Comparative Analysis of the Datasets with Multimodal Content

Maksym Shulha[1], Yuri Gordienko[2], Sergii Stirenko[3]
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine
[1]maksim0shulga@gmail.com, [2]yuri.gordienko@gmail.com, [3]sergii.stirenko@gmail.com

Abstract
Recent works have shown that multimodal content analysis is a very popular task in various applications including healthcare, security, marketing, etc. It can include a lot of subtasks, but joint vision and language understanding is one of the most trendy. It is needed to use some dataset to perform any machine learning task. Nowadays a lot of rich datasets have appeared. In this work we introduce comparison of datasets that are used in joint vision and language understanding tasks. We present a detailed analysis of the modern datasets, compare their basic characteristics, and describe their potential usage for some practical tasks, especially in the context of our previous works.

1.Introduction

Nowadays multimodal content analysis becomes more and more popular. A lot of research has already been conducted in this area. The information collected about objects can be presented in different formats and may not be related to each other. Multimodal data is any specific data about something, presented in different forms and formats – modalities [1, 2]. The sources of initial data for modalities can be, for example, diagrams, graphs, drawings, video recordings, audio recordings. All these data can describe any object. All such data, obtained in different ways, cause us a number of associations that build the so-called associative chains and associative networks.

When performing multimodal content analysis, there are datasets of different modalities in the same subject area, describing the same objects. First of all, in each dataset of individual modalities, homogeneous subsets of similar objects are distinguished. Further, it is necessary to map the obtained subsets of modalities to a single result set. At the same time, it is necessary to set the display rules in such a way as to obtain correspondences between elements of different modalities on the result set. As a result, all data are reflected on a single set, but depending on the presence or absence of data in certain modalities about the object, elements of the result set that are different in fullness can be obtained.

The main aim of this research is to make comparative analysis of the available multimedia datasets that can be used for multimodal content analysis on the basis of artificial intelligence (AI) methods by means of machine learning (ML) methods and hybrid deep learning (DL) models including convolutional (CNN), recurrent (RNN) and other dense (DNN) neural networks. These datasets will be analyzed and characterized in the context of their potential usage for the following practical applications like human health state estimation, multimodal sentiment analysis, emotional recognition in healthcare, security, marketing, etc.

The structure of this paper is as follows. The section *2.Background and Related Work* gives the very short outline of the multimodal content analysis. The section *3.Datasets with Multimodal Content* describes the datasets used here for multimodal content analysis by means of ML/DL approaches. The section *4.Comparison of the Datasets for Multimodal Content Analysis* contains the results obtained as to the comparative analysis of the before mentioned datasets. The section *5.Discussion and Conclusions* is dedicated to discussion of the results obtained and future work planned.

2.Background and Related Work

Multimodality refers to the interaction between a variety of representational modes, for example, images and texts. Multimodal representations arbitrate the sociocultural forms in which these modes are integrated in the process of communication. Nowadays, it is hard to set a limit between listening, watching, reading and writing. People are not only objectives of the communication process but also generators of messages and participants in meaning-making society and networks.

Multimodality as a new complex interdisciplinary research direction aims to study communication in its natural state, taking into account all its manifestations at all levels of interaction: verbal, visual, auditory, tactile, etc. and explores multimodal practices and their products in the form of representations. Multimodal practices can include directly multimodal interaction, necessarily chrono- and spatially conditioned, and discursive practices and their multimodal texts. This understanding of multimodality also requires several approaches to its understanding.

Nowadays, there is a lot of talk about natural language processing - and not only in scientific areas, where this concept is rightly considered fundamental for the further development of artificial intelligence. Natural language processing is a general direction of artificial intelligence and mathematical linguistics. It studies the problems of computer analysis and synthesis of natural languages. When applied to artificial intelligence, analysis means understanding a language, and synthesis means generating literate text. Solving these problems will mean creating a more convenient form of interaction between a computer and a person. The main areas of natural language processing include information retrieval, text sentiment analysis, answering questions, information retrieval, text generation, translation, etc.

