Dense-Captioning Events in Videos

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, Juan Carlos Niebles
Stanford University
{ranjaykrishna, kenjihata, fren, feifeili, jniebles}@cs.stanford.edu

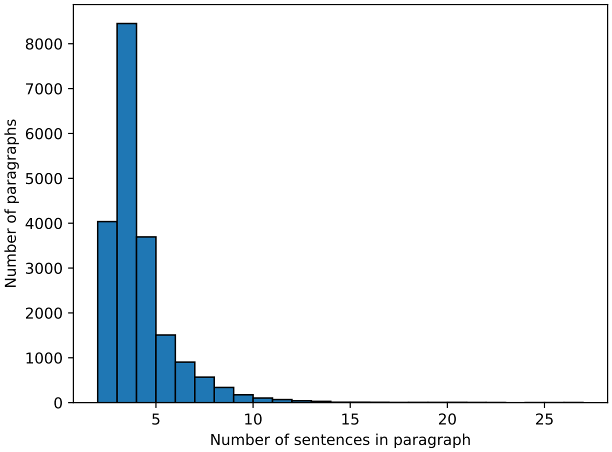


Figure 1: The number of sentences within paragraphs is normally distributed, with on average 3.65 sentences per paragraph.

1. Supplementary material

In the supplementary material, we compare and contrast our dataset with other datasets and provide additional details about our dataset. We include screenshots of our collection interface with detailed instructions. We also provide additional details about the workers who completed our tasks.

1.1. Comparison to other datasets.

Curation and open distribution is closely correlated with progress in the field of video understanding (Table 1). The KTH dataset [16] pioneered the field by studying human actions with a black background. Since then, datasets like UCF101 [18], Sports 1M [6], Thumos 15 [5] have focused on studying actions in sports related internet videos while HMDB 51 [9] and Hollywood 2 [10] introduced a dataset of movie clips. Recently, ActivityNet [1] and Charades [17] broadened the domain of activities captured by these datasets by including a large set of human activities. In an effort to map video semantics with language, MPII MD [13] and M-VAD [19] released short movie clips with descriptions. In an effort to capture longer events, MSR-VTT [20], MSVD [2] and YouCook [3] collected a dataset with slightly longer length, at the cost of a few descriptions than previous datasets. To further improve video annotations, KITTI [4] and TACoS [11] also temporally localized their video descriptions. Orthogonally, in an effort to increase the complexity of descriptions, TACos multi-level [12] expanded the TACoS [11] dataset to include paragraph descriptions to instructional cooking videos. However, their dataset is constrained in the "cooking" domain and contains in the order of a 100 videos, making it unsuitable for dense-captioning of events as the models easily overfit to the training data.

Our dataset, ActivityNet Captions, aims to bridge these three orthogonal approaches by temporally annotating long videos while also building upon the complexity of descriptions. ActivityNet Captions contains videos that an average of 180s long with the longest video running to over 10 minutes. It contains a total of 100k sentences, where each sentence is temporally localized. Unlike TACoS multi-level, we have two orders of magnitude more videos and provide annotations for an open domain. Finally, we are also the first dataset to enable the study of concurrent events, by allowing our events to overlap.

1.2. Detailed dataset statistics

As noted in the main paper, the number of sentences accompanying each video is normally distributed, as seen in Figure 1. On average, each video contains 3.65 ± 1.79 sentences. Similarly, the number of words in each sentence is normally distributed, as seen in Figure 2. On average, each sentence contains 13.48 ± 6.33 words, and each video contains 40 ± 26 words.

There exists interaction between the video content and the corresponding temporal annotations. In Figure 3, the number of sentences accompanying a video is shown to be positively correlated with the video's length: each additional minute adds approximately 1 additional sentence description. Furthermore, as seen in Figure 4, the sentence descriptions focus on the middle parts of the video more

