iPerceive: Multi-Modal Dense Video Captioning using Common-Sense Reasoning

Aman Chadha
Stanford University / Apple Inc.
amanc@stanford.edu

Abstract

Dense video captioning (DVC) is the task of localizing events from an untrimmed video and producing textual descriptions for each event. Most of the previous works in DVC, and more generally in visual understanding, rely solely on understanding the "what" (e.g., object recognition) and "where" (e.g., event localization), which in some cases, fails to describe correct contextual relationships between events or leads to incorrect underlying visual attention. Part of what defines us as human and fundamentally different from machines is our instinct to seek causality behind any association, say an event Y that happened as a direct result of event X. To this end, we propose iPerceive, a system capable of understanding the "why" between events in a video by building a common-sense knowledge base using contextual cues. Furthermore, while most prior art in DVC relies solely on visual information, other modalities such as audio and speech are vital for a human observer's perception of an environment. We formulate the captioning task as a machine translation problem that utilizes multiple modalities. Another common drawback of current methods is that they train the event proposal and captioning model either separately or in alternation, which prevents direct influence of the proposal based on the caption. To address this, we adopt an end-to-end Transformer architecture. By evaluating the performance of iPerceive on ActivityNet Captions and YouCookII datasets, we aim for our results to show that our approach furthers the state-of-the-art.

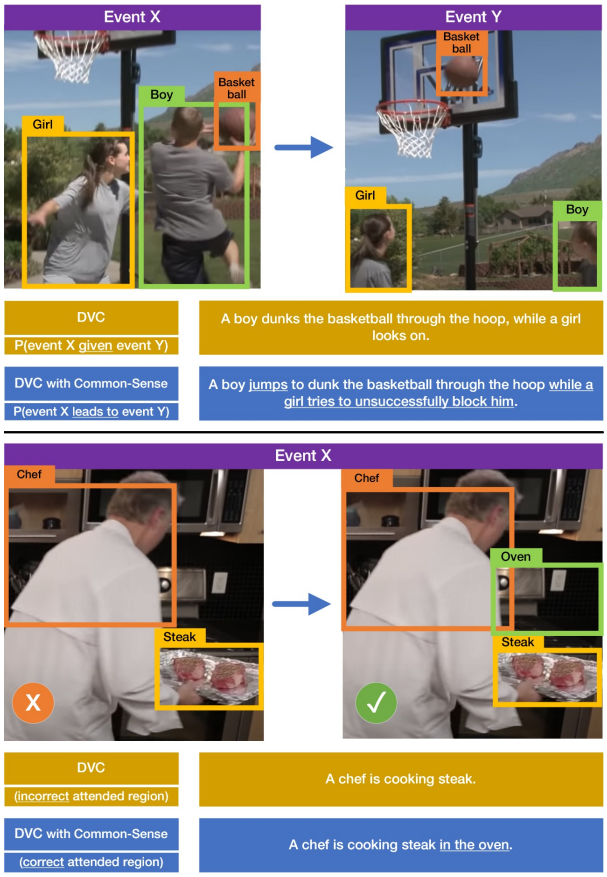

Figure 1. Top: An example of a "cognition error" in DVC. While the girl tries to block the boy's dunking attempt, him *jumping* (event X) eventually *leads* to him dunking the basketball through the hoop (event Y). Bottom: An example of an incorrect attended region where conventional DVC approaches correlate a chef and steak to the activity of cooking without even attending to the nearby oven. We used [24] as our DVC baseline as it is the current state-of-the-art.

1. Introduction

Video content has become an important source for humans to acquire information and knowledge. The task of describing a video using natural language and synthesizing a very compact and intuitive representation is typically referred to as video captioning.

A typical video contains numerous events, some separable in time, while others that occur in parallel. Producing a single description for an entire video might be impractical for long untrimmed footage. Instead, dense video captioning (DVC) [12] aims to temporally localize events and describe them using natural language. DVC is analogous to dense image captioning [10]; it describes videos and localize events in time whereas dense image captioning localizes and describes regions in space.

