Lecture Presentations Multimodal Dataset: Dataset Documentation

Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, Louis-Philippe Morency
MIT, Carnegie Mellon University
`https://github.com/dondongwon/LPMDataset`

We utilize the Datasheet for Dataset [7] framework to provide a detailed documentation MLP Dataset.

1 Motivation For Datasheet Creation

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. MLP Dataset was created as a step towards building AI systems that can understand lecture presentations, and enable intelligent AI tutoring systems that can explain and synthesize lecture slides. Previous lecture datasets [9, 6, 2, 5, 3, 14] do not offer aligned multimodal (language and visual) data. To the best of our knowledge, MLP Dataset is the first large-scale dataset which offers clean slide segmentation, figure annotations and aligned modalities (spoken language, slide text, and visual figures).

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The dataset has been created by researchers from the MultiComp lab at the Language Technologies Institute, Carnegie Mellon University.

Who funded the creation dataset? The dataset was funded by National Science Foundation (Awards #1750439 #1722822), National Institutes of Health and NTT Japan.

Any other comment? No.

2 Datasheet Composition

What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges) Each instance consists of a slide, its visual figures, on-slide text, and accompanying spoken language. An overview can be seen in Figure 1 in the main paper.

How many instances are there in total (of each type, if appropriate)? There are 9031 slides and spoken language segments. We have 10 total speakers, we list the number of slides for each speaker in the following: 'anat-1': 1142, 'anat-2': 84, 'bio-1': 717, 'bio-3': 629, 'bio-4': 1258, 'dental': 2588, 'ml-1': 894, 'psy-1': 138, 'psy-2': 275, 'speaking': 959. Our dataset also consists of 8598 figures. Among these figures, 3877 (45.1%) are natural images, 4018 (46.7%) are diagrams, 301 (3.5%) 108 are tables, 402 (4.6%) are equations.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? Our dataset consists of 10 different speakers in 6 different domains (anatomy, biology, dentistry, machine learning, psychology, speaking). The larger set would consist of all topics taught in all classrooms, therefore, our dataset is a sample. An automatic annotation pipeline can be used to extend the dataset using a Layout Parser such as [12], and a scene change detector such as [11, 8]. However, for our dataset, we offer manually annotated clean figures and slides.

Preprint. Under review.

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 in MLP dataset consists of the following:

- slide image: the raw image of the slide
- slide spoken language: the accompanying spoken language
- slide object character recognition (OCR): the written text on the slide (bullet points, text in diagrams, etc)
- slide mouse traces: the x,y coordinate of the mouse traces on the slide
- annotations of figures: the bounding boxes and labels (diagram, equation, table, natural image) of the figure on slide

Is there a label or target associated with each instance? If so, please provide a description. For figure-to-text retrieval, the source instance is the annotation of figure and the target is the slide spoken language. For text-to-figure retrieval the source instance is the slide spoken language and the target is the annotation of figure.

Is any information missing from individual instances? The following can be missing if they were unavailable: slide spoken language, object character recognition, mouse trace and annotations of figures.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? Yes, each speaker has a 'json' file (i.e 'anat-1.json') that compiles all of the above mentioned individual instances together such that each instance is given an id, with dictionaries which contain all of the above mentioned information. We also provide the specific speaker, lecture number and slide number for each instance as well.

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. In our baselines, we split the data for each speaker and implement an 80-20 train/test split, as we want to develop speaker specific models. For future research, the data can be split up such that it is trained on one speaker and tested on another, or trained on earlier set of lectures and tested on later lectures. We offer these information such that it is readily available for future use.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Yes, there are some sources of redundancies and noise in the dataset. Firstly, though rare, there may be redundant figures,,a speaker may use the same figure throughout multiple slides. All automatically extracted features could contain noise, such as the spoken language from google ASR [4], object character recognition from Tesseract [13] and automatically extracted mouse traces. We report the relevant validation metrics respectively in the main paper. The slide segmentation and annotation of figures, however, is internally corrected and validated and we are not aware of any errors or noise.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The dataset is self-contained.

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)? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? MLP Dataset contains lecture information from biology, anatomy, dentistry, which may contain explicit graphical information. Furthermore, there may be biases that are present in lectures.

