AI4media
ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "CROSS-MODAL AND MULTIMODAL REPRESENTATION, INDEXING AND RETRIEVAL"

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	**Frederic Precioso** (3IA Côte d'Azur) **Lucile Sassateli (**3IA Côte d'Azur)

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

Cross-modal and multimodal representation, indexing and retrieval

Current status

Multimedia content covers all modalities (visual, audio, text, etc.) and each modality is carrying its own specific piece of information. To gather all these pieces together and explore the whole information all at once, each modality requires to be represented in a common description space where they can all be combined and compared. This new digital representation has to preserve the essence of the information carried by the different modalities.

For decades, each modality representation was addressed with dedicated techniques designed by experts in acoustic for audio, in image processing for images and video, in linguistics for texts, etc. Once multimedia content was converted in these new representations, one had to define the organisation and the storage, i.e. the indexing, of this content such that information retrieval in all the multimedia content processed would be faster and easier.

During the last decade, the methods for representing multimedia content have been revolutionised by the emergence of deep representation learning. The paradigm of data representation has moved from precisely hand-crafted feature extraction to learning the representations as parts of training deep neural network architectures from data. These advances have particularly impacted how to efficiently represent audio data[1], visual data,[2,3,4] video data,[5] or textual data[6].

These new efficient methods have allowed to drastically reduce the transfer of multimedia analysis models from research labs to the market, simplifying the design of new multimedia retrieval systems even by non-experts. By leveraging most challenges of the last decades on multimedia content representation, these methods caused a shift of focus on more challenging tasks, moving from multimodal analysis to cross-modal analysis. In cross-modal analysis, only one modality is exploited to retrieve the content information in all modalities. Visual Question Answering (VQA), i.e. open-ended questions about images requiring an understanding of vision, language and common sense knowledge to answer, is one of these new challenges where the new multimedia content representations allow common sense cross-modality analysis[7].

[1] G. E. Dahl, M. Ranzato, A. Mohamed, G. E. Hinton, Phone recognition with the mean covariance restricted Boltzmann machine. In NIPS. pp. 469-477, 2010.

[2] A. Krizhevsky, I. Sutskever, and G. E Hinton. Imagenet classification with deep convolutional neural networks. NeurIPS, 25:1097–1105, 2012.

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.

[5] D. Tran, L.r Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks. In ICCV, pages 4489-4497, 2015.

[6] R. Collobert and J. Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In ICML, 2008.

[7] Visual Question Answering (VQA): https://visualqa.org/

All these recent advances open new opportunities for developing systems to analyze and retrieve multimedia content, with direct applications in the media industry. The current focus of researchers in the field is to design new gigantic models called transformers as unifying models that receive all available modalities of multimedia content and provide solutions for many different tasks all at once (Transformers are discussed in section "*Transformers for computer vision*" of this Roadmap). However, the design of such gigantic models is a challenge, even more when addressing multimodal data. An example of a deep transformer network for multimedia content is presented in Figure 1.

Until now, the advances in deep learning have not directly benefited the design of a new generation of search engines. Very few recent works investigate this possibility[8], even though the potential is clearly and unanimously identified.

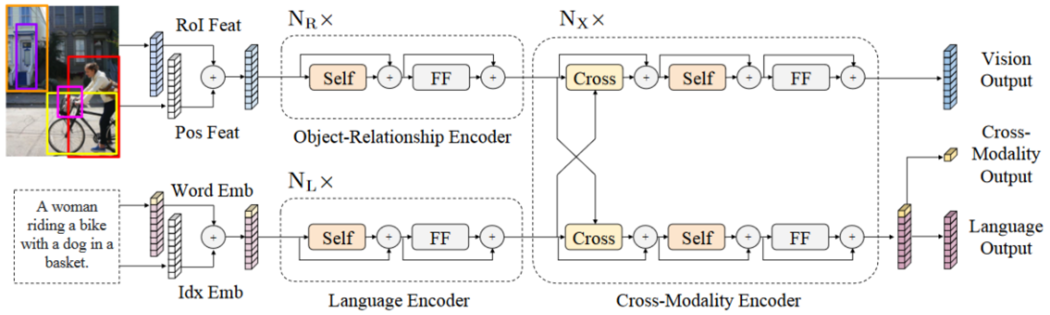

Figure 1: The LXMERT model for learning vision-and-language cross-modality representations. 'Self' and 'Cross' are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. 'FF' denotes a feed-forward sub-layer.

