AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "AI DATASETS"

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 datasets

Current status

In the past few years, deep learning has become the de-facto approach in the multimedia field, spanning all types of applications and covering each major type of data, i.e., visual, textual, and audio. This has been strongly encouraged by the superior performances that deep learning methods have over classical machine learning algorithms. Many deep learning theories offer almost ideal results, but they are impossible to put into practice. Similarly, many deep learning models work well on specific tasks, but their creators cannot offer a sound explanation for choosing a specific setup. Therefore, there is still a reasonable amount of mystery concerning this field, where mathematical models lack practical implementation and applications lack a theoretical justification. It is a field where major breakthroughs on one of the two sides can propel the next breakthrough on the other side.

No matter what algorithm we design or implement, we are required to show its applicability through extensive validation. Each state-of-the-art method must prove its superiority when compared against existing research, so it is mandatory to have a common ground for these methods to establish a ranking. Traditionally, this has been done by reporting results on openly available datasets. The likes of MNIST[1], CIFAR-10[2], ImageNet[3], celebA[4], PASCAL VOC[5], MS-COCO[6], SQuAD[7], GLUE[8], Penn Treebank[9], LibriSpeech[10], Universal Dependencies[11], VoxCeleb1[12], CheXpert[13] etc. have become representative benchmarks in their respective fields. They offer

[1] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324

[2] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

[3] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). ImageNet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). IEEE.

[4] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision (pp. 3730-3738).

[5] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2), 303-338.

[6] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[7] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

[8] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

[9] Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank.

[10] Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5206-5210). IEEE.

[11] Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... & Zeman, D. (2016, May). Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1659-1666).

[12] Nagrani, A., Chung, J. S., & Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612.

[13] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M.

the premises for both training and validating algorithms, but they often lag behind the immense diversity that applications have reached. Figure 1 presents some samples from several of these datasets.

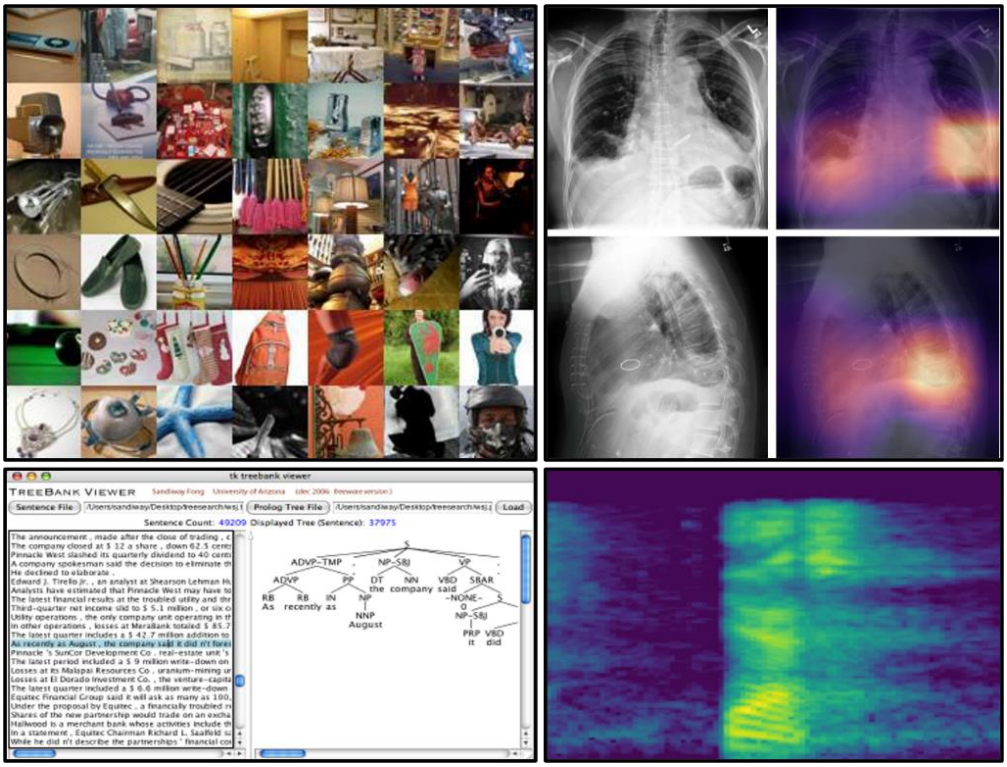

Figure 1: Samples extracted from large-scale public datasets. From left to right, top to bottom, samples from: ImageNet (images), CheXpert (chest X-rays), Penn Treebank (textual), LibriSpeech waveform (audio).

