Roadmap on AI technologies & applications for the Media Industry

Section: "Learning with scarce data"

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	**Giuseppe Amato** (Consiglio Nazionale delle Ricerche) **Alejandro Moreo Fernandez** (Consiglio Nazionale delle Ricerche) **Hannes Fassold** (Joanneum Research) **Werner Bailer** (Joanneum Research) **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

Learning with scarce data

Current status

The great success achieved during the last decade by deep learning based algorithms has been supported simultaneously by the availability of massive amounts of training data, and high performance GPU powered computing resources. However, there are applications where there is a lack of high quality annotated training sets and performance of AI algorithms in these scenarios is still a challenging issue.

Having a dataset of insufficient size for training usually leads to a model which is prone to overfitting and performs poorly in practice. In many real-world applications based on multimedia content analysis, it is simply not possible or not viable to gather and annotate such a large training dataset. This may be due to the prohibitive cost of human annotation, ownership/copyright issues of the data, or simply not having enough media content of a certain kind available.

This still open problem has been addressed so far using various solutions, which can be roughly classified into the following categories:

Transfer learning: Transfer learning aims at exploiting knowledge acquired while addressing a problem to solve a related but different problem[1] (Figure 1). In the deep learning scenario, this is generally obtained by using a pre-trained model (a deep neural network trained in a certain applicative scenario) to address a new application. For instance, in a multimedia retrieval setting, one can use a pre-trained deep neural network to extract features from images belonging to a scenario different from the one where the pre-training was carried out. In addition, the pre-trained model can be fine-tuned for the new scenario, by using just a few samples available. In fact starting from a pre-trained network, it is possible to achieve high performance in a new (yet similar) scenario with just a few annotated training samples. This procedure moves toward techniques of domain adaptation and few-shot learning described below.

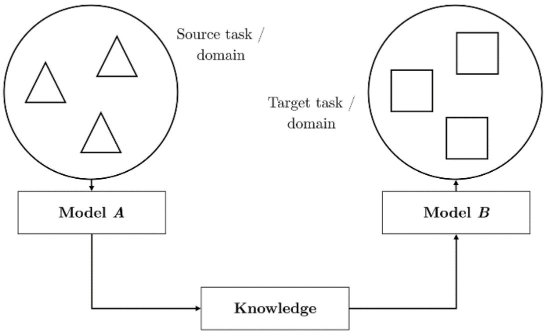

Figure 1: The process of transfer learning[2].

[1] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, Qing He, A Comprehensive Survey on Transfer Learning, arXiv:1911.02685v3, 2020

[2] Image source: S. Ruder, The State of Transfer Learning in NLP (2019) https://ruder.io/state-of-transfer-learning-in-nlp/index.html

Domain adaptation: Most deep learning based methods require a large amount of labelled data and make a common assumption: the training and testing data are drawn from the same distribution. The direct transfer of the learned features between different domains does not work very well because the distributions are different. Thus, a model trained on one domain, named *source*, usually experiences a drastic drop in performance when applied on another domain, named *target*. This problem is commonly referred to as *Domain Shift*[3] (Figure 2). Domain Adaptation is a common technique to address this problem. It adapts a trained neural network by fine-tuning it with a new set of labelled data belonging to the new distribution. However, in many real cases, gathering another collection of labelled data is expensive as well, especially for tasks that imply per pixel annotations, like semantic or instance segmentation. In this respect, solutions of *Unsupervised Domain Adaptation* (UDA)[4,5] can be used, that do not use labelled data from the target domain and rely only on supervision in the source domain. Specifically, UDA takes a source labelled dataset and a target unlabelled one. The challenge here is to automatically infer some knowledge from the target data to reduce the gap between the two domains.