Figure 1: The LXMERT model for learning vision-and-language cross-modality representations[9].

Research challenges

As mentioned in the previous section, very few research works have been recently investigating the potential of integrating deep learning in designing a new generation of search engines, even though this has already been identified as a promising field of research and development.

The recent focus on transformers as central models dealing with all available modalities and addressing many different tasks all at once may be a game-changer. Let us imagine news producers searching audio-visual archives to support a news story with selected video, film writers searching film or script archives to get ideas, users searching the internet to find content that they like or need, music producers searching music with specific characteristics to match with film scenes or a textual story maybe matching some of the lyrics, game developers searching 3D content to find visual assets for a game level, etc.

Two serious concerns with regard to transformer architectures are the resources required both in terms of data to train the models and in terms of computation and energy consumption. This

[8] T. Teofili. Deep Learning for Search. In Manning Publications, June 2019.
[9] Image source: H. Tan and M. Bansal, LXMERT: Learning Cross-Modality Encoder Representations from Transformers, In EMNLP, 2019.

is true for textual data[10] but it will be even more challenging with these models applied to every possible modality and context.

There are several challenges to tackle in order to really leverage deep learning based search engines, or multimedia content retrieval. First, there is not yet any network architecture that reaches consensus to address all the aforementioned standard search contexts as it could be the case now for convolutional neural networks for image classification.[2,3,4] The research works are still at their beginning both in academic labs and in companies (the most recent advances in transformers for natural language processing are led by companies such as OpenAI, Facebook, or Google).

The other challenge with regard to these gigantic models is related to the resources they require for training, as previously explained. One possible path to overcome this issue is to build hybrid models combining symbolic and non-symbolic approaches. This path is already explored, in particular for the visual question answering task, which can be addressed either as a multimodal retrieval task or as a cross-modality retrieval task, integrating external sources of knowledge.[11,12] Furthermore, using external knowledge may help to reduce the amount of required training samples. For instance, if a model can combine given grammatical rules with statistical analysis from textual data, less training data may be required.

Thus, the main research challenges are first to design and train sophisticated models which will require less annotations while combining several modalities, and second to integrate external knowledge to reduce the cost of building these models.

Societal and media industry drivers

Vignette: Multi-modal and cross-modal content search in vast multimedia data lakes

Chloé is a journalist searching a multimedia data lake for multimedia content that could be used for enriching her news story on the beginning of the French presidency of the EU. She is looking for content that could either be complementary sources of information on the main subject, or have a direct link with it, or at least be useful to illustrate the context of the story: she may look for a video of a historical discourse from the first French president of the Commission, Jacques Delors, or retrieve a sentence Jacques Delors has said on the importance of the EU during an international crisis (as it is currently the case with Ukraine and Russia). She may also look for pictures from the participation of President Emmanuel Macron in previous European Summits. She could also explore the recent discourses or recent interventions made by President Macron about the EU. Chloé will thus need several tools: first, a search engine very similar to the current search engines but with more sophisticated cross-modal capabilities, which would allow searching for images given a textual description of the queried content, providing an image to retrieve videos with similar visual content, or providing audio data to retrieve a corresponding

[10] Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and Policy Considerations for Modern Deep Learning Research. Proceedings of the AAAI Conference on Artificial Intelligence, 34(09), 13693-13696. https://doi.org/10.1609/aaai.v34i09.7123
[11] K. Ye, M. Zhang and A. Kovashka. Breaking Shortcuts by Masking for Robust Visual Reasoning. WACV, January 2021.
[12] Q. Wu, P. Wang, C. Shen, A. Dick, A. van den Hengel. Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge From External Sources. In IEEE CVPR, pp. 4622-4630, 2016.

video. A second tool would allow a multimodal analysis of the multimedia content to jointly exploit different modalities, hence providing richer content and a deeper analysis of this content. For instance, a sophisticated joint audio-visual speech recognition solution would allow to efficiently subtitle any conference or discourse and give access to its content by keyword search. This would allow Chloé to quickly find situations of tension between France and another European country in the history of the EU.