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 Our dataset consists only of speakers who are teachers.

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? It is possible to identify the speaker directly from the audio, or their video recording. However, all of the videos that are included in our dataset are publicly available data, according to the Terms of Service by Youtube which users agreed to when the videos were uploaded.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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)? No.

3 Collection Process

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 the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how We utilized the natural segments of slides for each instance, which includes spoken language, on-slide figures and text, as well as mouse traces that the presenter utilizes. These are directly observable from the raw lecture videos.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated? There were both manual human annotation and automatic preprocessing. We had humans (1) find lecture videos which used a slide-based presentation format, (2) find the slide segments, (3) manually mark with bounding boxes and label the different figure types. All manual annotations were were validated by an internal team of annotators at CMU. In terms of automatic processing, we used youtube-dl for downloading large-scale video data, Google ASR for automatic speech-to-text transcriptions, Tesseract for Object Character Recognition for on-slide texts, and we utilized the frames differences to automatically detect mouse traces. These automatic annotations were further validated and these result are reported in Section 3.3.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Candidate english educational videos were downloaded from YouTube spanning all topics. From the initial list, we filtered and curated a smaller list of 10 speakers according to the following criteria: (1) they must be presenting the material in a slide-based style, (2) the slides must be stationary (i.e. external video clips cannot be played), and (3) the speaker makes use of their mouse to refer to specific figures on the slide. After filtering, 334 videos remained.

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)? CMU Multicomp Lab's staff, students and crowdworkers were involved with the data collection process. CMU staff and students are paid approximately $12 per hour or receives academic units counting towards their course requirments. Crowdworkers were paid an hourly wage of $8 dollars per hour. For the task of slide segmentation, as annotators are simply required to scroll through the video to find transition points, we pay 50 cents for a 15 minute long video (i.e $2 for an hour long video). We pay annotators a total amount of $856.95 for this task. For the task of figure annotations, we pay the annotators 5 cents per slide, where annotators are expected to spend around 20 seconds per slide. As a result, we spent $451.55 for a total of 9031 slides.

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. The data was collected from January 2020 to December 2020.

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. No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? The video clips in our dataset was downloaded from YouTube.

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. Individual consent was not obtained, however, all of the videos that are included in our dataset are publicly available data, according to the Terms of Service by Youtube which users agreed to when the videos were uploaded. The copyright of all video clips used in our dataset belongs to YouTube and corresponding channels. However, we believe that our dataset is consistent with their "Fair Use" policy. Our dataset is very similar in nature with many other datasets that consist of Youtube videos [1, 10, 15, 16]. For manual annotations, through the Amazon Mechanical Turk Participation Agreement, MTurk Workers consented to have their responses recorded.

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). N/A

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.

4 Collection Process

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)? As outlined in the main paper Section 3.3, and Documentation Section 3, we had humans (1) curate a list of lecture videos which used a slide-based presentation format, (2) find the slide segments, (3) manually mark with bounding boxes and label the different figure types. For automatic preprocessing, we used youtube-dl for downloading large-scale video data, Google ASR for automatic speech-to-text transcriptions, Tesseract for Object Character Recognition for on-slide texts, and we utilized the frames differences to automatically detect mouse traces.

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. We offer a csv with links to the raw videos, which can be downloaded with youtube-dl. This csv can be found at our repository: `https://github.com/dondongwon/LPMDataset`. It is named 'raw_video_links.csv'.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point. The automatic preprocessing pipeline will become available at `https://github.com/dondongwon/LPMDataset`.

5 Uses

Has the dataset been used for any tasks already? If so, please provide a description. We utilize this dataset for the tasks of crossmodal retrieval (figure-to-text and text-to-figure). Detailed experiments can be found in Section 4 of the main paper.

Is there a repository that links to any or all papers or systems that use the dataset? Our repository can be found here: `https://github.com/dondongwon/LPMDataset`.

What (other) tasks could the dataset be used for? In addition to crossmodal retrieval, our dataset can be used for the following: (1) layout detection (2) automatic video scene segmentation (3) slide captioning (4) slide generation (5) masked language modelling (6) Span Selection QA.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? No, there is minimal risk from lecture videos.