Dataset	Domain	# videos	Avg. length	# sentences	Des.	Loc. Des.	paragraphs	overlapping
UCF101 [18]	sports	13k	7s	-	-	-	-	-
Sports 1M [6]	sports	1.1M	300s	-	-	-	-	-
Thumos 15 [5]	sports	21k	4s	-	-	-	-	-
HMDB 51 [9]	movie	7k	3s	-	-	-	-	-
Hollywood 2 [10]	movie	4k	20s	-	-	-	-	-
MPII cooking [14]	cooking	44	**600s**	-	-	-	-	-
ActivityNet [1]	human	20k	**180s**	-	-	-	-	-
MPII MD [13]	movie	68k	4s	68,375	✓	-	-	-
M-VAD [19]	movie	49k	6s	55,904	✓	-	-	-
MSR-VTT [20]	**open**	**10k**	20s	**200,000**	✓	-	-	-
MSVD [2]	**human**	**2k**	10s	**70,028**	✓	-	-	-
YouCook [3]	cooking	88	-	2,688	✓	-	-	-
Charades [17]	**human**	**10k**	30s	16,129	✓	-	-	-
KITTI [4]	driving	21	30s	520	✓	✓	-	-
TACoS [11]	cooking	127	**360s**	11,796	✓	✓	-	-
TACos multi-level [12]	cooking	127	**360s**	52,593	✓	✓	✓	-
ActivityNet Captions (ours)	**open**	**20k**	**180s**	**100k**	✓	✓	✓	✓

Table 1: Compared to other video datasets, ActivityNet Captions contains long videos with a large number of sentences that are all temporally localized and is the only dataset that contains overlapping events. (Loc. Des. shows which datasets contain temporally localized language descriptions. Bold fonts are used to highlight the nearest comparison of our model with existing models.)

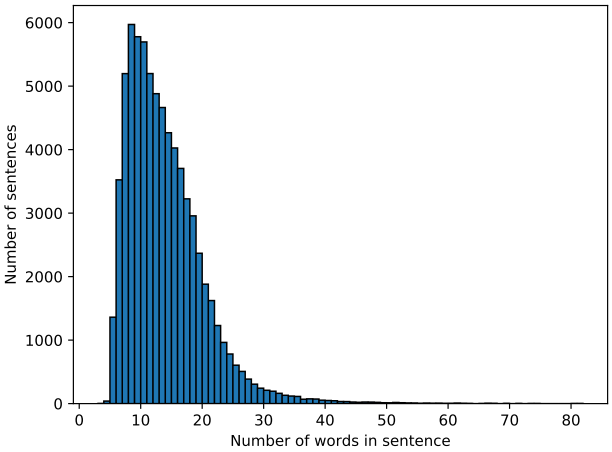


Figure 2: The number of words per sentence within paragraphs is normally distributed, with on average 13.48 words per sentence.

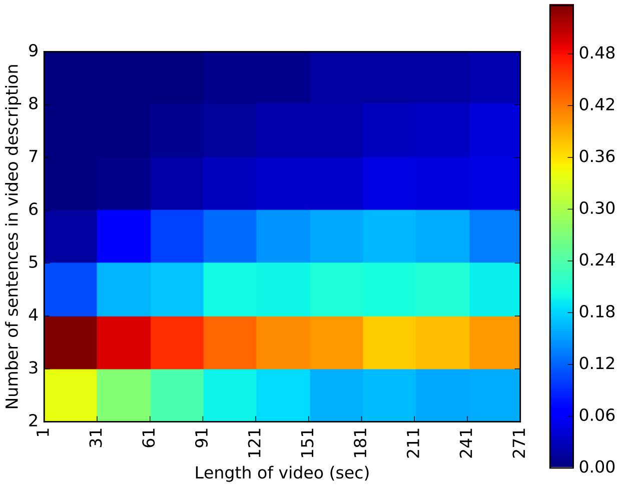


Figure 3: Distribution of number of sentences with respect to video length. In general the longer the video the more sentences there are, so far on average each additional minute adds one more sentence to the paragraph.

than the beginning or end.

When studying the distribution of words in Figures 5 and 6, we found that ActivityNet Captions generally focuses on people and the actions these people take. However, we wanted to know whether ActivityNet Captions captured the general semantics of the video. To do so, we compare our sentence descriptions against the shorter labels of ActivityNet, since ActivityNet Captions annotates ActivityNet videos. Figure 11 illustrates that the majority of videos in ActivityNet Captions often contain ActivityNet's labels in at least one of their sentence descriptions. We find that the many entry-level categories such as *brushing hair* or *playing violin* are extremely well represented by our captions. However, as the categories become more nuanced, such as

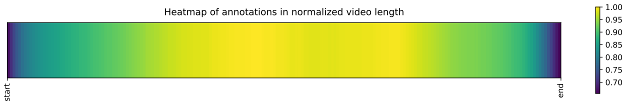


Figure 4: Distribution of annotations in time in ActivityNet Captions videos, most of the annotated time intervals are closer to the middle of the videos than to the start and end.