Information retrieval refers to the search in an unstructured or weakly structured document for specific facts of interest. Sentiment analysis implies the automatic determination of the emotional background of the text and the identification of the attitude of the person who wrote the text to the subject of discussion. Answering questions can fit both the so-called chatbots, which imitate real communication with people through the transmission of text messages, and special programs that first analyze a certain text, and then answer questions related to its content. Also, one of the most famous and frequently used areas of natural language processing is its translation from one natural language to another.

Among the most interesting and popular methods of natural language processing, one stands apart, which is called sentiment analysis. Sentiment analysis is the determination of the polarity of emotional assessments in the studied text, which contains opinions, judgments, emotions, the author's attitude to entities, personalities, issues, events, themes and their attributes. In simple terms, sentiment analysis answers the question "How does the author of the text relate to this topic?" In this case, the author's attitude can be positive, negative or neutral.

During the analysis of sentiment in the text, words and expressions are identified that have a positive, negative or neutral connotation. People recognize the sentiment of a given text not only on the basis of linguistic knowledge, but also based on the social context. Computers have learned to easily recognize language patterns, but if the interpretation of the sentiment of a text is highly context-dependent, the accuracy of such an analysis performed by a machine is not yet guaranteed.

Multimodal sentiment analysis is a novel approach of classical sentiment analysis which in addition to text analysis involves other modalities like audio and visual features [2, 3]. It can include various combinations of two modalities and be bimodal, or includes three modalities and be trimodal.

Multimodal sentiment analysis has the same major task as classical sentiment analysis. This task is sentiment classification, which categorizes words and expressions into groups that can be positive, negative or neutral. The main difficulty of analyzing text, audio and visual modalities to

accomplish this task requires the application of various fusion techniques, such as feature-level, decision-level and hybrid fusion.

3.Datasets with Multimodal Content

Here we consider tasks that are connected to video-image and language processing. Well known tasks contain visual reasoning, referring expression comprehension, visual commonsense reasoning, visual question answering, visual entailment and visual dialog. In this article we analyze datasets that are used in joint vision and language understanding and existing solutions of this task.

J Liu et al. (2020) introduced a large-scale dataset – VIdeO-and-Language INference (VIOLIN) [4], built upon natural video content with rich temporal dynamics and social interactions. Video clips are collected from diverse sources to cover realistic visual scenes, and statements are collected from crowdsource workers via Amazon Mechanical Turk (AMT), who watched the videos accompanied by subtitles. They wanted to provide a dataset that can test a model's cross-modality reasoning skills over both video and textual signals. They asked AMT workers to write statements based on joint understanding of both video and subtitles, which not only describe explicit information in the video, but also reveal in-depth comprehension of complex plots. Their dataset is differ from others because statements lack an explicit factual description of video scenes, but have deeper inference about it. Also they needed to collect high-quality negative statements without artificial cues or biased priors, they used two strategies for data gathering:

- changing just a few words or phrases in a positive statement, to get negative statement and ensure that the style and length of the statement remain unchanged;
- performing adversarial matching: for each video, select challenging and confusing statements from the statement pool of other videos as the negative ones.

The first strategy guarantees that the collected statements can test a model's in-depth inference ability, because a small part of a positive statement is changed, which makes the model to differentiate highly similar statements with other meanings. The second strategy focuses on testing a model's global understanding of the video, to differentiate statements with high-level scene diversity between videos.

They collected the videos from different sources to increase the coverage and versatility, including 4 popular TV shows of different genres and YouTube movie clips from thousands of movies.

The VIOLIN dataset contains 15,887 video clips, and each video clip is annotated with 3 pairs of positive/negative statements and is 35.2 seconds long on average, resulting in 95,322 triplets in total. Each statement has 18 words on average, and the lengths of positive and negative statements are almost the same, showing no significant bias in length.

