describes 96.7% of the whole video, demonstrating that each paragraph annotation covers the majority of the video. Furthermore, we found that 22% of the temporal descriptions overlap, showing that our annotation allows co-occurring or simultaneous events. We divide our dataset in training (17098 videos), validation (999), test (1011) and blind test splits (1039). Each video in the training set has a single annotation, but videos in validation, test, and blind test splits have three temporally localized paragraph annotations, for evaluation. While the test set can be used to compare model variants in a paper, only the best model per paper should be evaluated on the blind test set annotations, which will only be possible on an evaluation server. Annotations for the blind test set will not be released.

To explore the different domains in our dataset vs. ActivityNet captions we use the normalized pointwise mutual information to identify the words most closely associated with each dataset. Highest ranked words for ActivityNet are almost exclusively sports related, whereas in our dataset they include animals, baby, and words related to social events such as weddings. Most dominant actions in ActivityNet are either sports or household activity related whereas actions in our dataset are related to social activities such as laughing, waving, cheering etc. Our analysis of the distribution of POS categories show that nouns are the most dominant category observed in the VideoStory captions dataset with 24% of the total tokens followed by verbs (18.5%), determiners (15.9%), adjectives (4.36%), adverbs (5.16%) and propositions (5.04%). We also observe the similar distribution of POS categories in ActivityNet captions.

Figure 2: Distribution of annotations in time in VideoStory dataset. Most of the videos have majority of it annotated except the first few and last few seconds—which, in our analysis, correlated with the page/logo information.

We also find that ActivityNet has 50% of the videos where at least one segment in the video describes more than half of the video duration whereas in our dataset only 30% of videos have that phenomenon. In Figure 2, we show the distribution of sentence/segment annotations in time. The average number of (temporally localized) sentences is 4.67 compared to 3.65 in ActivityNet, despite having shorter videos, indicating the high information content of our videos.

In Table 3 we present all three paragraph annotations for a video showing a wedding ceremony. Out of 3 annotations, Annotation 2 is more descriptive compared to 1 and 3. However, it misses details about the presence of the photographer and taking the pictures.

Temporal Analysis. High quality video descriptions are more than bags of single-sentence captions; they should tell a coherent story. To identify the importance of sentence ordering or temporal coherence in our video paragraphs, we train a neural language model (Merity et al., 2017) on the training paragraphs of the VideoStory dataset and report perplexity on the correct order of sentences vs. randomly shuffled order of sentences in the descriptions created to understand the importance of temporal coherence in the video descriptions of our dataset. Results in Table 2 show that shuffled sentences have higher perplexity scores, demonstrating that order of sentences in the paragraphs are important for the coherence in the story.

3 Baseline Captioning Models

We explore learning to caption the videos using ground truth video segments.

Image Captioning Models. To understand if the temporal component of the video is contributing

Annotation 1: A bride walks down the aisle to her waiting bridegroom. As the bride walks, a photographer captures photos. At the end of the aisle the man giving the bride away shakes hands and hugs the bridegroom. The bride and bridegroom then interlock arms and face forward together.

Annotation 2: A large group of people have gathered inside of a room for a wedding. A woman walks down the aisle with a man slowly as people watch. The two of them get to the end of the aisle where a groom stands waiting.The man shakes hands with the groom and gives the woman a kiss on her forehead. The man who walked her down the aisle steps away towards the side of the room as the couple take each others arms.

Annotation 3: A groom is standing at the end of an aisle as a photographer takes a photo. The bride and father then come into view and walk down the aisle to the waiting groom. They stop at the grooms spot and the bride's father then shakes the grooms hand and gives a hug and walks to his spot. The groom then holds arms with the bride to begin the wedding ceremony.

Table 3: Example video description annotations in our VideoStory set. Each video has multiple paragraphs and localized time-interval annotations for every sentence in the paragraph.

to the description, we trained image captioning models on a frame sampled from the middle of the each segment of a video. We use the Show and Tell (Vinyals et al., 2015) image captioning architecture to generate captions.