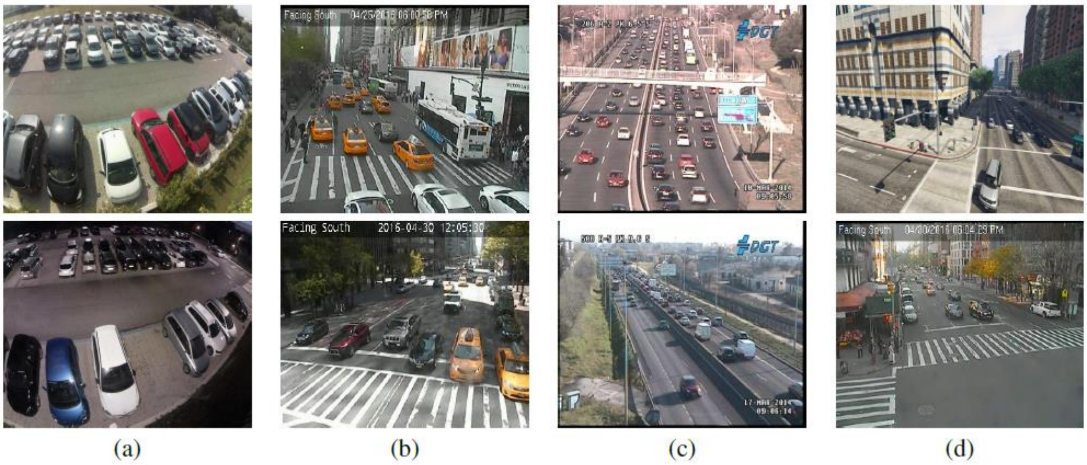

Figure 2: Examples of domain shifts. (a) Day to Night, (b,c) Different points of view, (d) Synthetic to real data[6].

Few-shot learning: Few-shot learning denotes deep learning approaches which are *explicitly* designed to learn from only a few samples per class, starting from a pre-trained model (trained on *base classes*)[7]. Typically, one to ten samples per class are provided for the *novel classes* for which training data is scarce. Few-shot learning methods can be applied to different tasks, like image classification, object detection or semantic segmentation. The existing approaches can be

[3] Torralba, A. and Efros, A. (2011). Unbiased look at dataset bias. InCVPR 2011, pp. 1521–1528

[4] Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. Proceedings of Machine Learning Research, pp. 1180–1189, 2015.

[5] L. Ciampi, C. Santiago, J.P. Costeira, C. Gennaro, G. Amato, Domain Adaptation for Traffic Density Estimation. VISIGRAPP (5: VISAPP), 185-195, 2021

[6] Source image from: L. Ciampi, C. Santiago, J.P. Costeira, C. Gennaro, G. Amato, Domain Adaptation for Traffic Density Estimation. VISIGRAPP (5: VISAPP), 2021, 185-195

[7] Yaqing Wang, Quanming Yao, James Kwok, Lionel M. Ni, Generalizing from a Few Examples: A Survey on Few-Shot Learning, arXiv:1904.05046v3 , 2020

categorised roughly into data augmentation methods, metric learning methods and meta-learning methods. Data augmentation / hallucination methods generate more samples from the few existing ones, e.g. via image synthesis with GANs (Generative Adversarial Networks). Metric learning / embedding methods work by embedding the sample (feature) into a metric space and doing the classification in this space. Meta-learning / optimisation methods aim to pre-train a learner (classifier) so that it can be quickly transformed to the new task setting (e.g., different classes, different domain) in few training steps, enabling it to classify the novel classes.

Self-supervised Learning: Self-supervised learning (SSL) aims at learning from unlabelled data, in a way which is similar to how most of the knowledge learned by humans is believed to be acquired. The idea is to perform a pre-training phase of the network parameters by resorting to an auxiliary task. Examples of these tasks may include trying to reconstruct randomly masked patches from real images[8], trying to identify randomly masked terms from natural language sentences or deciding whether two sentences are consecutive or not[9], to name a few. A notorious case of SSL is called "*Contrastive SSL*" [10,11],and consists of solving an auxiliary task defined upon positive and negative pair examples which are generated in an unsupervised way around the concepts of "same" and "different". The method consists of producing variations of data items (e.g., in the realm of computer vision, by applying rotations, colour or tone variations, etc.) that are assumed to preserve the original class label (e.g., an image of a rotated bird is still to be classified as a bird) despite the fact that the label itself is unknown (e.g., even if the image is not tagged as containing a bird). The problem is then translated to learn to distinguish between the "same" and "different" concepts, by minimising the distance between positive pairs and maximising the distance between negative pairs (the contrastive loss). While the self-supervised phase is typically carried out using as many unlabelled data as possible, the resulting network is ultimately fine-tuned using the (likely few) annotated data at one's disposal.

Synthetic data generation: Although a large amount of annotated data is already available and successfully used to produce important academic results and commercially viable products, there is still a huge number of scenarios where laborious human intervention is required to produce high-quality training sets. To address this problem and make up for the lack of annotated examples, the research community has begun to increasingly leverage the use of programmable virtual scenarios to generate synthetic visual data sets as well as associated annotations. For example, in image-based deep learning techniques, the use of a modern rendering engine (i.e. capable of producing photo-realistic images) has proven to be a valuable tool for the automatic generation of large data sets[12] (Figure 3). The advantages of this approach are remarkable. In addition to making up for the lack of data sets in some particular application domains, these synthetic datasets do not create problems with existing laws about the privacy

[8] Hinton, G. E., Osindero, S., and Teh, Y.-W. A fast learning algorithm for deep belief nets. Neural computation, 18(7):1527– 1554, 2006.