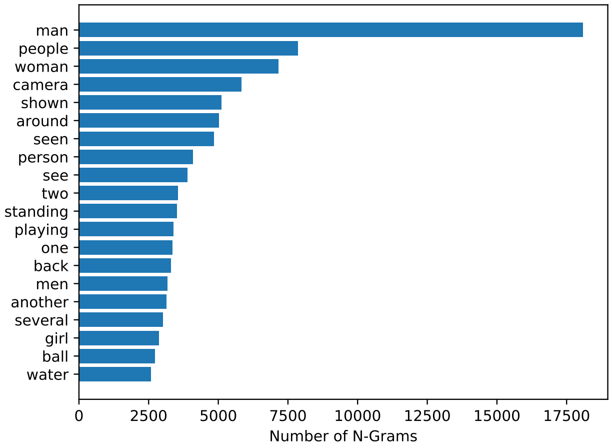


Figure 5: The most frequently used words in ActivityNet Captions with stop words removed.

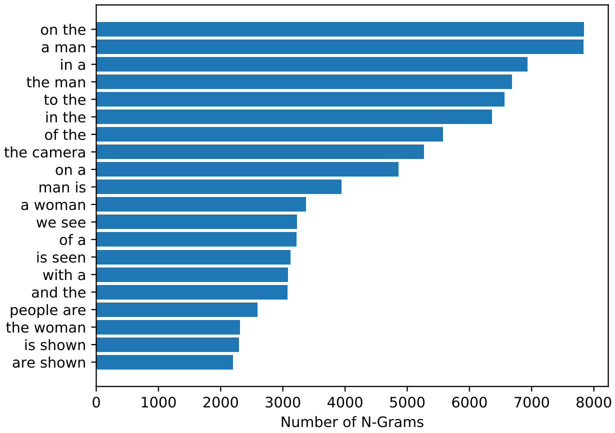
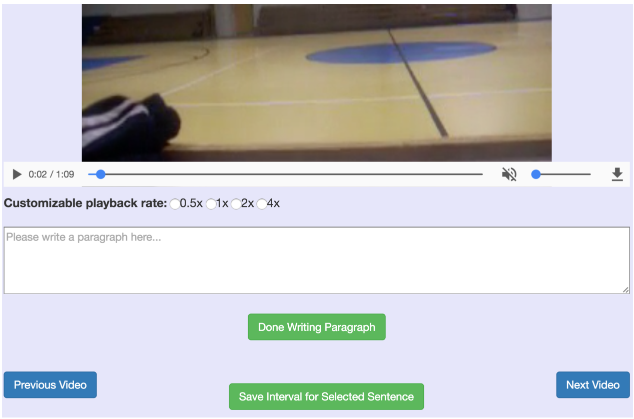


Figure 6: The most frequently used bigrams in ActivityNet Captions .

powerbocking or *cumbia*, they are not as commonly found in our descriptions.

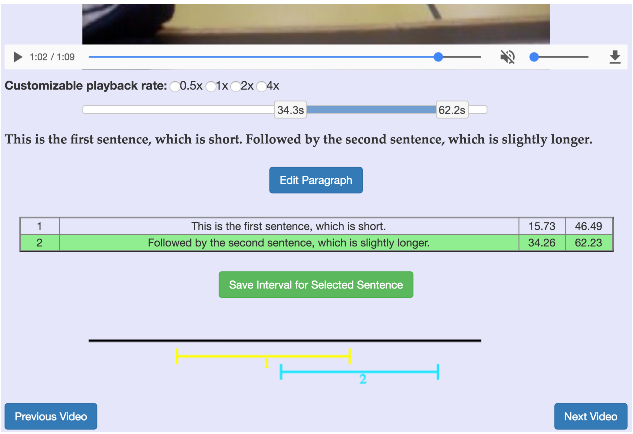
1.3. Dataset collection process

We used Amazon Mechanical Turk to annotate all our videos. Each annotation task was divided into two steps: (1) Writing a paragraph describing all major events happening in the videos in a paragraph, with each sentence of the paragraph describing one event (Figure 7; and (2) Labeling the start and end time in the video in which each sentence in the paragraph event occurred (Figure 8. We find complementary evidence that workers are more consistent with their video segments and paragraph descriptions if they are asked to annotate visual media (in this case, videos) using natural language first [7]. Therefore, instead of asking workers to segment the video first and then write individual sentences, we asked them to write paragraph descriptions first.