Video Captioning Models. We study various video captioning models. First, we use sequence to sequence (seq-seq) recurrent neural network (RNN) model which has a two-layer encoder RNN to encode video features and a decoder RNN to generate descriptions. In the seq-seq approach we treat each description/segment individually and use an RNN decoder to describe each segment of the video, similar to Venugopalan et al. (2015), but using Gated Recurrent Units, GRUs, (Cho et al., 2014) for both the encoder and decoder.

In most videos, events are correlated with previous and future events. For example, for the first video description shown in Figure 1 once the girl is thrown into the water, she gets hold of herself, and the horse shakes off water on her. To capture such contextual correlations, we incorporate context from previous segment description into the captioning module. We build a model (seq-seq + context) which takes current segment video features and hidden representation of previous segment's sentence generation RNN at every timestamp in the decoder. For a given video segment, with hidden encoded video representation h_i^v and hidden representation of previous segment h_{i-1}^s, the concatenation of (h_i^v, h_{i-1}^s) is fed as input to the decoder that describes the segment (shown in Figure 3). Prior work has shown using previous video context has improved generated captions (Krishna et al., 2017).

Visual representation. For the image captioning models, we used features extracted from pre-trained ResNet-152 on ImageNet (He et al., 2016). For video captioning models we extract features from pre-trained 3D convolution ResNext-101 architecture trained on Kinetics (Kay et al., 2017), denoted as R3D, which achieved state-of-the-art results on various activity recognition tasks (Hara et al., 2018). Since a significant percentage of our videos has objects other than humans (e.g., animals) we also experiment with image-video fusion features(denoted by RNEXT, R3D) i.e., concatenation of ResNext-101 features extracted from pre-trained ImageNet with R3D features described above. We extract image features from the same frames which were used to extract R3D features.

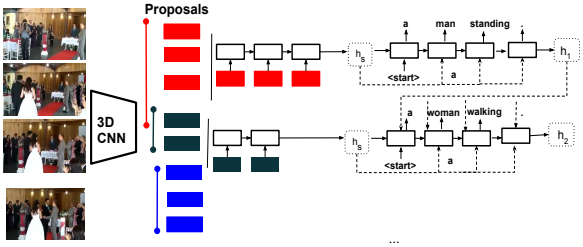

Figure 3: Our seq-seq+context model

4 Experiments and Results

For every segment, we set the maximum number of the sequence of features to 120 (i.e., 16X120 frames from the video) and maximum sentence length to 30. We trained using Adam optimizer with learning rate 0.0001. We use GRU as recurrent architecture to encode frames and decode captions with 512 dimensional hidden representation. We measure the captioning performance with most

GT (Ground Truth):
A baby is playing outside with two dogs.
The baby rolls the ball and the dog brings the ball back to the baby.
The baby tosses the ball again and again for the dogs.
One of the dogs walk away but the other stays and plays with the baby.

I (Image):
The dog is standing on the bed .
The dog is looking at the dog .
The dog is walking on the ground .
The dog is walking around the room .

seq-seq (RNEXT,R3D):
A dog is walking in the water with a baby .
The dog runs up and down the water .
The dog runs up and down the slide .
A baby is walking around the house with a baby .

seq-seq+context (RNEXT,R3D):
A dog is standing in the middle of a house .
The dog runs around the room and the dog jumps up and down .
The dog is walking on the floor and the dog walks away .
the girl runs around the house and the other dog runs away .

Table 4: Qualitative results: Descriptions generated by all variations of our baseline models.

