Datasheet for FIRE Dataset

Text-to-Video Retrieval FIRE Dataset

By: Pedro Rodriguez <me@pedro.ai>, Mahmoud Azab, Becka Silvert, Linzy Labson, Renato Sanchez, Hardik Shah, and Shane Moon

As part of this study that assesses the validity of multimodal retrieval benchmarks, we collected a dataset that is related to the MSR-VTT and MSVD datasets. These datasets have caption-video pairs. We collected binary annotations indicating whether alternate caption-video pairs were valid; we did not collect additional captions or videos. We call this the FIRE dataset.

Motivation

1. **For what purpose was the dataset created?** *(Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)*

This dataset was created to test assumptions about the validity of current text-to-video retrieval benchmarks based on the frequency of false negatives in the data. Using human annotators to assess the relevance of arbitrary video and caption pairs allows us to assess the scope of the false-negative problem, and propose mitigations to improve the validity of evaluation benchmarks, both in these and similar data sets.

1. **Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

This dataset was created by Meta employees in FAIR, MultiModal Voice Assistant, and Reality Labs. The dataset was a collaboration between Research Scientists, Engineers, Linguists, and Data Specialists. The project was led by Pedro Rodriguez.

1. **Who funded the creation of the dataset?** *(If there is an associated grant, please provide the name of the grantor and the grant name and number.)*

The work was funded by Meta.

1. **Any other comments?**

Composition

1. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *(Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.)*

The data consists of a list of relevance labels. Each relevance label refers to whether a caption is relevant to a specific video. The content from the MSR-VTT and MSVD datasets are only reproduced in our data to the extent needed to link the labels to unique caption and video identifiers (e.g., video IDs or the text of captions when no caption IDs exist).

2. **How many instances are there in total (of each type, if appropriate)?**

The dataset contains 683K instances.

3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *(If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).)*

The data set uses 579K unique caption video pairs drawn from MSR-VTT and MSVD as the two most prevalent data sets used in retrieval. For each, we obtained the top ten test set predictions from three models: CLIP4CLIP, SSB, and Teach-Text. Additional annotations were obtained for approximately 10% of jobs to verify consistency and quality of annotation. In one sense, the dataset contains all possible instances: where all is defined by the predictions of all compared models. In another sense, the dataset is a non-uniform sample of all possible caption-query pairs. Since the objective of our paper is to analyze model predictions, and we do not attempt to characterize how representative the subset is to all caption-query pairs, this is ok.

4. **What data does each instance consist of?** *("Raw'' data (e.g., unprocessed text or images)or features? In either case, please provide a description.)*

Each instance consists of a video ID, a caption ID (or when not available the caption), and a binary relevance label. The dataset is stored in JSON with a schema specified `fire_schema.json` which is provided in supplemental materials. This schema is generated automatically from the python code via `FireDataset.schema_json(indent=2)` To view these, you can use online tools like https://navneethg.github.io/jsonschemaviewer/ or https://codebeautify.org/jsonviewer/c6a219

5. **Is there a label or target associated with each instance? If so, please provide a description.**

Each instance is labeled either as relevant or not relevant. A caption is considered relevant to a video if every concept present in the caption appears in the video, including people, objects, locations, activities, adjectives, quantifiers, and qualifiers.

6. **Is any information missing from individual instances?** *(If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.)*

No information has been removed or redacted from the instances

7. **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** *(If so, please describe how these relationships are made explicit.)*

There are no relationships between individual instances.

8. **Are there recommended data splits (e.g., training, development/validation, testing)?** *(If so, please provide a description of these splits, explaining the rationale behind them.)*

The dataset is relevant to the test splits of MSR-VTT and MSVD as defined in the cited text-to-video retrieval literature. It does not per se fall into a specific fold, but for the purpose of this document consider it the test fold.

9. **Are there any errors, sources of noise, or redundancies in the dataset?** *(If so, please provide a description.)*

Some instances were reviewed multiple times in order to verify annotation quality, as well, some additional duplicate labels were obtained from a failure to deduplicate caption video pairs from two model outputs for the MSVD dataset. There is no evidence of data quality issues.

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** *(If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)*

The dataset is not totally self-contained. By design, the dataset minimally reproduces the MSR-VTT and MSVD dataset (e.g., we do not host videos). There are no guarantees that the externally hosted versions of these datasets will continue to exist and to our knowledge there are no official archival versions. We are not aware of explicit restrictions on the use of these datasets.

