Appendix:

Datasheets for YOVO-3M and YOVO-10M

Motivation

1. For what purpose was the dataset created?
A: The two web video datasets are created to enable research on weakly-supervised video representation learning. The searched query and the video title are included to provide additional supervision for model training. The two datasets are created intentionally with the task in mind.

2. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
A: The two datasets were created by Fuchen Long and Zhaofan Qiu at JD AI Research.

3. Who funded the creation of the dataset?
A: Funding was provided from the National Key R&D Program of China.

Composition

4. What do the instances that comprise the dataset represent?
A: The instances are the video clips paired with the searched queries and video titles. The video clip is in AVI format. The query and video title are in text format.

5. How many instances are there in total?
A: There are 2,958,092 instances in YOVO-3M and 10,023,532 instances in YOVO-10M.

6. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
A: The two datasets are samples of instances. The instances are sampled from the large pool of the YouTube uploaded videos.

7. What data does each instance consist of?
A: Each instance consists of the video clip, the searched query and the corresponding video title. The queries and most of the titles are in English.

8. Is there a label or target associated with each instance?
A: There is no explicit label or target, and the annotation is weak.

9. Is any information missing from individual instances?
A: Yes. Small quantities of videos might be missing or taken down from YouTube.

10. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?
A: No. All user information is removed and only the information of video itself, i.e., query and video title, is remained.

11. Are there recommended data splits (e.g., training, development/validation, testing)?
A: There is a recommended training and validation split for video feature learning.

12. Are there any errors, sources of noise, or redundancies in the dataset?
A: Yes, there is noise in the weak supervision of web videos. Please see preprocessing.

13. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
A: The two datasets are not self-contained. They rely on the resources of YouTube videos.

14. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
A: A minority of the raw videos/titles may contain little of inappropriate contents, e.g., violence, and we remove them in the datasets during the dataset cleaning.

15. Does the dataset identify any subpopulations?
A: No.

16. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
A: No. All the personal information of the video owner has been removed in the data preprocessing.

17. Does the dataset contain data that might be considered sensitive in any way?
A: No. The data with inappropriate contents has been be removed.

Collection Process

18. How was the data associated with each instance acquired?
A: We collect the queries from the labels of existing datasets and the Oxford English Dictionary. After that, we issue each query to YouTube search engine, and download the searched videos with the corresponding video titles. The clip is random sampled in the video.

19. What mechanisms or procedures were used to collect the data?
A: The API of YouTube search engine.

20. If the dataset is a sample from a larger set, what was the sampling strategy?
A: When searching videos through one query in YouTube, the top ranked videos are sampled or downloaded in priority.

21. Who was involved in the data collection process and how were they compensated?
A: The dataset creators are involved in the query, title and video processing for the collection of the two datasets.

22. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?
A: As described above, the two datasets were collected from YouTube instead of collecting from individuals.

23. Were the individuals in question notified about the data collection?
A: No. The data was crawled from public web sources. The video owners presumably knew that their videos would be public, but the owners were not explicitly informed that the videos will be used in this way.

24. Did the individuals in question consent to the collection and use of their data?
A: No. Please see the previous question.

Preprocess/cleaning/labeling

25. Was any preprocessing/cleaning/labeling of the data done?
A: We collect the queries from the labels of existing datasets and the Oxford English Dictionary. Given the extra queries from the Oxford English Dictionary, we employ the Profanityfilter toolbox with a blacklist of the sensitive words to filter out the inappropriate queries. For the searched videos, we first exploit Profanityfilter on the title to remove the video if the title contains sensitive contents. Meanwhile, we parse each word based on a comprehensive vocabulary of BERT to remove the non-English words that do not appear in the list to obtain a new title. Besides, a deep neural network is adopted to detect the adult and violent contents on the key frames of each video, and we remove the video cases containing sensitive contents. Finally, we employ the clip deduplication approach to remove video clips occurring anywhere in the downstream datasets from both of the YOVO-3M and YOVO-10M datasets.

26. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data?
A: No. The datasets only contain the processed video clips with the searched queries and video titles.

27. Is the software that was used to preprocess/clean/label the data available?
A: No.

Uses

28. Has the dataset been used for any tasks already?
A: The YOVO-3M dataset has been used in the workshop challenge of "Pre-training for Video Understanding" with ACM International Conference on Multimedia (ACM MM) 2021 and 2022.

29. What (other) tasks could the dataset be used for?
A: In addition to the weakly-supervised video representation learning, the two datasets could be employed for cross-modality video retrieval, e.g., text-to-video retrieval.

30. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
A: In view that the personal information of YouTube users has been completely removed in our video collection, there is no potential risk privacy or social impact. Nevertheless, some videos might be removed from YouTube making the two datasets smaller. The datasets will be updated on the project website.

31. Are there tasks for which the dataset should not be used?
A: The two datasets should not be used for the unauthorized video-related tasks, e.g., unauthorized surveillance detection.

Distribution

32. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
A: Yes, the two datasets are publicly available on the internet.

33. How will the dataset be distributed (e.g., website, API or GitHub)?
A: The datasets are distributed on the GitHub: https://github.com/FuchenUSTC/BCN/tree/master/datasets.

34. When will the dataset be distributed?
A: YOVO-3M was first released in 2021 and YOVO-10M was first released in 2022.

35. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
A: The two datasets are licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

36. Have any third parties-imposed IP-based or other restrictions on the data associated with the instances?
A: No.

Maintenance

37. Who will be supporting or hosting or maintaining the dataset?
A: Fuchen Long is supporting or maintaining the two datasets.

38. How can the owner/curator/manager of the dataset be contacted?
A: The curator of the datasets, Fuchen Long, can be contacted with the email address: longfc.ustc@gmail.com.

39. Is there an erratum?
A: The initial release (v1.0) is the latest version and there is no explicit erratum.

40. Will the dataset be updated?
A: This will be updated on the Git: https://github.com/FuchenUSTC/BCN/tree/master/datasets.

41. Will older versions of the dataset continue to be supported/hosted/maintained?
A: The older version will be kept if the new version has been updated.

42. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
A: Others can do so and should contact the original authors about the incorporation of fixes or extensions.