0:02 / 1:09
Please write a paragraph here...
Done Writing Paragraph
Previous Video
Save Interval for Selected Sentence
Next Video

Customizable playback rate: ○0.5x ○1x ○2x ○4x

Figure 7: Interface when a worker is writing a paragraph. Workers are asked to write a paragraph in the text box and press ”Done Writing Paragraph” before they can proceed with grounding each of the sentences.



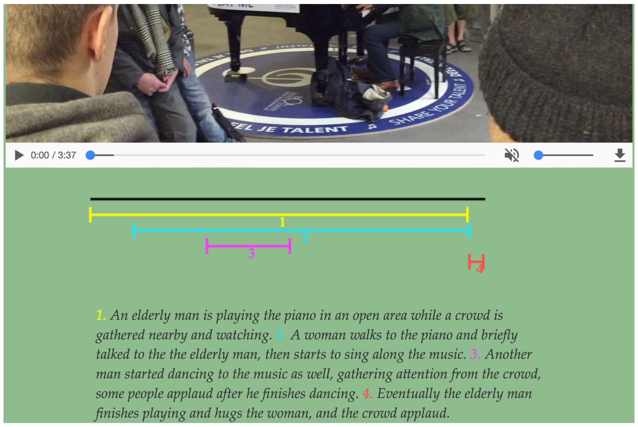
Customizable playback rate: ○0.5x ○1x ○2x ○4x

This is the first sentence, which is short. Followed by the second sentence, which is slightly longer.

1	This is the first sentence, which is short.	15.73	46.49
2	Followed by the second sentence, which is slightly longer.	34.26	62.23

Figure 8: Interface when labeling sentences with start and end timestamps. Workers select each sentence, adjust the range slider indicating which segment of the video that particular sentence is referring to. They then click save and proceed to the next sentence.

Workers are instructed to ensure that their paragraphs are at least 3 sentences long where each sentence describes events in the video but also makes a grammatically and semantically coherent paragraph. They were allowed to use



1. An elderly man is playing the piano in an open area while a crowd is gathered nearby and watching. 2. A woman walks to the piano and briefly talked to the the elderly man, then starts to sing along the music. 3. Another man started dancing to the music as well, gathering attention from the crowd, some people applaud after he finishes dancing. 4. Eventually the elderly man finishes playing and hugs the woman, and the crowd applaud.

Figure 9: We show examples of good and bad annotations to workers. Each task contains one good and one bad example video with annotations. We also explain why the examples are considered to be good or bad.

co-referencing words (ex, *he*, *she*, etc.) to refer to subjects introduced in previous sentences. We also asked workers to write sentences that were at least 5 words long. We found that our workers were diligent and wrote an average of 13.48 number of words per sentence. Each of the task and examples (Figure 9) of good and bad annotations.

Workers were presented with examples of good and bad annotations with explanations for what constituted a good paragraph, ensuring that workers saw concrete evidence of what kind of work was expected of them (Figure 9). We paid workers $3 for every 5 videos that were annotated. This amounted to an average pay rate of $8 per hour, which is in tune with fair crowd worker wage rate [15].

1.4. Annotation details

Following research from previous work that show that crowd workers are able to perform at the same quality of work when allowed to video media at a faster rate [8], we show all videos to workers at 2X the speed, i.e. the videos are shown at twice the frame rate. Workers do, however, have the option to watching the videos at the original video speed and even speed it up to 3X or 4X the speed. We found, however, that the average viewing rate chosen by workers was 1.91X while the median rate was 1X, indicating that a majority of workers preferred watching the video at its original speed. We also find that workers tend to take an average of 2.88 and a median of 1.46 times the length of the video in seconds to annotate.

At any given time, workers have the ability to edit their paragraph, go back to previous videos to make changes to their annotations. They are only allowed to proceed to the next video if this current video has been completely annotated with a paragraph with all its sentences timestamped. Changes made to the paragraphs and timestamps are saved when ”previous video or ”next video” are pressed, and reflected on the page. Only when all videos are annotated can the worker submit the task. In total, we had 112 workers who annotated all our videos.

