VERIFIED: A Video Corpus Moment Retrieval Benchmark for Fine-Grained Video Understanding (Dataset Document)

A Datasheets

1. **What is the name of the dataset?**
   The names of the datasets are Charades-FIG, DiDeMo-FIG, and ActivityNet-FIG. They contain fine-grained video moment captions for videos collected in previous Charades-STA [1], DiDeMo [2], and ActivityNet Captions [3]. They have the same organizational structure. The following descriptions are suitable for each of the datasets.
2. **For what purpose was the dataset created?**
   The dataset was created to advance the research for fine-grained video understanding, including but not limited to video (corpus) moment retrieval, dense video captioning, and video question answering. It should be beneficial to controllable video generation as well and more work exploring this is highly encouraged.
3. **What do the instances that comprise the dataset represent?**
   Each instance consists of a fine-grained text annotation with a confidence score (having selected the annotation with the highest score) for its corresponding video moment.
4. **How many instances are there in total?**
   Charades-FIG contains 16128 instances (12408 in the train split and 3720 in the test split); DiDeMo-FIG contains 41091 instances (32903 in the train split, 4175 in the validation split, and 4013 in the test split); ActivityNet-FIG contains 71656 instances (37261 in the train split, 17429 in the validation 1 split, and 16966 in the validation 2 split).
5. **How was the data collected?**
   The dataset is annotated on the videos collected by the previous dataset.
6. **How was the data labeled?**
   The dataset is labeled by the proposed VERIFIED pipeline.
7. **Was the raw data saved?**
   The access to the raw videos should comply with the previous video collector, but we will provide ResNet152 features extracted by ourselves, which are accessible.
8. **What are the intended uses of the dataset?**
   The dataset is primarily intended for training and evaluating the fine-grained video corpus moment retrieval task but is also suitable for other video tasks.
9. **Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?**

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Yes, the dataset will be available under a Creative Commons license.

10. **Who is supporting/hosting/maintaining the dataset?**
The dataset is maintained by the research lab that created it.

B Metadata

The dataset is organized by JSONL, where each instance occupies a JSON body. For each instance, the field descriptions contain:

1. "desc_id": a unique ID for each instance.
2. "video": the URL of the video.
3. "duration": the duration of the video.
4. "time": the start and end timestamps(unit: second) of the annotated video moment.
5. "cog_desc": the previous coarse-grained annotation text.
6. "fig_desc": the fine-grained annotation text generated by our proposed VERIFIED pipeline.
7. "fig_desc_score": the confidence score for the fine-grained annotation text.

D Data License

References

[1] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[3] 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] https://prior.allenai.org/projects/charades
[2] https://github.com/LisaAnne/LocalizingMoments
[3] http://activity-net.org/