2. Problem Statement

Today's computer vision systems are good at telling us the "what" (e.g., classification [11], segmentation [5]) and "where" (e.g., detection [13], localization [20], tracking [29]). Common-sense reasoning [18], which leads to the interesting question of "why", is a thinking gap in today's pattern-learning-based systems which are based on the likelihood of observing object Y given object X, $P(Y|X)$.

Failing to factor in causality leads to the unfortunate conclusion that the co-existence of objects X and Y might be attributed to spurious observational bias [6, 15], e.g., if a keyboard and mouse are often observed on a table, the model learns to develop an "association" between the two. The underlying common-sense that the keyboard and mouse are parts of a computer would not be inferred, and infact the duo would be wrongly associated as being part of a table. In the event that a keyboard and mouse are observed outside of a tabular setting, the model can commit a "cognition error" due to the lack of common-sense.

Next, we take upon the task of enhancing the gamut of telemetry that can be ingested by our model to fare better at the task at hand. The vast majority of research in the field of dense captioning generates captions purely based on visual information [12, 31, 23]. However, given the fact that auditory feedback is an essential aspect of human communication, unsurprisingly, almost all videos include an audio track and some also include a speech track, both of which could provide vital cues for understanding the context of the event. Inspired by the implementation in [9], our model consumes video frames, the raw audio signal and the speech content for the caption generation process.

DVC can be decomposed into two parts: event detection and event description. Existing methods tackle this using a module for each of these sub-tasks, and exploit two ways to combine them for DVC. One way is to train the two modules independently and generate descriptions for the best event proposals with the best captioning model [3]. The other way is to alternate between training [12] the two modules, i.e., alternate between (i) training the proposal module only and (ii) training the captioning module on the positive event proposals while fine-tuning the proposal module. However, either case inhibits directly influencing the proposal module based on the quality of the generated caption.

Another challenge for DVC, and more broadly for sequence modeling tasks, is the need to learn a representation that is capable of capturing long term dependencies. The Transformer architecture [22] implements a fast self-attention mechanism that has demonstrated its effectiveness in machine translation. Since the Transformer does not require unrolling across time, and therefore trains and tests much faster as compared to RNN based models. We thus deploy a Transformer in both the encoder and decoder of our model.

Our key contributions are three-fold.

(a) Common-sense reasoning for videos: We reimagine the concept of common-sense knowledge base generation for videos, building upon [24] as baseline. [24] tackles the issue of observational bias in the context of images, but applying a similar set of ideas to videos comes with its own set of challenges distinct from the image case. One observation is that events in videos can range across multiple time scales and can even overlap. Also, events can have causal relationships between themselves $(X \Rightarrow Y)$ that humans subconsciously perceive without any visible acknowledgment/feedback. Humans naturally learn common sense in an unsupervised fashion by exploring the physical world, and until machines imitate this learning path, there will be a "gap" between man and machine. This requires us to build a knowledge base and acquire contextual information from all temporal events in a video sequence to determine inherent causal relationships. These "context-aware" features can improve both the accuracy of contextual relationships as well as steer attention to the right entities.

We modify the approach proposed in [24] to suit the aforementioned nuances specific to videos. Further, since our common-sense features can be generated in a self-supervised manner, they can easily be adapted for other high-level vision tasks such as video question answering (VideoQA) [32, 27, 28].

(b) Blending multi-modal DVC with end-to-end Transformer-based learning: While [9] presents a detailed study of the merits of using multiple modalities for DVC, they do not implement an end-to-end trainable system. This restricts model "wiggle", i.e., since events in a video sequence and the generated language are closely related, the language information should ideally be able to help localize events in the video. To this end, we utilize an end-to-end trainable model similar to [31] – this enforces consistency between the content in the proposed video segment and the semantic information in the language description.

(c) Using only past context or past+future context depending on whats available: While most DVC studies work on stored videos where both backward and forward context is available [31, 23, 9, 16, 12, 25], some look into dense captioning steaming videos [26] and thus consider only the "past context. iPerceive can work through both scenarios – with streaming video streams where only past context can be looked up, as well as stored videos where both past and future context is available at each time step. To the best of our knowledge, our model is the first one that uses common-sense reasoning as its underlying backbone to analyze and build contextual relationships in videos and applies it to multi-modal DVC with end-to-end Transformer-based learning.