Model	visual feat frame,video	B-3	B-4	M	R	C
I (single-frame)	RN-152, –	1.99	0.52	7.87	18.99	23.00
seq-seq	–,R3D	2.33	0.60	8.33	19.59	26.48
seq-seq + context	– ,R3D	2.78	0.78	9.20	21.24	30.80
seq-seq	RNEXT,R3D	2.63	0.79	8.44	19.89	27.64
seq-seq + context	RNEXT,R3D	**3.37**	**1.20**	**9.37**	**21.52**	**33.88**
trained on ActivityNet Captions						
seq-seq + context	RNEXT,R3D	1.68	0.49	8.48	19.40	22.12

Table 5: Captioning results from VideoStory Dataset using ground-truth test video segments. We report BLEU (B) and METEOR (M), ROUGE-L(R) and CIDEr (C). Best scores are in bold.

commonly-used evaluation metrics: BLEU$\{3,4\}$, METEOR, ROUGE-L, and CIDEr following previous works of image and video captioning (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005; Vedantam et al., 2015).

In Table 5, we present the performance of our baseline models on VideoStory test dataset. We observe that models that consider context (seq-seq+context) from the previously generated sentence have better performance than the corresponding models without context (seq-seq), with both 3D convolution based features (R3D) as well as image-video fusion features (RNEXT,R3D). This indicates that our model benefited from contextual information, and that sentences in our stories are contextual, rather than independent.

To validate the strength of our baseline model, we train our best performing model on ActivityNet Captions. It achieves 10.92 (METEOR) and 43.42 (CIDEr) on the val set, close to state-of-the-art results of 11.06 and 44.71 by Zhou et al. (2018b), indicating that it is a strong baseline. However, when evaluating our ActivityNet model on our VideoStory dataset (Table 5, last row), we see significantly lower performance compared to a model trained on our dataset, highlighting the complementary nature of our dataset.

Our image only (single frame) model has the lowest scores across all metrics suggesting that a single image is not enough to generate contextual descriptions. We observed that our fusion models consistently outperform models with video-only R3D features, indicating features extracted using pre-trained ImageNet complement activity based R3D features. We show qualitative results from the variants of our models in Table 4. We observe that single frame models tend to repeat same captions and seq-seq model without context repeats phrases in the descriptions.

5 Conclusions

This paper introduces a dataset which we sourced from videos on social media and annotated with multi-sentence descriptions. We benchmark strong baseline approaches on the dataset, and our evaluations show that our dataset is complementary from prior work due to more diverse topics and the selection of engaging videos which tell a story. Our VideoStory dataset can serve as a good benchmark to build models for story understanding and multi-sentence video description.

Acknowledgements

We would like to thank Ranjay Krishna for providing the annotation interface used in Krishna et al. (2017) which we adapted to collect our dataset. We would also like to thank Haoqi Fan, Boris Vassilev, Jamie Ray, Sasha Sheng, Nikhila Ravi, and Evan Numbers for their help collecting the dataset, Devi Parikh for feedback on the annotation interface and Anna Rohrbach for useful feedback on drafts of this paper.

References

Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72.

David Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 190–200.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111.

Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 2634–2641.

Sandra Eliza Fontes De Avila, Ana Paula Brandão Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. 2011. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pages 18–22.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

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

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Patterm Recognition (CVPR)*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*.

Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. TGIF: A new dataset and benchmark on animated GIF description. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4641–4650.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Michael Mateas and Phoebe Sengers. 2003. *Narrative intelligence*. J. Benjamins Pub.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and Optimizing LSTM Language Models. *arXiv preprint arXiv:1708.02182*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *TACL*, 1:25–36.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition - 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings*, pages 184–195.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3202–3212.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*, 123(1):94–120.

Gunnar A Sigurdsson, Xinlei Chen, and Abhinav Gupta. 2016a. Learning visual storylines with skipping recurrent neural networks. In *European Conference on Computer Vision*, pages 71–88. Springer.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016b. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 510–526.

Atousa Torabi, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. 2015. Using descriptive video services to create a large data source for video annotation research. *CoRR*, abs/1503.01070.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296.

Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. 2016. Generation for user generated videos. In *European conference on computer vision*, pages 609–625. Springer.

Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018*.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748.