AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "AI BENCHMARKS"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Authors	**Mihai Gabriel Constantin** (University Politehnica of Bucharest) **Mihai Dogariu** (University Politehnica of Bucharest)

This report is part of the deliverable D2.3 - *"AI technologies and applications in media: State of Play, Foresight, and Research Directions"* of the AI4Media project.

You can site this report as follows:

F. Tsalakanidou et al., Deliverable 2.3 - AI technologies and applications in media: State of play, foresight, and research directions, AI4Media Project (Grant Agreement No 951911), 4 March 2022

This report was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media (A European Excellence Centre for Media, Society and Democracy).

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Copyright

© Copyright 2022 AI4Media Consortium

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the AI4Media Consortium. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.
All rights reserved.

AI benchmarks

Current status

The latest trends in AI show an increased attention for the creation and wide adoption of benchmarking competitions. Media data has been constantly featured in different conferences that are dedicated to the creation of common benchmarking systems[1,2,3], that integrate common subject matter definition, data, train-development-test splits, metrics, annotations and pre-processed auxiliary data.

While organisers and participants create and use common metrics in understanding system performance, an interesting development is that benchmark organisers are also interested in the discovery of general trends that go beyond just ranking the systems according to their performance. These trends may be defined by ideas, approaches, or parts of approaches that not only tend to influence the performance of methods, but that also shed some light into the studied concept itself. To this effect, the MediaEval[3] website declares interest in studying the "*Quest for Insight: List several research questions related to the challenge, which the participants can strive to answer in order to go beyond just looking at evaluation metrics.*", while the ImageCLEF[2] website declares that "*The main goal is not to win the competition but compare techniques based on the same data, so everyone can learn from the results. Everyone who has been to a workshop can tell that the discussions at the posters are always very vivid and many approaches that work or do not are discussed. It is a chance for everyone to learn from each other*".

The current status of AI benchmarking systems is also dictated by the emergence and adoption of Evaluation as a Service (EaaS) systems. EaaS is defined as a cloud computing based architecture that can be used for assessing or evaluating a series of systems, by providing the necessary tools for system evaluation[4,5]. This type of approach attempts to increase the ***reliability, trustworthiness and reproducibility*** of the AI methods, as well as ensure a fair chance for each participant to the benchmarking competitions. While in no way an exhaustive list, just a few of the most important and popular EaaS platforms are Kaggle[6], AIcrowd[7] or Codalab[8]. Most of them are open source, allowing their integration in other projects, as well as changes and improvements brought by interested parties, and present many useful functions. Some examples would be:

- ***Leaderboards and scoring***: allowing organisers to create and use their own metrics and methods of scoring participant runs and displaying their performances.

[1] CLEF: http://clef2021.clef-initiative.eu/
[2] ImageCLEF: https://www.imageclef.org/
[3] MediaEval: https://multimediaeval.github.io/
[4] Hanbury, A., et al. (2015). Evaluation-as-a-service: overview and outlook. arXiv preprint arXiv:1512.07454.
[5] Wasik, S., Antczak, M., Badura, J., & Laskowski, A. (2018). Evaluation as a Service architecture and crowdsourced problems solving implemented in Optil. io platform. arXiv preprint arXiv:1807.06002.
[6] Kaggle: https://www.kaggle.com/
[7] AIcrowd: https://www.aicrowd.com/
[8] Codalab: https://codalab.org/

- ***Sub-tasks, multi-phase, and scheduling***: allowing organisers to create multiple sub-tasks with different targets, create multiple phases for the tasks and schedule data publication accordingly.
- ***Computing workers and containerisation***: allowing organisers to create worker cloud-based virtual stations that can be used for the task, and allowing participants to submit containerised versions of their method, thus supporting reproducibility of the results.

Some examples of previous, current or upcoming AI benchmarking campaigns published by the AI4Media project include the following:

- **Interestingness10k**[9] provides participants with image and video samples extracted from Hollywood-like movies, annotated for visual interestingness. The organisers also provide common metrics, data splits, pre-extracted features, as well as an in-depth analysis of previously used methods for this set of data.
- **MediaEval Predicting Media Memorability** 2020[10] and 2021.[11] This task addresses short- and long-term memorability for videos. The authors also provide a common set of metrics, data splits and pre-extracted features. Also, for the 2021 edition of the benchmarking campaign, two different sets of data are provided, and a generalisation and EEG-based subtask are available for interested participants.
- **ImageCLEFaware** 2021[12] and 2022[13] proposes a task where participants are asked to infer the effect of media content sharing on several real-world situations, such as: asking for a bank loan, getting an accommodation, searching for a job as a waiter/waitress, and looking for a job in IT. The organisers provide a common set of metrics, anonymised user profile data, and automatically extracted predictors.
- **ImageCLEFfusion** 2022[14] proposes a benchmarking task for comparing the performance of late fusion or ensembling systems, applied to two media processing tasks, namely media interestingness and image search result diversification. The organisers provide a common set of metrics, data splits, as well as the outputs from a pre-processed set of inducers that will be used by participants as inputs for their fusion systems.