To investigate in more detail, for each pair of positive and negative statements, they categorized it into 6 types of reasoning skills required. The types of "visual recognition" (21.3%), "identifying character" (15%) and "action recognition" (17.7%) are more focused on explicit information and require relatively low-level reasoning. "Human dynamics" (18.7%) includes inferring human emotions/relations/intentions, etc. "Conversation reasoning" (20.3%) requires performing inference over characters' dialogues and other forms of interactions (body language, hand gestures, etc.). And "inferring reasons" (7%) is about inferring causal relations in complex events. Overall, "explicit information recognition" makes up 54% of the dataset, and "commonsense reasoning" makes up the remaining 46%, making their dataset a balanced one, imposing new challenges on multi-facet video-and-language understanding.

Compared to other datasets, VIOLIN dataset is more focused on reasoning rather than surface-level grounding. For instance, in TVQA [6] dataset which we will consider further, only 8.5% of the questions require reasoning. J Lei et al. (2019) collected a new large-scale dataset TVQA that is built on natural video content with rich dynamics and realistic social interactions, where question-answer pairs are written by people observing both videos and their accompanying dialogues, encouraging the questions to require both vision and language understanding to answer.

TVQA dataset, built on 6 popular TV shows spanning 3 genres: medical dramas, sitcoms, and crime shows. On this data, they collected 152.5K human-written QA pairs. There are several major benefits of their dataset:

- It is large-scale and natural, containing 21,793 video clips from 925 episodes. On average, each show has 7.3 seasons, providing long range character interactions and evolving relationships. Each video clip is associated with 7 questions, with 5 answers (1 correct) for each question.
- Their video clips are relatively long (60-90 seconds), thereby containing more social interactions and activities, making video understanding more challenging.
- They provided the dialogue (character name + subtitle) for each QA video clip. Understanding the relationship between the provided dialogue and the question-answer pairs is crucial for correctly answering many of the collected questions.
- Their questions are compositional, requiring algorithms to localize relevant moments (START and END points are provided for each question).

With the above rich annotation, their dataset supports three tasks: QA on the grounded clip, question-driven moment localization, and QA on the full video clip.

J Lei et al. (2020) proposed a new dataset TV show Retrieval (TVR) [7]. They selected TV shows as their data resource as they typically involve rich social interactions between actors, involving both activities and dialogues. In total, they collected 108,965 high-quality query-moment pairs on 21,793 videos from 6 long-running TV shows across 3 genres (sitcom, medical, crime), producing the largest dataset of this kind. Videos are paired with subtitles and are on average 76.2 seconds in length. Moments have an average length of 9.1 seconds, and are annotated with tight start and end timestamps, enabling training and evaluating on more precise localization. Compared to existing datasets, TVR has relatively shorter moments and longer queries, it also has greater linguistic diversity. 66% of TVR queries involve at least two people and 67% involve at least two actions, both of which are significantly higher than those of other datasets.

A Miech et al. (2019) introduced HowTo100M [8]: a large-scale dataset of 136 million video clips sourced from 1.22M narrated instructional web videos from YouTube, with activities from domains such as cooking, hand crafting, personal care, gardening, etc, depicting humans performing and describing over 23k different visual tasks. Each video is associated with a narration available as subtitles that are either written manually or are the output of an Automatic Speech Recognition (ASR) system.

HowTo100M is several orders of magnitude larger than existing datasets and contains an unprecedented duration (15 years) of video data. However, unlike previous datasets, HowTo100M does not have clean annotated captions. As the videos contain complex activities, they are relatively long with an average duration of 6.5 minutes. On average, a video produces 110 clip-caption pairs, with an average duration of 4 seconds per clip and 4 words per caption.