11. **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?** *(If so, please provide a description.)*

The data set consists of public, open-source data with no privacy restrictions.

12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *(If so, please describe why.)*

The data set was vetted for any material that might be offensive or inappropriate, such as the following: Suicidal Content Self Harm Credible Threats (Non terrorism / terrorism) Human Trafficking Sextortion Grooming CEI (Child Exploitation) Sex Offender Prostitution involving minors only Regulated goods

Since we do reproduce captions where caption IDs are not available, we cannot 100% guarantee that captions do not contain data of the type described in the question. The vetting referred to above represents our best effort, which was implemented by training raters to reject annotation jobs meeting any of these conditions (the resulting rejections were not included in our dataset).

13. **Does the dataset relate to people?** *(If not, you may skip the remaining questions in this section.)*

The data set contains videos of people.

14. **Does the dataset identify any subpopulations (e.g., by age, gender)?** *(If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.)*

The descriptive captions may at times refer to persons by age, gender, etc. With regards to sensitive categories, the raters were instructed to accept the caption as accurate unless they had compelling, concrete reasons to believe otherwise (e.g. a little baby should be not considered old, and octogenarians with white hair and wrinkled skin should not be considered young); raters should not attempt to make more fine-grained distinctions. In particular, they were instructed not to make any assumptions about gender and accept the gender described by the caption.

15. **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?** *(If so, please describe how.)*

There is usually no personally identifying information in the MSR-VTT and MSVD videos, such as names, addresses, phone numbers, or other identifying characteristics. That said, there are videos where one could plausibly identify specific individuals. For example, some videos are from a music competition TV show, so contestants and judges could be identified.

16. **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** *(If so, please provide a description.)*

Given the scale of the original datasets, we cannot guarantee the absence of content since we did not exhaustively annotate for this purpose. However, dataset at large does not contain personal information beyond what can be visually inferred from the videos, e.g. hair color, apparent gender, age, etc.

17. **Any other comments?**

Collection Process

1. **How was the data associated with each instance acquired?** *(Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.)*

The data instances are composed of a few parts:

1. The video_id corresponding to the video, these are from the original MSR-VTT and MSVD datasets
2. The caption_id or caption, these are from the original MSR-VTT and MSVD datasets
3. The relevance label, which annotators create.
4. Additionally, which (caption, video_id) pairs were annotated was deciding by taking the top ten predictions from three models: TeachText, SSB, and CLIP4CLIP (info in associated paper). The TeachText model was taken from the checkpoints on their project site at https://github.com/albanie/collaborative-experts. The CLIP4CLIP and SSB models were re-trained until convergence to previously published results.
5. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** *(How were these mechanisms or procedures validated?)*

The judgments were collected by trained human evaluators on an internal annotation platform (screenshot of UI in paper). In addition to spot checks by the paper's authors, 10% of the instances were submitted for a second evaluation, if the evaluations did not agree then a third evaluation was done to resolve the judgment. Agreement between evaluators was over 90%.

1. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

The selected caption-video pairs were selected by taking the top 10 predictions of recent state-of-the-art text-to-video retrieval models.

1. **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Crowd annotators annotated the data. The paper authors are Meta full time employees or contractors. We cannot disclose compensation as this is confidential information.

1. **Over what timeframe was the data collected?** *(Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.)*

MSR-VTT was collected in 2016. MSVD in 2010. The judgments in our dataset (the FIRE dataset) were collected in between February 2022 and the end of May 2022. The data collection was not continuously running throughout this period, but proceeded as additional jobs were prepared for annotation.

1. **Were any ethical review processes conducted (e.g., by an institutional review board)?** *(If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.)*

A Meta internal review board that oversees privacy and data collection reviewed our experiment.

1. **Does the dataset relate to people?** *(If not, you may skip the remaining questions in this section.)*

Although the data we publish does not contain videos of people, it is tied to MSR-VTT and MSVD which do contain videos (and captions) of people.

1. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

Our dataset was collected via crowdsourcing. The MSR-VTT and MSVD were collected via third parties. Specifically, copies of the datasets are available at: (1) MSVD: https://www.cs.utexas.edu/users/ml/clamp/videoDescription/ and (2) MSRVTT: https://github.com/ArrowLuo/CLIP4Clip

1. **Were the individuals in question notified about the data collection?** *(If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.)*

We do not know if individuals in the MSR-VTT and MSVD videos were notified.

