AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "AI AT THE EDGE"

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	**Symeon Papadopoulos** (Centre for Research and Technology Hellas – Information Technologies Institute) **Emmanouil Krasanakis** (Centre for Research and Technology Hellas – Information Technologies Institute)

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).

Copyright

AI at the edge

Current status

Technological improvements of the last decades have led to a widespread adoption of smart devices, such as mobile phones and sensors, in everyday life. For example, there are over 3 billion Android devices[1] that run software catering to a variety of needs, such as web surfing, creation and consumption of multimedia, social networking, and analysis of sensor data that range from weather readings to biometric ones. Many software products make use of AI breakthroughs to enhance user experience, for instance by recommending multimedia content or social interactions and automatically generating short description or tag summaries. These operations are often supported by central services, which are accessed through internet endpoints and store and process user data, for example for the purposes of retrieving those upon request or performing AI inference.

The above-described dependence on central endpoints makes software reliant on third-party infrastructure and services that make users hand over control of their private data. However, in our increasingly digital societies, the needs for data privacy, confidentiality and ownership, as well as for secure data exchanges with trustworthy parties are of paramount importance. To this end, an increasingly popular alternative to centralised data processing that addresses these concerns is to perform ***in-device data processing*** and employ ***privacy-aware communication schemes*** between devices that do not expose internal user data. Additional perks of this approach include robustness against downtime of centralised infrastructure (e.g., the 2021 Facebook outage had serious ramifications around the globe[2]) and the ability to deploy software and its accompanying AI to places with limited or restricted internet access (e.g., areas stricken by natural calamities, warzones, regimes where internet activity is monitored). Since devices lie at the "edge" of communication networks (e.g. of the Internet, but the same principles hold true in local networks without external connectivity) the paradigm of (partial or full) in-device data processing is referred to as ***edge computing*** (Figure 1).

[1] A. Kranz, There are over 3 billion active Android device (2021): https://www.theverge.com/2021/5/18/22440813/android-devices-active-number-smartphones-google-2021
[2] A. Asher-Schapiro and F. Teixeira, Facebook down: What the outage meant for the developing world (2021): https://news.trust.org/item/20211005204816-qzjft/

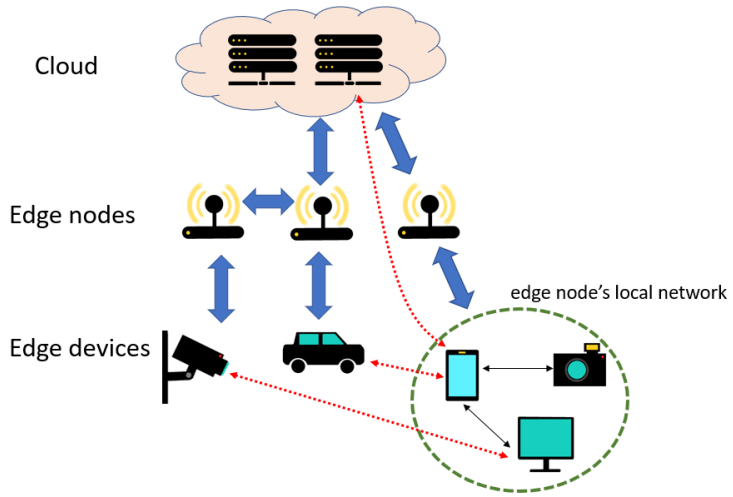

Figure 1: AI at the Edge basic concept. Cloud servers and edge devices can perform local computations or help each other learn through implicit communication paths (dashed red arrows) even if they are not on the same local networks.