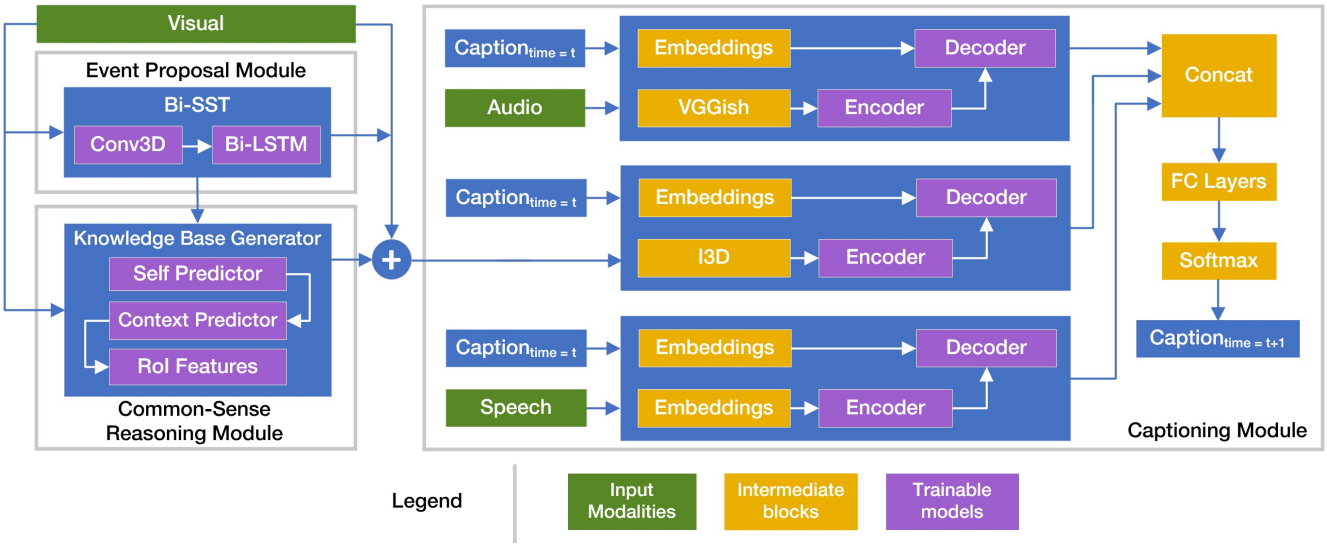

Figure 2. Architectural overview of iPerceive. iPerceive generates "common-sense" vectors from the temporal events that the proposal module localizes (left). Features from all modalities are sent to the corresponding encoder-decoder Transformers (middle). Upon fusing the processed features we finally output the next word in the caption using the distribution over the vocabulary (right).

3. Proposed Framework

3.1. Top Level View

Fig. 2 outlines the goals of iPerceive: (i) temporally localize a set of events in a video, (ii) build a knowledge base for common-sense reasoning and, (iii) produce a textual description using audio, visual and speech cues for each event. To this end, we apply a three-stage approach.

3.2. Event Proposal Module

First, we localize the temporal events in the video using the bidirectional single-stream (Bi-SST) network proposed in [23]. Bi-SST applies 3D convolutions (C3D) [21] to video frames and passes on the extracted features to a Bi-directional LSTM [8] network. The LSTM accumulates visual cues from past and future context over time and predicts the endpoints of each event in the video along with its confidence scores. The output of the LSTM feeds the common-sense reasoning module to convey the set of events in the input video to extract "common-sense" features for.

3.3. Common-Sense Reasoning Module

Common-sense reasoning enables self-supervised feature representation learning to serve as an improved visual region encoder for subsequent processing. In the context of developing a common-sense knowledge base, one of the biggest challenges involved is determining causal context: how do you figure out the cause and effect relationship between objects and thus re-align context for the model? The proxy training objective of the common-sense reasoning module, in essence, is to predict the contextual objects of an event. Common-sense reasoning requires $P(Y|do(X))$, which is fundamentally different from what prior work in the domain of DVC, and video understanding in general, uses, the conventional likelihood $P(Y|X)$.