Are there tasks for which the dataset should not be used? If so, please provide a description. Our dataset should not be used to generate fake and misinforming educational videos.

6 Distribution

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. Yes, the dataset is publicly available at `https://github.com/dondongwon/LPMDataset`. The repository will be updated with automatic pre-precoessing pipelines and baseline codes by the Neurips 2022 Camera-ready.

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 github and google drive: `https://github.com/dondongwon/LPMDataset`. We will make the DOI available after finalizing the GitHub repository.

When will the dataset be distributed? The dataset is already available.

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.

Unless noted otherwise, we are providing the contents of our repository under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License License (for data-like content) and/ or BSD-2-Clause License (for software-type content).

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. At the time of release, all videos included in this dataset were being made available by the original content providers under the standard Youtube License.

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. No.

7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The dataset will be supported, hosted, and maintained by Dong Won Lee and the MultiComp Lab at CMU.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? All comments and questions regarding this dataset can be sent to the data manager email: datasetmlp@gmail.com or raise an issue in the github repository.

Is there an erratum? If so, please provide a link or other access point All changes will be announced on `https://github.com/dondongwon/LPMDataset`.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the 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 data will be made available according to the Youtube 'Fair Use' Policy.

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 dataset consumers All versions of MLP dataset will be supported and maintained at `https://github.com/dondongwon/LPMDataset`. Any updates will be posted in the Github repository.

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 dataset consumers? If so, please provide a description. Others can contact the author of the paper for extension or contribution. The proposed extension and/or contribution will be reviewed and discussed to confirm the validity. If so, we will share a new version of the dataset on Github.

8 Metadata

Title	Multimodal Lecture Presentations (MLP) Dataset
Description	A dataset to develop AI models capable of understanding multimodal information present in lecture slides
Keyword	multimodal learning, crossmodal retrieval, education, vision and language, intelligent tutoring systems
modified	06/15/22
publisher	MultiComp Lab, Carnegie Mellon University
contactPoint	Dong Won Lee, datasetmlp@gmail.com
accessLevel	public
license	CC BY-NC-SA 4.0

Table 1: Metadata of MLP Dataset

Table 1 displays the metadata of MLP dataset.

9 Responsibility

The authors bear all responsibility in case of violation of rights, etc. and confirm that the dataset is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. Vlengagement: A dataset of scientific video lectures for evaluating population-based engagement. *arXiv preprint arXiv:2011.02273*, 2020.

[3] Huizhong Chen, Matthew Cooper, Dhiraj Joshi, and Bernd Girod. Multi-modal language models for lecture video retrieval. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1081–1084, 2014.

[4] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE, 2018.

[5] Kartik Dutta, Minesh Mathew, Praveen Krishnan, and C. V. Jawahar. Localizing and recognizing text in lecture videos. In *ICFHR*, 2018.

[6] Damianos Galanopoulos and Vasileios Mezaris. Temporal lecture video fragmentation using word embeddings. In *International Conference on Multimedia Modeling*, pages 254–265. Springer, 2019.

[7] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[8] Igor S Gruzman and Anna S Kostenkova. Algorithm of scene change detection in a video sequence based on the threedimensional histogram of color images. In *2014 12th International Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*, pages 1–1. IEEE, 2014.

[9] Irene Li, Alexander R Fabbri, Robert R Tung, and Dragomir R Radev. What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6674–6681, 2019.

[10] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[11] Bindu Reddy and Anita Jadhav. Comparison of scene change detection algorithms for videos. In *2015 Fifth International Conference on Advanced Computing & Communication Technologies*, pages 84–89. IEEE, 2015.

[12] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. Layoutparser: A unified toolkit for deep learning based document image analysis. *arXiv preprint arXiv:2103.15348*, 2021.

[13] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007.

[14] Nhu Van Nguyen, Mickal Coustaty, and Jean-Marc Ogier. Multi-modal and cross-modal for lecture videos retrieval. In *2014 22nd International Conference on Pattern Recognition*, pages 2667–2672. IEEE, 2014.

[15] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.

[16] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021.