Existing research to support AI on the edge can be roughly categorised into three directions, which differ in their degree of autonomy:

On-device inference. This aims to deploy AI models with pre-trained parameters to edge devices by replicating inference computations on data residing there. For example, this may take the form of image processing software that performs object recognition or automated tagging without external dependencies. Effectively, inference endpoints are made obsolete by bringing respective computations inside devices, where the latter make inferences autonomously but rely on central services to deploy the trained models. Related research aims to support the deployment of AI on device hardware with new compatibility frameworks (e.g., TensorFlow Lite for mobile GPUs[3]) and create models that fit device resources, for example by supporting low-end hardware or "compressing" the number of trained parameters to reduce memory and processing requirements.[4]

Distributed learning. This organises AI model training across several devices by making it independent of where computations are being performed. For instance, a popular paradigm of doing so is *federated learning*[5], which designates one device (e.g., a centralised service) as the trained AI model's host that orchestrates a learning process and lets other devices perform training operations (e.g., gradient calculations) in their own local data. Devices then share back parameter updates to be combined by the orchestrator and are sent copies of the updated model. Distributed learning can be considered as semi-autonomous, because it requires an initialisation process to organise the communication network, but devices run independently.

[3] TensorFlow Lite: https://www.tensorflow.org/lite
[4] Ogden, Samuel S., and Tian Guo. "{MODI}: Mobile deep inference made efficient by edge computing." {USENIX} Workshop on Hot Topics in Edge Computing (HotEdge 18). 2018.
[5] McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.

At the same time, it is popular for leveraging the high computational power of relay or cloud servers to learn at scales and speeds unimaginable by centralised computing. An important consideration many distributed systems already address is that devices performing computations may have gathered confidential data, such as medical records, in which case privacy-preserving protocols are employed to make it practically impossible to replicate source data at other devices or the model's host.

Decentralised learning. In this paradigm, fragments of AI models are trained on devices to approximate the outcome of centralised training. Devices do not follow predetermined communication topologies but create unstructured communication links, i.e. which devices communicate is not known at the algorithm design time but only once AI tools are deployed. Existing decentralised learning protocols either consider fixed high-throughput communication overlays (unknown at design time)[6,7] or require the ability to communicate between devices and randomly selected others[8,9], a design referred to as *gossip learning*. In both cases, communicating devices perform model fragment training based on local data and repeatedly average trained parameters between neighbors. Thanks to the conceptual simplicity of this practice, decentralised learning algorithms are often deployed in peer-to-peer communication networks to train model fragments that tightly approximate centralised learning.

Research challenges

AI at the edge provides a promising alternative to existing technological solutions that cope well with the increase in data scale and privacy concerns. However, there remain a lot of open questions over how to support it in real-world scenarios. Below, we outline promising directions that future research can address to support widespread adoption of AI at the edge beyond specialised environments (e.g., distributed learning of cloud or relay servers) to edge devices that see everyday use.

Accountability. Distributed and decentralised AI are trained across multiple devices. Thus, determining accountability is a pressing issue, as there is no single entity responsible for the outcome. Without accountability, even highly accurate AI is difficult to port to high-stakes settings, such as for example in automated medical diagnosis systems[10]. This task is made doubly challenging compared to centralised AI accountability, because different devices could make different conclusions for the same data. At the very least, it is important to dissuade "lazy" practices that lead to harmful (e.g. discriminatory) AI behaviour due to replicating and even accentuating real-world biases (see data heterogeneity challenges).

[6] Koloskova, Anastasia, Sebastian Stich, and Martin Jaggi. "Decentralized stochastic optimization and gossip algorithms with compressed communication." International Conference on Machine Learning. PMLR, 2019.
[7] Niwa, Kenta, et al. "Edge-consensus learning: Deep learning on P2P networks with nonhomogeneous data." Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020.
[8] Hegedűs, István, Gábor Danner, and Márk Jelasity. "Decentralized learning works: An empirical comparison of gossip learning and federated learning." Journal of Parallel and Distributed Computing 148 (2021): 109-124.
[9] Hu, Chenghao, Jingyan Jiang, and Zhi Wang. "Decentralized federated learning: A segmented gossip approach." arXiv preprint arXiv:1908.07782 (2019).
[10] Yetisen, Ali K., et al. "A smartphone algorithm with inter-phone repeatability for the analysis of colorimetric tests." Sensors and actuators B: chemical 196 (2014): 156-160.

***Bandwidth limits*.** Distributed and decentralised learning require continuous model gradient or parameter exchanges. At the same time, pre-trained model sizes grow proportionally to the number of their parameters. Hence, more complex machine learning models with billions of parameters may be impractical to deploy through traditional platforms or learning through non-centralised computing. This issue could be projected to the future too, if new computing technologies evolve before communication ones, as has been the trend so far. Preliminary works in this direction address bandwidth limits for gossip learning by performing information exchanges in smaller chunks at the cost of slower learning. Compressing model information during decentralised learning remains an open challenge that can help support more complex models.