Following the approach in [24], we compare the difference between the conventional likelihood $P(Y|X)$ and causal intervention $P(Y|do(X))$ [18]. Intuitively, we carry out the following deliberate "borrow-put" experiment for a given event extracted from the proposal module: (1) "borrow" non-local context, say an object Z from another event, (2) "put" Z in the context of object X and object Y, and then (3) test if object X still causes the existence of object Y given Z. This experiment helps determine if the chance of Z is independent on X or Y. Therefore, by using $P(Y|do(X))$ as the learning objective instead of $P(Y|X)$, the observational bias from the "apparent" context could be alleviated.

Our visual world contains several confounders $z \in Z$ that add spurious observational bias around the existence of objects X and Y and hinder common-sense development. This is due to us limiting the model's learning using the traditional likelihood $P(Y|X)$, which can be formally defined [24] using Bayes' rule as:

$$P(Y|X) = \sum_{z} P(Y|X,z)P(z|X) \tag{1}$$

where the confounder z introduces the observational bias via $P(z|X)$.

Similarly, we define the "do" operation by disrupting the causal link between z and X (and thus de-biasing X) as,

$$P(Y|do\,(X)) = \sum_{z} P(Y|X,z)P(z) \tag{2}$$

These "common-sense" features are stacked with the corresponding video input and sent down-stream for further processing.

3.4. Captioning Module

Given an event proposal and its "common-sense" vectors, the captioning module generates a caption using audio, visual and speech modalities. We use inflated 3D convolutions (I3D) [1] to process visual modalities and the VGGish network [7] for audio modalities. We deploy an automatic speech recognition (ASR) system [4] to extract temporally-aligned speech transcriptions. These are juxtaposed alongside the video frames and the corresponding audio track, and fed in as input to our model. We formulate the captioning task as a machine translation problem and utilize the recently proposed Transformer architecture [22] to convert multi-modal input data into textual descriptions [31]. Features from each modality are fed to individual Transformer models along with the words of a caption from the previous time steps. The output of each Transformer is fused and a probability distribution is estimated over the vocabulary.

3.5. Loss functions

iPerceive uses a four-fold training loss: (i) cross entropy MCE as proposal loss L_p to balance positive and negative proposals, (ii) negative log likelihood as common-sense reasoning loss L_{cs}, (iii) binary cross entropy BCE as mask prediction loss L_m, and (iv) cross entropy MCE across all words in every sentence as captioning loss L_c.

$$L_p = MCE(c, t, X, y) \tag{3}$$

where, c is the prediction score at time t; X is the input video; y is the ground truth label with an acceptable intersection over union (IoU).

$$L_{cs} = L_{self}(e, p, x^c, t) + \frac{1}{K}\sum_i L_{context}(e, p_i, y_i^c, t) \tag{4}$$

where, e is an event localized by the proposal module; p is the probability distribution output of the self-predictor [24] at time t; x^c is the ground-truth class of the object within the region of interest (RoI) X at time t; p_i is $P(Y_i|do(X))$ where X is the "center" object in event e while Y_i is one of the K "context" objects with ground-truth label y_i^c.

$$L_m = BCE(Bin(S_a, E_a, t), f_M(S_a, E_p, S_a, E_a, t)) \tag{5}$$

where, $Bin(.)$ is 1 if $t \in [S_a, E_a]$, 0 otherwise; f_M is a differentiable mask for time t; S_p and E_p are the start and end times of the proposal; S_a and E_a are the start and end positions of the anchor.

$$L_c = MCE(w'_t, w_t) \tag{6}$$

where w'_t is the ground truth word at time t. The final loss L is a combination of the individual losses,

$$L = \lambda_1 L_p + \lambda_2 L_{cs} + \lambda_3 L_m + \lambda_4 L_c \tag{7}$$

where λ_{1-4} weigh the individual loss components.

4. Experiments

4.1. Dataset

We train and assess iPerceive using the ActivityNet Captions [12] dataset, using a train/val/test split of 0.5:0.25:0.25. We report all results using the validation set (since no ground truth is available for the test set).

4.2. Metrics

We evaluate the performance of iPerceive using BLEU@N [17] and METEOR [2]. We regard the METEOR as our primary metric as it has been shown to be highly correlated with human judgement especially in a situation with a limited number of references. It is worthwhile to note here that the BLEU n-gram order that has the highest correlation with monolingual human judgements is four [17]. We use the official evaluation script provided in [12].