[9] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[10] Jaiswal, Ashish, et al. "A survey on contrastive self-supervised learning." Technologies 9.1 (2021): 2.

[11] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

[12] M. Di Benedetto, F. Carrara, E. Meloni, G. Amato, F. Falchi, C. Gennaro, Learning accurate personal protective equipment detection from virtual worlds, Multimedia Tools and Applications 80 (15), 23241-23253

of individuals related to facial detection, such as the General Data Protection Regulation (GDPR). These techniques have been successfully used in scenarios that include self-driving cars, the detection of safety equipment, the detection of fire arms, etc. They can also be of great help in the media sector to quickly develop solutions able to analyse data to classify and recognise events, objects, actions, which were not considered before and for which no training sets are readily available.

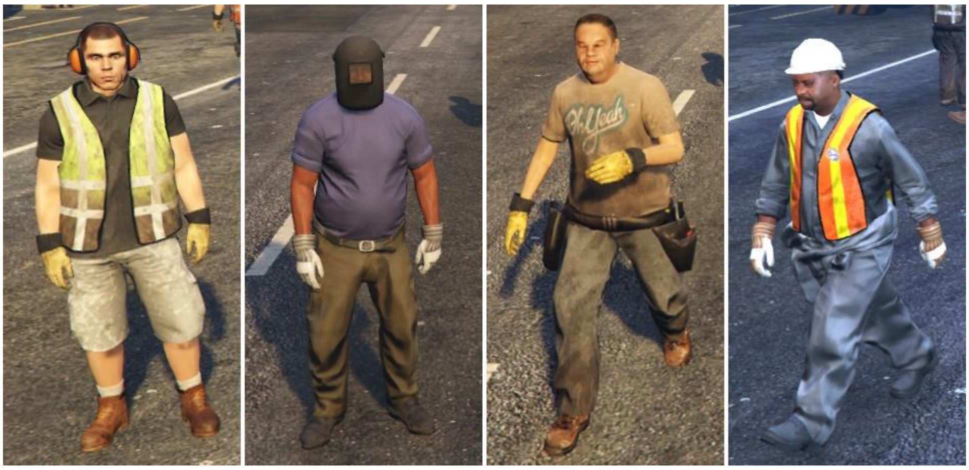

Figure 3: Examples of synthetically generated images that can be used for training AI models[13].

Research challenges

Research during the last decades has provided significant advancements in AI. However learning in scenarios with scarce data, still represents a serious challenge and an opportunity for further research directions.

Nowadays, plenty of pre-trained models specialised to work with specific media and specific applicative scenarios are available. One difficulty, typically encountered, is the choice of the best pre-trained model to be used as a starting point to execute transfer learning or domain adaptation. Different models can lead to different performance in different application domains and in some cases several trial-and-error attempts have to be executed to achieve satisfactory results.

In addition, a challenging research direction might be transferring knowledge inferred from a media, to a different one. For instance transferring knowledge acquired in image analysis to analyse different media, such as 3D models, or even text. Cross-modal solutions are being used to allow searching for images using text, or vice versa. Can we extend these methods to cross-modal transfer learning? For instance, is it possible to learn how to extract knowledge from images, once we know how to extract knowledge from text?

In real applications, we also observe a gap between the common practice used to set up benchmarks for research purposes and the real scenarios. For instance, in case of few-shot learning techniques, we can observe three main aspects, where the setup of benchmarking

[13] Image source: M. Di Benedetto, F. Carrara, E. Meloni, G. Amato, F. Falchi, C. Gennaro, Learning accurate personal protective equipment detection from virtual worlds, Multimedia Tools and Applications 80 (15), 23241-23253

problems (and thus the methods described in literature, as well as the existing implementations available from the research community) deviate from the practical requirements of using few-shot learning in media use cases:

- The typical setup of the problem is posed as *n*-way *k*-shot, i.e. a problem with *n* classes and *k* samples per shot. However, in practice the number of samples per class that are provided may differ.
- There is not a fixed predefined dataset, but the set for base classes will contain a mixture of third party and maybe own proprietary data for some classes, while the novel classes are mined from own or third party media content (e.g., web sites). Thus, the concept of a dataset is fluid and the available data will evolve over time.
- Classes need to be added incrementally, which requires creating balanced training sets, but approaches should aim to keep the training effort low. This again means that there is no fixed notion of a dataset, but the dataset needs to be updated on the fly via some sort of incremental learning.