***Data heterogeneity*.** Most distributed and decentralised learning approaches consider homogeneous distributions of data across edge devices (e.g., spreading data samples to devices without biases of which device gets which data). However, in practice, devices could differ in terms of the data they collect, for example due to placement of sensors on different physical locations or different preferences of mobile device users. Thus, research must take care to prevent imbalances in the types of local data from becoming biases of local AI model fragments. To make matters worse, these could also be difficult to detect with macro-evaluation (e.g., averaging) of model fragment results.. Gossip learning systemically addresses this challenge by making sure that random pairs of devices exchange parameters, but this comes at the steep cost of requiring constant device availability (accentuating the impact of dynamic behaviour challenges).

***Domain transfer*.** Transfer learning[11] is a widespread paradigm in which trained AI models are repurposed towards different predictive tasks by keeping large chunks of their parameters (e.g. most neural layers) constant and training only the rest. This often helps learn new high-quality models from limited data based on training on similar but larger datasets. For example, a popular practice is to transfer image feature extraction layers of state-of-the-art models to new tasks. This paradigm can be of particular interest for the deployment of pre-trained AI at the edge that can be used to adapt to problems encountered by the device's user. For example, transfer learning can be used to locally turn object recognition software into a recommender system that learns from a mobile phone user's stored image to locally refine image web search results, for instance by re-ordering them, without exposing their data to others.

***Dynamic behaviour*.** Current research on non-centralised AI either assumes fixed communication topologies of beneficial characteristics or the ability to randomly communicate with other nodes. However, communication links in the real world may be formed based on the belowmentioned concept of homophily, in which case topologies are fixed but are unlikely to exhibit the desired theoretical characteristics that lead to tight approximation of equivalent centralised model training. At the same time, communication links between devices can be unstable, for example due to users irregularly going online or offline or evolving social

[11] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2009): 1345-1359.

relations[12]. Thus, future research needs to address the evolving nature of communication networks (e.g., of peer-to-peer networks), especially those whose links change with rates comparable to learning convergence speed, and in which edge device availability is uncertain.

***Homophily*.** Homophily refers to the tendency of complex network nodes (e.g., social media users) to link with each other based on common attributes[13]. For example, social media friends often have similar hobbies. Thus, device communications (e.g., in peer-to-peer networks) can suffer from a relational type of bias. However, contrary to other risks, homophily, or other relational properties for that matter, can also be leveraged by graph-based AI tools (e.g., graph signal processing, graph neural networks) to make them more accurate. Overall, future research on AI on the edge needs to acknowledge potential homophilous communications and either use this property to improve predictions, or safeguard against potential biases. Notably, leveraging homophily can even support hybrid approaches, where it used by decentralised devices to improve pre-trained inference models

***Hyper-parameter selection*.** This is a concern for decentralised distributed learning and gossip learning approaches only, where there does not exist one central overseer to dictate model hyper-parameters (e.g., the number of neural layers, latent feature dimensions), for example, by comparing alternatives on a validation subset of data. In the case of distributed or gossip learning, hyper-parameters can be selected a-priori through experiments on similar datasets, but may not port well to new data once deployed in the wild. An elegant alternative would be for decentralised AI to also learn its hyper-parameters on-the-fly through additional decentralised processes. When doing this, it is important to create hyper-parameter selection protocols of low computational complexity that do not require untenable training times.

***Unequal device resources*.** The computing capabilities of devices on which AI is deployed may vary and even be unknown at design time. Thus, a promising trend is to create adaptive models that can make the best use of device resources, for example by providing many models for in-device inference. In case of gossip learning, making use of many devices to train fragments of models means that training is lightweight enough to be supported by even older devices. However, resource allocation may still be unequal in terms of available bandwidth.[14] Overall, research on resource usage needs to make sure that AI on the edge is not as weak as the lowest computing capabilities of devices expected to run computations.