4.Comparison of the Datasets for Multimodal Content Analysis

Here we considered a few datasets that are used in joint vision and language understanding tasks and below we will describe some of their main characteristics in the context of their potential usage for the following practical applications like human health state estimation, multimodal sentiment analysis, emotional recognition in healthcare, security, marketing, etc.

Table 1. Comparative analysis of multimodal datasets (RC — Reasoning Categories, AM — Annotation Method, FD — Factual description, Ref — Reference)

Dataset	Modalities - Input	Modalities - Used	Clips (Annotations)	RC	AM	FD	Ref
VIOLIN	video (images + sound) / subtitles / text annotations	video (images) / subtitles / text annotations	15,887 (95,322)	6	Human (AMT*) / Negation / Adversarial	-	[4]
TVQA	video (images + sounds) / subtitles / text annotations / transcripts	video (images) / subtitles / text annotations	21,793 (152,545)	8	Human (AMT*) / Negation / Adversarial	-	[6]
TVR	video (images + sound) / subtitles / text annotations	video (images) / subtitles / text annotations	21,793 (108,965)	-	Human (AMT*)	-	[7]
HowTo100M	video (images + sound) / subtitles	video (images) / subtitles	1.2M (136.6M)	12	Narration	+	[8]
HowTo100M-R	video (images + sound) / subtitles / text annotations	video (images) / subtitles / text annotations	29,843 (67,542)	12	Human (AMT*)	+	[5]
HowTo100M-QA	video (images + sound) / subtitles / text annotations	video (images) / subtitles / text annotations	29,843 (67,542)	7	Human (AMT*) / Negation / Adversarial	-	[5]
ActivityNet Captions	video (images + sound) / text annotations	video (images) / text annotations	20K (100K)	-	Human (AMT*) / Narration	+	[9]
TACoS	video (images + sound) / text annotations	video (images) / text annotations	127 (11,796)	22	Human (AMT*) / Narration	+	[10]
Charades-STA	video (images + sound) / text annotations	video (images) / text annotations	9,848	157	Semi-automatic	+	[11]

VIOLIN is based on TV shows and Youtube videos. It contains videos with human interaction and activities. It is a new dataset so we believe that it will be useful for a lot of tasks because it has rich annotations and videos with diverse human communication. VIOLIN can be used for sentiment analysis, emotional recognition, video-and-language understanding. TVQA is based on trendy TV shows covering 3 genres: medical dramas, sitcoms and crime shows. It brings comprehensive character communications and evolving relationships. Videos in this dataset include more activities and social communications. So this dataset can be used for different tasks like human health estimation, crime scene investigation, security or emotional recognition because it has rich annotation and has medical, crime and human communication background. TVR is based on TVQA, so they share videos. Compared to other datasets except TVQA, TVR has great linguistic diversity, because

almost every query is unique, and includes more characters and actions in each video. TVR can be used for the same tasks as TVQA. HowTo100M is a very rich dataset. It includes a big amount of videos and is larger than any existing video-query dataset. Visual activities from different domains are presented in this dataset which make it useful for usage in a wide number of tasks like sentiment analysis, marketing, promoting goods or advertising. Only videos with physical interaction with objects are presented in this dataset. Both HowTo100M-R and HowTo100M-QA [5] are based on HowTo100M, so they can be used for shared areas. ActivityNet Captions is very useful for detecting and describing events, video retrieval and event localization. Dataset focuses on verbs and actions in videos so it shifts annotations from being object-centric to action-centric. ActivityNet Captions [9] can be used for tasks that include finding interactions between characters or describing different activities in security, sentiment analysis, studying areas. TACOS [10] includes videos with different activities in the cooking area. So it can be used for object or action detection in home/kitchen helper applications. Charades-STA [11] based on videos with a big amount of daily indoor activities. Is it important for security, resource planning and human interaction spheres of life. In the context of our previous works dedicated to estimation of human state and health by wearable electronics [12-15], the multimodal context analysis can be the very promising alternative due to its unobtrusive nature of monitoring and leveraging the more complex semantic meaning of various signs of complex human behavior due to application of ML/DL approaches. As to the above mentioned datasets, from the semantical point of view VIOLIN dataset seems to be the richest one and the most promising background for the human state estimation in health and elderly care applications.