Research challenges

Research in the AI field faces several challenges when it comes to working with datasets. Below, we list the most important ones:

Low dataset diversity: Most AI algorithms have a limited applicability due to the fact that there are no large and diverse enough datasets that correspond to their needs. This leads to training done on other datasets than what is specifically needed for the task at hand. For example, there are numerous algorithms that are pre-trained on ImageNet, even though they do not aim to classify images. One could argue that ImageNet is the most complete large-scale dataset which helps best initialise a network's weights, but it is far from perfect for each computer vision task. In response to this, researchers might opt to create their own datasets, which leads to the following issue.

Complex data gathering process: Creating a relevant dataset involves a great deal of effort from the involved team. Gathering the dataset samples is just the start of a tedious process and, even at this point, critical problems may arise. For example, satellite images are not easy to acquire

P., & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590-597).

since it implies tremendous costs to launch a satellite for this purpose. Therefore, researchers have to deal with whatever resource is available for them. Moreover, even if researchers have hypothetical access to unlimited samples, they still need to annotate them. An accurate annotation of the dataset is critical for the outcome of any application. However, the human resources needed to accomplish such a task is expensive because it requires many persons to perform this process and they also need to be accurately trained. Moreover, the annotator's subjectivity may also interfere with the goodness of the annotations. After such a lengthy process, a large part of the institutions or groups that gather such datasets may not be too keen to share them with others, especially if it gives them a competitive edge, which leads to the next issue.

Closed datasets: The high costs of creating a good dataset may deter researchers from freely sharing their work. This is, most often, the case for companies involved in research that want to ensure an advantage over their competitors. It is by all means understandable that they choose to do so, since this contributes to their source of revenue, but the research results that they report are difficult to validate by external researchers. Even more, a large number of datasets are inaccessible due to privacy concerns, which raises the following point.

GDPR concerns: With the introduction of the GDPR regulations in 2018, most processing that involves personal data has hit a wall in development. While these regulations came to the aid of individuals, they have seriously hindered the gathering of data for numerous datasets. Some examples include face detection or speaker verification. Identity obfuscation becomes, this way, a very appealing field that can help the creation of large-scale datasets.

Bias issues: AI algorithms are known for offering objective results from the decision point of view. However, there is no warranty that this objectivity transfers to the system's global decision philosophy, especially since any bias in the training dataset will be visible in the output. This will skew the system's capabilities in favor of the most frequent/prominent samples of the dataset, leading to falsely accurate results (see section on "*AI fairness*"). For example, Shankar et al.[14] examined two large-scale datasets, Open Images and ImageNet, and discovered that these datasets are highly biased towards the countries where they were gathered. They are highly US and Euro-centric, leaving other geographical areas, such as the entire African continent, severely under-represented. The same authors present a pie-chart of the distribution of geographically identifiable images in these two datasets, shown in Figure 2.

[14] Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:1711.08536.

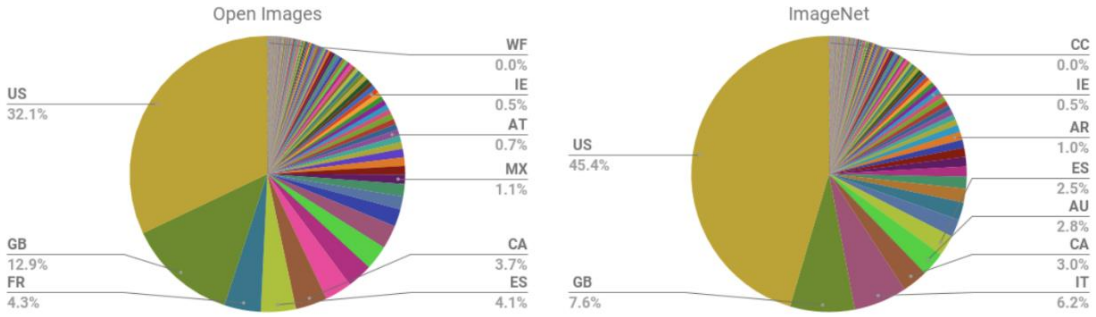

Figure 2: Fraction of Open Images and ImageNet images from each country [Shankar et al, 2017][14].