Societal and media industry drivers

Vignette: Debunking fake news under an authoritarian regime using AI at the edge

Ann, Bob and Cale are journalists stationed inside the territory of an authoritarian regime. The regime closely monitors internet activity and keeps producing disinformation content in order to spread propaganda in its populace. All three journalists come across fake media on a daily

[12] Berta, Árpád, Vilmos Bilicki, and Márk Jelasity. "Defining and understanding smartphone churn over the internet: a measurement study." 14-th IEEE International Conference on Peer-to-Peer Computing. IEEE, 2014.
[13] McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. "Birds of a feather: Homophily in social networks." Annual review of sociology 27.1 (2001): 415-444.
[14] Musaddiq, Arslan, et al. "Reinforcement learning-enabled cross-layer optimization for low-power and lossy networks under heterogeneous traffic patterns." Sensors 20.15 (2020): 4158.

basis, although these are only a small portion of the regime's continuous efforts at disinformation, i.e., there are more pieces of fake media they do not come across. The journalists aspire to support AI tools that, based on their annotations, would learn to identify more pieces of disinformation to warn the populace. However, due to ongoing monitoring of internet activity, doing so runs the risk of them being caught.

Luckily, all three journalists are users of a decentralised fake media detection platform. In this, the users annotate (i.e. tag) media examples as fake or not and these annotations are fed to decentralised learning algorithms to learn to distinguish fake media similar to the annotations. The platform runs on a peer-to-peer network (e.g. that already circumvents part of monitoring by encrypting its communication) that maintains privacy by fully obfuscating how users contribute to the fake media detection algorithm. Furthermore, the platform is designed to send user data only to trusted others.

Thus, the journalists can feel safe in providing high-quality annotations, which decentralised algorithms will then collectively process so that the devices of all platform users would hold fragments of predictive models that learn to distinguish whether viewed media content is fake or not.

Future trends for the media sector

Edge computing can be a game changer in the way the media sector deploys AI models to enrich media content with metadata and develop new user experiences. We highlight some of these opportunities with an eye to AI4Media use cases:

- **Create collectively trained AI models** that process and learn from continuously generated real-world data. For instance, these models could interact with the users to obtain feedback on classification or recommendation goals; if only a portion of users are willing to manually annotate data meant for their own consumption, then all devices can make use of this information. We expect the increased privacy of AI to encourage users to engage in this way, perhaps through software design opportunities (e.g., like and dislike buttons in mobile applications) that would not be possible for fully centralised systems.
- **Reduce development and upkeep costs of AI tools** by making use of the combined computing power of their users' devices to run calculations, i.e., media companies can avoid inference costs and -in the cases of distributed or decentralised learning- model maintenance costs (e.g., training with new data and re-deploying). This also means that AI training can become more environmentally-friendly, as the already running resources of edge devices are used.
- **Ensure data privacy and ownership** by not allowing data to leave user devices. This can help significantly promote trust of content users.
- **Create highly personalised media applications** that take into account many aspects of user lives that would be difficult to gather and get back with centralised architectures.
- **Personalise AI** based on local user feedback.
- **Help fight disinformation** by creating collectively managed environments that make use of AI without the tampering of overseers.

- **Perform ongoing training** that quickly adapts to changes in real-world data in safety-critical systems. This is particularly useful against adversarial attacks that aim to fool AI tools by circumventing current learned models. For example, decentralised learning could provide protection against evolving disinformation and deepfake techniques by leveraging new content flagged by users. Users could flag new types of fake content and decentralised AI could immediately integrate this information in model training to also flag similar content. This would be achieved without waiting for software updates (bearing new versions of models) that take too long to deploy and could help stop disinformation attacks long before they reach a critical mass of users to become popular.
- **Allow smaller media organisations to compete** on the AI front without requiring expensive machinery or extensive data collection processes.

Goals for next 10 or 20 years

In the next 10 or 20 years, AI at edge will be able to make use of most data generated by edge devices to capture multifaceted aspects of people's lives without violating their privacy. Thus, highly personalised and well-scaling AI will be able to enrich a new generation of human-machine interactions, where trainable models (including those used in media applications) learn from the collective experience of their target audiences and support learning tasks that are not feasible by existing centralised computing.

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