Societal and media industry drivers

Vignette: Tagging and learning new classes for audio-casual content search

Valerie and Theodor are working in the archive of a large broadcaster. Valerie is responsible for supporting the journalists and editors in the newsroom of the broadcaster looking for content, whereas Theodor documents content arriving in the archive with automatically and manually generated metadata. Based on the incoming queries from the newsroom for material of a certain kind (e.g. showing a certain person / location / object / entity), Valerie does a search in the broadcaster’s archive for proper material and returns matching video clips to the newsroom editors. Due to the large size of the video archive (> 1 million hours), an exhaustive content-based search cannot be done in the provided timeslot. Therefore, usually the search in the archive is done by searching for certain (textual) metadata tags derived from the associated query. For a successful search, this means that it is crucial that new content which is added to the video collection is annotated properly (documented with the proper concepts). Of course, a concise manual annotation of a video is a very time-consuming process, therefore Theodor employs a semi-automatic workflow with the help of an AI tagging engine. The AI tagging engine employs automatic video analysis methods (object detection, face recognition etc.) internally and proposes a list of metadata tags based on the result of the video analysis. These proposed metadata tags are now inspected by Theodor and either accepted or discarded by him.

In February 2020, a new virus named SARS-CoV-2 has appeared suddenly and is spreading very fast throughout the world. Therefore, nearly every day the broadcaster is reporting in its news program about this new virus, which has the potential for a pandemic. Consequently, also Valerie is getting queries for content related to SARS-CoV-2 every day from the newsroom. Besides other related content (such as queries for street shots with people wearing face masks), the editors in the newsroom are also interested in video content which is showing visualisations of the SARS-CoV-2 virus with the characteristic spikes on its surface. As this kind of object is not existing yet in the AI tagging engine, it is not able yet to automatically detect and tag the object “SARS-CoV-2 virus” in a video. Which means more work for Theodor, as he has to tag the new object class manually. But the AI tagging engine provides a comfortable mechanism for incrementally adding new object classes, via few-shot learning. For this, all Theodor has to do is to provide a few (five to ten) images of the desired object class that he has annotated in recent days. After that, the new object class is incrementally trained in very short time (less than 15 minutes), and the updated AI tagging engine is now able to detect and automatically tag also the object class "SARS-CoV-2 virus shape" in video content. He will now tag a few of those face masks, to also support it in the object detector, and being able to discriminate between street scenes where people wear masks or not. As face masks are more challenging to identify visually, more samples are needed to increase the robustness, but those can be obtained from tracking faces throughout the video, and thus obtain a variety of poses of the masks.

Future trends for the media sector

In the media sector, technological solutions should constantly be aligned to the evolution of global affairs. Consider, for instance, news production: Media companies, in this case, should immediately intercept new emerging trends, new hot topics, and new relevant events. Accordingly, an AI system, analyzing the continuous flow of multimedia digital information constantly generated, cannot rely on pre-determined tools of analysis, but they have to be constantly updated and trained with the scarce and noisy data at hand.

As an intuitive example, imagine a face recognition tool that has to intercept pictures of a person who just recently became popular for some reason and who was unknown until yesterday. A pre-trained model of a face recogniser would not be able to recognise unknown people. Therefore, analysis tools (classifiers, face recognisers, etc.) should be immediately updated to be able to react quickly to this. Similar examples can be done, considering text, image, and video classifiers, that are supposed to identify occurrences of media related to new emerging topics or to detect fakes and disinformation in social media.

Effective solutions able to learn with scarce data allow media companies to immediately react to the occurrence of new emerging topics and be able to effectively detect media related to them, with minimum effort, minimum delay, and high accuracy.

Goals for next 10 or 20 years

In the future, we will need AI tools that are able to learn from the user with as little effort as possible. Nowadays, most of the effort needed to produce accurate AI tools relies in the preparation of a high-quality training set. In the future, solutions that learn from noisy and evolving data, in a trustable and reliable way, might represent the real goal. Solutions based on reinforcement learning and learning with scarce data go in this direction and can provide a significant step forward in AI.

Humans learn from their daily experiences and from interactions with other humans. Once we become experienced in a field, we require little effort and training to adapt to new related tasks. In addition, working in a team can further improve our skills.

We can imagine adopting a similar paradigm also in AI systems in the next decade. We can imagine an ecosystem of AI tools that interact and continuously improve their performance by autonomously learning from other tools and from available (possibly scarce) data, by constantly staying up to date with the most recent and emerging skills they need for their assigned tasks.