References

[1] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[2] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Portland, OR, June 2011.

[3] P. Das, C. Xu, R. F. Doell, and **J. J.. Corso**. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[5] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://www.thumos.info/`, 2015.

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. 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.

[7] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International Journal on Computer Vision (IJCV)*, 2017.

[8] R. A. Krishna, K. Hata, S. Chen, J. Kravitz, D. A. Shamma, L. Fei-Fei, and M. S. Bernstein. Embracing error to enable rapid crowdsourcing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3167–3179. ACM, 2016.

[9] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011.

[10] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[11] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36, 2013.

[12] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description

with variable level of detail. In *German Conference on Pattern Recognition*, pages 184–195. Springer, 2014.

[13] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1194–1201. IEEE, 2012.

[15] N. Salehi, L. C. Irani, M. S. Bernstein, A. Alkhatib, E. Ogbe, K. Milland, et al. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1621–1630. ACM, 2015.

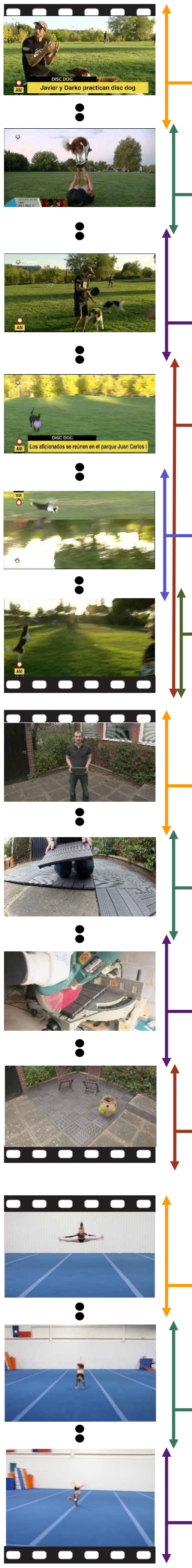
[16] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[17] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, 2016.

[18] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[19] A. Torabi, C. Pal, H. Larochelle, and A. Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015.

[20] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016.



Ground Truth	No Context	With Context
A man sits with his dog in the grass and holds out discs in his hands.	A man is seen speaking to the camera while holding a dog around him.	A man is seen speaking to the camera while standing in a field with a dog.
The man balances his dog on his feet then throws Frisbee discs for him.	The woman continues to swing around with the frisbee as well as performing tricks.	The dog is seen in several clips performing tricks with his dog and running all around the yard.
The man spins his dog and holds it in his arms.	The man then begins to do tricks with the dog while the camera follows him.	The man then begins walking around with a frisbee.
Different trainers throw Frisbee discs for the dogs to catch while performing tricks.	A woman is seen walking out onto a field with a dog.	The dog runs around in circles on the field with the dog.
A woman throws discs to her dog that jumps from her back.	The dog jumps off the girl and the dog jumps to the dog.	The dog runs around in circles on the field with the frisbee.
The woman throws multiple discs in a row for her dog to catch.	The dog jumps off the girl and the dog jumps to the dog.	The dog runs around in circles on the grass as he chases the frisbee.

Ground Truth	No Context	With Context
A man is standing outside holding a black tile.	a man is seen speaking to the camera while holding up a tool and begins to cut.	a man is seen speaking to the camera while holding up a bucket and begins painting the wall.
He starts putting the tile down on the ground.	the man then puts a on the floor and begins putting into the tire and.	a man is seen kneeling down on a roof and begins using a tool on the carpet.
He cuts the tile with a red saw.	the man then puts a on the floor and begins putting tiles on the sides and.	a man is seen speaking to the camera and leads into him holding knives and sharpening a board .
He sets chairs and flowers on the tile.	a person is seen pushing a puck down a floor with a rag and showing the camera.	the person then walks around the table and begins painting the fence.

Ground Truth	No Context	Full Context
A little girl performs gymnastics jumping and flipping in the air.	A girl in a black shirt is standing on a mat.	The girl then begins flipping around the beam and ends by jumping off the side and walking away.
The little girl performs three back flips in the air, after she jumps.	A girl in a black shirt is standing on a mat.	The girl then flips herself over her feet and does several back flips on the mat.
The girl flips but she falls, then she stands and does cartwheels and continues doings flips and dancing.	A girl in a red shirt is standing in a large room in a large gymnasium.	The girl then flips herself over her feet and does several flips and tricks.