4.3. Comparison with Baseline Methods

Tab. 1 shows a comparison of the baseline implementation of iPerceive with the state-of-the-art. Algorithms were split into the ones which "saw" all training videos and others which trained on partially available data (since some YouTube videos which were part of the ActivityNet Captions dataset are no longer available).

Note that iPerceive's current implementation uses a multi-modal dense captioning model with end-to-end training. The implementation of our common-sense reasoning module is work-in-progress. Apart from the ActivityNet Captions dataset, we plan to train and assess iPerceive using the YouCookII dataset [30]. Also planned are ablation studies for (i) common-sense reasoning and, (ii) end-to-end training to determine their performance impact.

Table 1. Evaluation of iPerceive on the ActivityNet Captions validation set using BLEU @ 3, 4 (B@3, B@4) and METEOR (M).

Method	GT Proposals			Learned Proposals		
	B@3	B@4	M	B@3	B@4	M
Seen full dataset						
Krishna et al. [12]	4.09	1.60	8.88	1.90	0.71	5.69
Wang et al. [23]	-	-	10.89	2.55	**1.31**	5.86
Zhou et al. [31]	5.76	2.71	11.16	2.42	1.15	4.98
Li et al. [14]	4.55	1.62	10.33	2.27	0.73	6.93
Seen part of the dataset						
Rahman et al. [19]	3.04	1.46	7.23	1.85	0.90	4.93
Iashin et al. [9]	4.12	1.81	10.09	2.31	0.92	6.80
iPerceive	4.23	1.94	10.49	2.58	1.07	7.03
Iashin et al. (all modalities)	5.83	2.86	11.72	2.60	1.07	7.31
iPerceive (all modalities)	**5.91**	**2.95**	**12.01**	**2.73**	1.19	**7.52**

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2017. 4

[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*, pages 376–380, 2014. 4

[3] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, Ranjay Khrisna, Victor Escorcia, Kenji Hata, and Shyamal Buch. Activitynet challenge 2017 summary. *arXiv preprint arXiv:1710.08011*, 2017. 2

[4] Google. YouTube Data API Video Captions. https://developers.google.com/youtube/v3/docs/captions, 2008. [Accessed May 21 2020]. 4

[5] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa. Efficient hierarchical graph-based video segmentation. In *2010 ieee computer society conference on computer vision and pattern recognition*, pages 2141–2148. IEEE, 2010. 2

[6] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018. 2

[7] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 4

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[9] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. *arXiv preprint arXiv:2003.07758*, 2020. 2, 4

[10] Justin Johnson, Andrej Karpathy, and Fei-Fei Li. Densecap: fully convolutional localization networks for dense captioning. corr abs/1511.07571 (2015). *arXiv preprint arXiv:1511.07571*, 2015. 1

[11] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[12] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1, 2, 4

[13] Liyuan Li, Weimin Huang, Irene YH Gu, and Qi Tian. Foreground object detection from videos containing complex background. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 2–10, 2003. 2

[14] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7492–7500, 2018. 4

[15] Varun Manjunatha, Nirat Saini, and Larry S Davis. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9562–9571, 2019. 2

[16] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6588–6597, 2019. 2

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. 4

[18] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2, 3

[19] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8908–8917, 2019. 4

[20] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 3

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 4

[23] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7190–7198, 2018. 2, 3, 4

[24] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. *arXiv preprint arXiv:2002.12204*, 2020. 1, 2, 3, 4

[25] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 468–483, 2018. 2

[26] Huijuan Xu, Boyang Li, Vasili Ramanishka, Leonid Sigal, and Kate Saenko. Joint event detection and description in continuous video streams. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 396–405. IEEE, 2019. 2

[27] Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. Videoqa: question answering on news video. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 632–641, 2003. 2

[28] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering.

In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[29] Da Zhang, Hamid Maei, Xin Wang, and Yuan-Fang Wang. Deep reinforcement learning for visual object tracking in videos. *arXiv preprint arXiv:1701.08936*, 2017. 2

[30] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 4

[31] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 2, 4

[32] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421, 2017. 2