5.Conclusions

We have introduced comparison of datasets that are used in joint vision and language understanding tasks. We have described not only well known datasets, but new ones that have great potential for usage in various tasks. Based on the comparison made we can conclude that TACOS, Charades-STA and ActivityNet Captions can be used for tasks connected with object and action localization or object and action detection. TVR and TVQA are suitable for human interaction detection, emotional recognition and sentiment analysis in a wide number of areas. Taking into account the fact that HowTo100M is a very rich dataset and includes a large amount of videos and annotations, it can be used to solve multiple huge tasks in different areas. VIOLIN is a new dataset and still hasn't been used for ambitious tasks, but we believe that it has a good chance to become popular and widely used, because it is also a rich dataset with a lot of people communications, dynamic actions and social interactions. Here we considered a few datasets that are used in joint vision and language understanding tasks and we will describe some of these tasks in our next publication.

References

[1] Poria, S., Cambria, E., Howard, N., Huang, G. B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, *174*, 50-59.

[2] Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, *53*(6), 4335-4385.

[3] Poria, S., Hussain, A., & Cambria, E. (2018). *Multimodal sentiment analysis* (Vol. 8). Cham, Switzerland: Springer.

[4] Liu, J., Chen, W., Cheng, Y., Gan, Z., Yu, L., Yang, Y., & Liu, J. (2020). VIOLIN: A Large-Scale Dataset for Video-and-Language Inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10900-10910).

[5] Li, L., Chen, Y. C., Cheng, Y., Gan, Z., Yu, L., & Liu, J. (2020). HERO: Hierarchical Encoder for Video+ Language Omni-representation Pre-training. *arXiv preprint arXiv:2005.00200*.
[6] Lei, J., Yu, L., Bansal, M., & Berg, T. L. (2018). Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.
[7] Lei, J., Yu, L., Berg, T. L., & Bansal, M. (2020). Tvr: A large-scale dataset for video-subtitle moment retrieval. *arXiv preprint arXiv:2001.09099*.
[8] Miech, A., Zhukov, D., Alayrac, J. B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE international conference on computer vision* (pp. 2630-2640).
[9] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Carlos Niebles, J. (2017). Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision* (pp. 706-715).
[10] Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., & Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, *1*, 25-36.
[11] Gao, J., Sun, C., Yang, Z., & Nevatia, R. (2017). Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision* (pp. 5267-5275).
[12] Gang, P., Hui, J., Stirenko, S., Gordienko, Y., Shemsedinov, T., Alienin, O., ... & González, E. A. (2018, April). User-driven intelligent interface on the basis of multimodal augmented reality and brain-computer interaction for people with functional disabilities. In *Future of Information and Communication Conference* (pp. 612-631). Springer, Cham.
[13] Gordienko, Y., Stirenko, S., Alienin, O., Skala, K., Sojat, Z., Rojbi, A., ... & Coto, A. L. (2017, May). Augmented coaching ecosystem for non-obtrusive adaptive personalized elderly care on the basis of Cloud-Fog-Dew computing paradigm. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 359-364). IEEE.
[14] Gordienko, Y., Stirenko, S., Kochura, Y., Alienin, O., Novotarskiy, M., & Gordienko, N. (2017). Deep learning for fatigue estimation on the basis of multimodal human-machine interactions. *arXiv preprint arXiv:1801.06048*.
[15] Gang, P., Zeng, W., Gordienko, Y., Rokovyi, O., Alienin, O., & Stirenko, S. (2019, December). Prediction of Physical Load Level by Machine Learning Analysis of Heart Activity after Exercises. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 557-562). IEEE.