Figure 10: More qualitative dense-captioning captions generated using our model. We show captions with the highest overlap with ground truth captions.

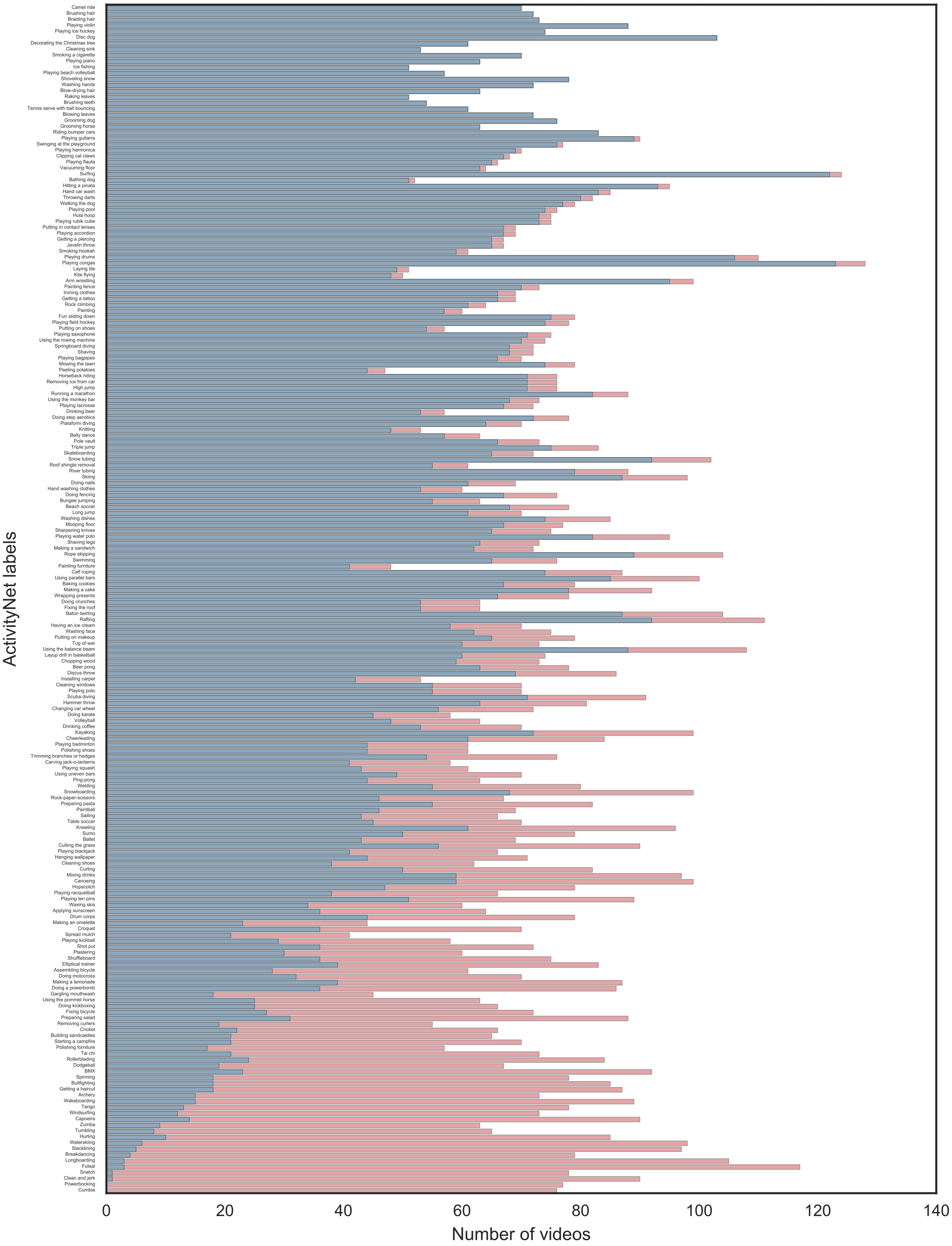


Figure 11: The number of videos (red) corresponding to each ActivityNet class label, as well as the number of videos (blue) that has the label appearing in their ActivityNet Captions paragraph descriptions.