To search for multimedia documents, Chloé has access to a professional multimedia data lake where content has been enriched with a lot of metadata (e.g. location and date of acquisition, topics present in the document, event related, media origin, etc.). It is actually easier to retrieve content using such metadata but many documents of the data lake have not been yet enriched with similar metadata. Furthermore, if Chloé wants to retrieve photos similar to the photos she provides to the system, metadata may not be the most relevant modality and a system directly working on the visual content will be more efficient. However, reasoning on metadata may help shortcut exploring the whole data lake and spending too much time on retrieving the best documents through different multimedia modalities. Thus, when Chloé finds new documents without metadata (e.g. a new photo, multimedia documents coming from another data lake, etc.), and she finds matching documents among the ones enriched with metadata, she can then propagate the metadata to the newly found documents (the location may be the same, the persons or the concepts pictured could be the same, etc.). Searching in multimedia content associated with metadata requires more sophisticated techniques combining reasoning and deep learning. This is a challenging new field for research.

Future trends for the media sector

Up to now, most of the contributions from private companies on new AI techniques to exploit multimedia content are made by big tech companies such as the GAFAM (Google-Amazon-Facebook- Apple-Microsoft) or the BATX (Baidu-Alibaba-Tencent-Xiaomi). Media companies are more usually customers of the solutions developed by Facebook and Google. Thus, media companies depend on the GAFAM and do not handle the processes at the core of the information retrieval system; they get what the algorithms designed by the GAFAM provide. It is urgent that media companies invest in this field to drive innovations more precisely towards their own specific needs and vision. Indeed, if the algorithms to retrieve information in huge multimedia lakes are biased, the search will not be accurate or even valid and may carry wrong information. Handling their own algorithms improves the possibility to control their weaknesses.

An efficient solution for audio-visual speech recognition, benefiting from both modalities, would allow to convert the speech in videos (interviews on TV, conferences, etc.) into a huge amount of textual content, a lot easier to search into and retrieve information from it. Another possible output of these multi-modal/cross-modal contexts, would be new information systems and search engines, able to jointly exploit prior knowledge and existing metadata with statistical models. By a simple text query a user could find/retrieve all relevant multimedia content no matter the available modality: a textual query could be matched with a video without sound; music producers searching music with specific characteristics (of rhythm and Minor tune) to match with film scenes; film writers searching film or script archives to get ideas.

Media and entertainment industry (news, film/TV, music, publishing, social media, games, etc.) would directly benefit from efficient content indexing and search. These companies are the ones producing the content but currently they need powerful intermediates to exploit that content. Being independent from external companies, which are also coming on the market of multimedia content producers, sounds both reasonable and also more efficient since no one better than the experts who produced the content could design a more accurate, fast and reliable information search engine.

This would also allow media companies to offer better services for the user to access a targeted content more accurately and fast, while preserving users' privacy. Working on designing specific systems to better exploit multimedia content would allow to automatise tedious tasks such as annotating new content, propagating new information associated to the multimedia content, confronting the search engine to the consistency of data. This would finally allow to enhance current multimedia content analysis workflows, and for each media company to more efficiently exploit and monetise their own content or UGC.

Goals for next 10 or 20 years

A significant goal is the emergence of new search engines for multimedia data that benefit from the advances in deep learning to solve multimodal and cross-modal queries, for instance to solve the well-known problem of visual question answering. Based on current progress in the field, such solutions could be available on the market within the next few years.

Other applications could also concern a better "dialog" between multimedia (i.e. unstructured content) and metadata (structured content) with the emergence of brand new approaches allowing a cross-fertilisation of these two information resources, resulting in enriched and improved content representation (in both unstructured and structured space). This phase will require more time than the previous one and some preliminary solutions may appear on the market within the next 5 to 10 years.

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