1. **Did the individuals in question consent to the collection and use of their data?** *(If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.)*

We do not know if individuals in the MSR-VTT and MSVD videos consented.

1. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** *(If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).)*

We do not know if individuals in the MSR-VTT and MSVD videos were given such a mechanism.

1. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** *(If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.)*

No.

1. **Any other comments?**

Preprocessing/cleaning/labeling

1. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** *(If so, please provide a description. If not, you may skip the remainder of the questions in this section.)*

No preprocessing of the dataset. We made every effort to preserve the original captions and videos.

1. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *(If so, please provide a link or other access point to the "raw" data.)*

Yes, our data pipeline follows these rough steps: (1) collection on an internal annotation platform, (2) storage on internal data platform, (3) extraction of low-level annotation jobs via SQL (no aggregation, row-level, job-level extraction) which still contains internal identifiers, and (4) conversion of internal version of data that strips internal-only identifiers (such as annotator IDs) and into the JSON schema specified earlier. Excluding the sanitizing of internal-only identifiers, the conversion is non-destructive and reversible. This process occurs in `mmr.tasks.ProcessHaloAnnotations`.

1. **Is the software used to preprocess/clean/label the instances available?** *(If so, please provide a link or other access point.)*

We aim to open source code and data by the June 16 supplemental materials deadline, pending internal publication review. Information will be available at https://www.pedro.ai/multimodal-retrieval-evaluation

2. **Any other comments?**

Uses

1. **Has the dataset been used for any tasks already?** *(If so, please provide a description.)*

Beyond the uses described in our paper, the FIRE dataset has not been used before. In this work, it is used to assess validity of the existing benchmark data sets for text-video retrieval

1. **Is there a repository that links to any or all papers or systems that use the dataset?** *(If so, please provide a link or other access point.)*

There is no repository, as this work introduces the FIRE dataset.

1. **What (other) tasks could the dataset be used for?**

The FIRE dataset in principle could be used to train models, although this would be unusual since it would mean training on the test set.

1. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** *(For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?)*

As discussed in the paper, the FIRE dataset provides annotations for a specific set of model predictions. Were a new model to be created and compared to models used in this work (e.g., CLIP4CLIP), it is possible that even if the new model is truly better, the metrics would not reflect this since the new model may have false negatives in its predictions. Therefore, comparing new models using our data should be done with care.

1. **Are there tasks for which the dataset should not be used?** *(If so, please provide a description.)*

The dataset should likely not be used to train new models, since that means training to test data.

2. **Any other comments?**

Distribution

1. **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?** *(If so, please provide a description.)*

The dataset will be released along with the paper for public review

1. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** *(Does the dataset have a digital object identifier (DOI)?)*

The dataset will be distributed via: (1) an open source github repository and (2) https://www.pedro.ai/multimodal-retrieval-evaluation.

1. **When will the dataset be distributed?**

2. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *(If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.)*

The code will be licensed under Apache 2 and the data will be licensed under a Creative Commons Attribution-ShareAlike 4.0 International license.

1. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** *(If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.)*

2. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** *(If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.)*

Not to our knowledge

1. **Any other comments?**

Maintenance

1. **Who is supporting/hosting/maintaining the dataset?**

The dataset will be maintained by the first author (Pedro Rodriguez) and is supported by Meta.

1. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

They primary contact (Pedro Rodriguez) is reachable at me@pedro.ai

1. **Is there an erratum?** *(If so, please provide a link or other access point.)*

Currently, no. As errors are encountered, future versions of the dataset may be released (but will be versioned). They will all be provided in the same github location.

1. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances')?** *(If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?)*

Errors in the dataset can be reported via GitHub and new versions of the dataset will be released there as well.

1. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** *(If so, please describe these limits and explain how they will be enforced.)*

The dataset does not have retention controls beyond that which is in the original MSR-VTT and MSVD datasets.

1. **Will older versions of the dataset continue to be supported/hosted/maintained?** *(If so, please describe how. If not, please describe how its obsolescence will be communicated to users.)*

Yes; all data will be versioned.

1. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** *(If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.)*

Errors may be submitted via the bugtracker on github. More extensive augmentations may be accepted at the authors' discretion.

1. **Any other comments?**