Proper benchmarking: Large-scale datasets often come with a validation protocol that researchers 'should' follow. Not too rare is the case when researchers will force different training + validation scenarios such that their proposed approach will top the existing state-of-the-art. The current state of worldwide research emphasises outperforming the state-of-the-art to the extent that researchers are tempted to alter the validation process just so they obtain results better than the state-of-the-art. This is usually noticed when trying to replicate the experiments, and end up with unexpected results when following the dataset's official validation protocol.

Future trends for the media sector

AI algorithms evolve at an astonishing pace. It is crucial that these algorithms are trained and validated adequately on large-scale datasets. These datasets lay at the very heart of all learning algorithms and they play a vital role. With people relying more and more on technology, especially on AI systems, we will start feeling the flaws of these datasets in our lives. The media sector can help in this regard since they have access to or are the owners of large data archives, which they could share partly or fully with the research community. Researchers would benefit tremendously from such datasets since they are at least weakly annotated (having a title, short description, metadata available etc.). These data collections also have the advantage of spanning large time periods, so a certain chronological evolution of its constituent samples can be observed, e.g. how fashion changed in time, how the media-specific terms and phrases differ now from the past, how image quality improved, etc.

The multimedia research community is directly involved in the development of most smart applications that are pushed into production, so it should also take responsibility for its drawbacks. In the future, there will be a great deal of emphasis placed on individual privacy and AI invasiveness. Therefore, it is the media sector's responsibility to ensure that any attempt to violate privacy rights will be identified and reported. Investigative journalism can have a major impact on holding accountable companies or states that take invasive actions towards their clients or citizens.

Lastly, tech giant companies will compete for the best large-scale datasets, since they are amongst the few organisations that hold the necessary resources to carry on such an effort (e.g. Microsoft's Common Objects in COntext dataset, Google Open Images, Habitat - Matterport 3D, etc.). There are numerous start-ups, spin-offs and independent businesses that are being born

every day and a small part of them become visible with innovative products. This draws the attention of tech giant competitors which might buy and integrate them into their portfolio, e.g., WhatsApp acquisition by Facebook. Instead of focusing on the record-breaking transactions by these tech giants, the media sector should try to emphasise more the role that smaller companies have in the entire AI tech ecosystem. This way, the effort of rising visionaries will not be diluted under the umbrella of big companies.

Goals for next 10 or 20 years

The future of AI applications is changing with each passing day. Applications that seemed to be unattainable during our lifetime are now already obsolete and the future holds as many new exciting applications as the human imagination can fathom. As previously stated, datasets play a vital role in designing performant algorithms. Therefore, to overcome some of the current setbacks, AI datasets should deal with several aspects:

Open data: The first and most important characteristic of AI datasets is that they should be open for research. The biggest advances in any field came when research was shared with other peers. There are several attempts at the moment to centralise access information on all large-scale multimedia datasets, but the future might offer a centralised way of creating, uploading and sharing datasets among researchers. Data openness (under particular legal constraints) is a must for the future of AI.

Automatic annotations: Simply put, humans do not have the ability to annotate media content at the same rate that it is being produced. Automatic annotations have the potential to solve the limited availability that large-scale datasets have.

Synthetic data generation: Another way of solving data scarcity is to generate completely synthetic samples that resemble or complete the characteristics of the original dataset. This is already a promising approach with the rise of GANs.

Privacy: datasets should be built with the idea of private personal information in mind. This problem seems to become more and more prevalent in every application since reports about consumer applications violating individual privacy are surfacing every day.

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