Research challenges

The use of AI benchmarking competitions has increased over the last 10 years, and popular competitions like ILSVRC[15] ended up defining some of the most impactful trends in AI over this timeframe. Some of the most important challenges in this domain can be summed up as follows:

Reproducibility: It is of utmost importance to ensure that the results submitted by participants can be reproduced according to the requirements for the testing set data. This can be ensured

[9] Interestingness10k: https://www.interdigital.com/data_sets/interestingness-dataset
[10] MediaEval PMM 2020: https://multimediaeval.github.io/editions/2020/tasks/memorability/
[11] MediaEval PMM 2021: https://multimediaeval.github.io/editions/2021/tasks/memorability/
[12] ImageCLEFaware 2021: https://www.imageclef.org/2021/aware
[13] ImageCLEFaware 2022: https://www.imageclef.org/2022/aware
[14] ImageCLEFfusion 2022: https://www.imageclef.org/2022/fusion
[15] ImageNet Large Scale Visual Recognition Challenge (ILSVRC): https://www.image-net.org/challenges/LSVRC/

by the integration of containerisation or open source submission that would allow participants to submit their proposed model and allow the community to check their results.

Trustworthiness: Given that many of the most popular AI methods and models are created or gain their popularity during AI benchmarking campaigns, it is important to ensure their trustworthiness, including concepts like explainability, robustness, fairness, transparency and privacy. It is important for future competitions to ensure that these aspects are thoroughly treated by all the participants of the competitions.

Continuity: While it is important to have a clear schedule and draw the main conclusions during the timeframe of the competition, it may also be a good idea to continue and keep the resources open after the competition is done. In this way, organisers can ensure that future users of the dataset attached to the competition respect the same conditions as all the other participants.

Fair access: The growth of AI resource availability helped a lot of researchers test their work; however, there are cases where hardware resources are not available, either through students with low access to GPU computing capabilities or cases when participating in the competition is not a priority at that time.

Efficiency and green computing: This represents another new dimension for measuring the performance of AI systems. Along with measures of trustworthiness and traditional performance metrics, an evaluation of how impactful training and running a system is on the environment creates a better and more accurate picture for system integrators, private companies and authorities with regards to the real performance of AI methods and models.

Future trends for the media sector

Future trends will undoubtedly take into account some of the current challenges and problems currently faced by this domain. The media sector in particular can be helped by the contributions and cooperation of large media companies, social media platforms, either through sponsoring, endorsing or contributing with data to the AI benchmarking competitions.

One important aspect with regards to media companies that are interested in organising benchmarking competitions is represented by data sharing anonymity. While current regulations and industry standards provide guidelines for user data protection, future developments may present a great opportunity for the introduction of fully or partly synthetic datasets, therefore creating a very powerful layer of anonymity.

The media industry will also represent an important factor in setting the definition and use case scenarios for the target concepts. One such current example is the Interestingness10k dataset, where the industry co-organiser proposes a use case scenario that is important from a practical standpoint for their applications or business model, namely helping content creator professionals select the most interesting frames or video excerpts from their movies in order to list them on Video-on-Demand platforms.

Also, it is important to note the responsibility of AI benchmark organisers to ensure that the competitions take into account all the aforementioned aspects. In this regard, there is a clear need for more emphasis on the trustworthiness aspect of AI models.

Goals for next 10 or 20 years

The future of AI benchmark competitions will be defined by the future of EaaS technologies, as these aspects will be closely interconnected and applying EaaS technologies to AI benchmarks will ensure a fair and open competition for all participants. While cloud-based computing has been developed for a long time, the next period may see an increase in its presence in computer vision, natural language processing, and artificial intelligence in general, as the technology may become more ubiquitous and may see a decrease in costs. Therefore, it may become possible for competitions to become more complex, either by integrating certain trustworthiness metrics as a supplementary dimension of performance or by exploring several concepts at the same time and considering their correlations. Also, cloud-based computing power may increase to shared GPU computing that can target fair chances of access to participation and may increase the rate of reproducible methods. From a software standpoint, it will be important for organisers to create open APIs that help participants submit their models in a unified way that allows for countless variations with regards to the software packages they employ, but ensure a standard input and output, therefore easing reproducibility.

AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu