AI4media
ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

www.ai4media.eu

@ai4mediaproject

AI4Media Technological Highlights

Discover AI4Media's Key Research Outcomes on Trustworthy AI

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

Introduction

This booklet showcases the highlights of four years' work towards research and development in Trustworthy AI.

More information on these and other research outcomes of WP4 "Explainability, Robustness and Privacy in AI" can be found in relevant public deliverables available here.

The relevant work package developed new techniques and algorithms for the application of trustworthy AI for media industry use cases, across four key pillars of interest: (i) adversarial robustness (covering novel attacks and defences for evasion and data poisoning, and certifying and verifying the robustness of AI models), (ii) explainability and interpretability (including assessing the explainability of multi-modal deep-fake detection classifiers), (iii) privacy and security (where privacy and security is preserved on the training data of AI models), and, (iv) fairness (concerning the detection and mitigation of bias in AI models across a variety of data modes and types of AI models).
Also included in this booklet is an outline of a novel AI benchmarking tool that was built as part of AI4Media.

This booklet gives a summary of the quantity and quality of work undertaken by AI4Media partners and collaborators, and the lasting impact it will have on trustworthiness for media use cases in the years to come.

#1

Mitigating Robust Overfitting via Self-Residual-Calibration Regularisation

Partner organisations involved

UNITN

→ **Paper**

A few words about this technology

Our technology addresses the important problem of overfitting in adversarial training and our analysis reveals that when evaluating the defence performance of several calibration methods 1) a well-calibrated robust model is decreasing the confidence of a robust model; 2) there is a trade-off between the confidences of natural and adversarial images. These issues bring us to designing an effective regularisation, called Self-Residual-Calibration (SRC) that can effectively mitigate the overfitting problem while improving the robustness of state-of-the-art models. Importantly, SRC is complementary to various regularisation methods. When combined with them, we are capable of achieving the top-rank performance on the AutoAttack benchmark leaderboard.

Who can benefit

Deep neural networks (DNNs) are very susceptible to adversarial examples, which have been demonstrated to be threatening in various domains such as computer vision, natural language processing, and speech recognition. These specific examples may be generated using various adversarial attack methods, causing the DNNs to behave incorrectly. Enhancing the adversarial robustness of DNNs could be an essential task for all the researchers in the community.

Impact and added value for the media industry

The results of our analysis are particularly interesting to the media industry given that they need to deal with several modalities and as such achieving adversarial robustness is paramount. In our research we mainly refer to the model's robustness to the perturbations of input data, i.e., a model is robust when it can defend against most kinds of adversarial attackers. We show that by injecting our SRC into an adversarial training algorithm, we can effectively avoid robust overfitting and thus enhance robustness/generalisation. Extensive experiments verify that the proposed method can help better defend cutting-edge adversarial attack methods.

Future developments on the technology

Recent studies show that multimodal models like CLIP can substantially improve the robustness of a model.
This advantage is mainly benefited from the large-scale, diverse, multimodal data. In our future development we will focus on how to handle robust overfitting and enhance adversarial accuracy against adversarial attacks with the help of multimodal models.

#2

Matching Pairs: Attributing Fine-Tuned Models to their Pre-Trained Large Language Models

Partner organisations involved
IBM

→ **Code**

A few words about this technology

Large Language Models (LLMs) or more generally Foundation Models are an emerging technology of general-purpose AI models trained on large volumes of data which can be fine-tuned for a wide range of downstream tasks. LLMs, in particular, can generate novel high-quality text and help drive many downstream applications including machine translation, question answering, and text summarization systems. However, training these models is challenging as it requires access to vast amounts of data (text corpus) and computation. This has led to a market where developers, who often don't have access to such resources, source LLMs from third-parties (which we refer to as base models) and fine-tune them for specific domains/tasks. There are growing threats like violation of model licences, model theft, and copyright infringement. Moreover, recent advances have shown that generative AI is also capable of producing harmful content which only exacerbates the problems of accountability within ML supply chains. As approaches like watermarking are shown to be easily bypassed, there's a need for developing general purpose solutions to help with forensics.

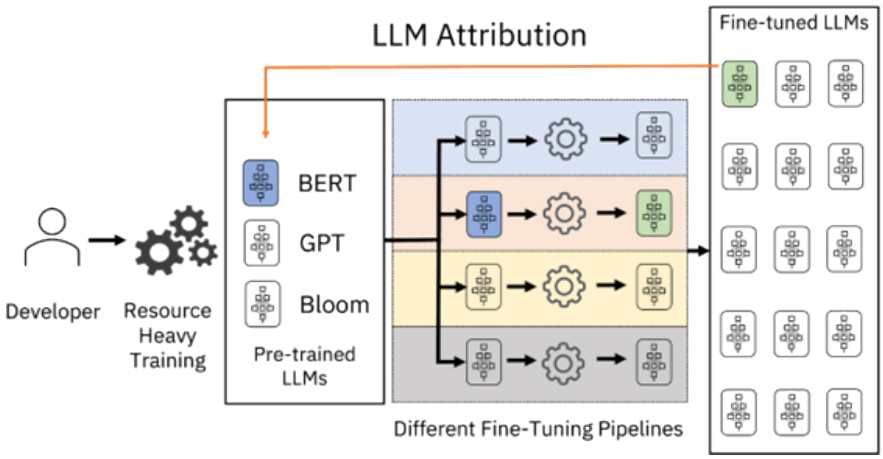
This work formalises the role of attribution within the supply chain and presents heuristic and ML based approaches for attribution under different knowledge levels. Furthermore, it shows how such methods can be made more efficient for constrained settings where an attributor may have limited access to the model and/or its API.

Who can benefit

LLM attribution aims to link an arbitrary fine-tuned LLM to its pre-trained base model using information such as generated responses from the models. Through LLM attribution, regulatory bodies can trace instances of intellectual property theft or influence campaigns back to the base model. Attribution is an important step in making the AI model supply chains more robust

Impact and added value for the media industry

Rapid adoption of machine learning across all industries has raised challenges on model ownership and traceability. For instance, one is likely to face issues on model theft or copyright infringement. This is particularly relevant for sectors like the media industry where generative models are being used to create content. Given access to an API for generating content, how can one trust the source of the API? In this work, we provided a method to establish this relationship and make ML supply chains more robust.



#3

SemAug: Semantic Generative Augmentations for Few-shot Counting

Partner organisations involved

CEA

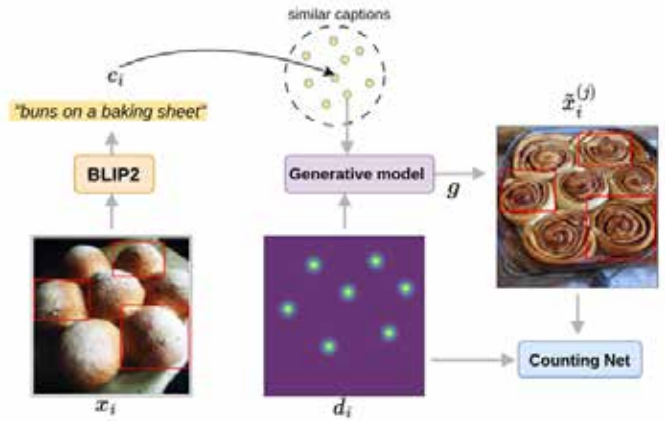
→ Code

A few words about this technology

With the availability of powerful text-to-image diffusion models, recent works have explored the use of synthetic data to improve image classification performances. These works show that it can effectively augment or even replace real data. In this work, we investigate how synthetic data can benefit few-shot class-agnostic counting. This requires generating images that correspond to a given input number of objects. However, text-to-image models struggle to grasp the notion of count. We propose to rely on a double conditioning of Stable Diffusion with both a prompt and a density map in order to augment a training dataset for few-shot counting. Due to the small dataset size, the fine-tuned model tends to generate images close to the training images. We propose to enhance the diversity of synthesised images by exchanging captions between images thus creating unseen configurations of object types and spatial layout.
By relying on explicit semantic content to control the generation of the extended training datasets, it provides a better interpretation of the visual content that is used to learn discriminative models.

Who can benefit

The developers of AI tools for the media industry can benefit from this asset to create synthetic data that will enrich their training datasets. The current version of thetool has been tested in the context of few-shot class-agnostic counting, that is the ability to count some object of any type in an image, by showing to the tool only few (e.g 3) examples.



Impact and added value for the media industry

The adoption of AI tools in the media industry raised several challenges, regarding their performance but also their reliability and to which extent the tools can be trusted, that is their trustworthiness. The current tool can contribute to address these challenge by several ways (1) by augmenting the training datasets of AI tools, it will contribute to improve their performance (2) since the augmentation is made with synthetic data which generation is controlled, to some extent, with a human instruction, the resulting dataset can be enriched in order to reduce their potential biases toward particular classes of individuals, while better understanding why these biases are reduced.

Future developments on the technology

The technology aims at being developed and tested on other types of task, such as object recognition, detection and segmentation.

#4

Disentangling Neuron Representations with Concept Vectors

Partner organisations involved
HES-SO

→ Code

A few words about this technology

Mechanistic interpretability is a pioneering research field that aims to clarify how neural networks store and process information. Traditional techniques to understanding neural networks frequently rely on investigating individual neurons. However, the discovery of polysemantic neurons, which respond to several, unrelated aspects, has complicated the process.
To overcome this, researchers have redirected their attention from individual neurons to concept vectors in activation space. Concept vectors incorporate different features, making them a more detailed and interpretable unit of analysis in neural networks.
This approach disentangles polysemantic neurons into concept vectors, allowing for a more granular and understandable representation of the features learned by the model. By examining these vectors, researchers can gain insight about how models store complicated notions, thereby boosting our capacity to anticipate and alter models behaviour.

Who can benefit

Mechanistic interpretability provides major benefits to AI researchers and developers by allowing them to better comprehend and refine neural network models. It is especially useful in domains that need high interpretability, such as healthcare, finance, and autonomous systems, where understanding the decision-making process is essential. Furthermore, educators and students of AI and machine learning can use these insights to improve their learning and contribute to the creation of more transparent models.

Impact and added value for the media industry

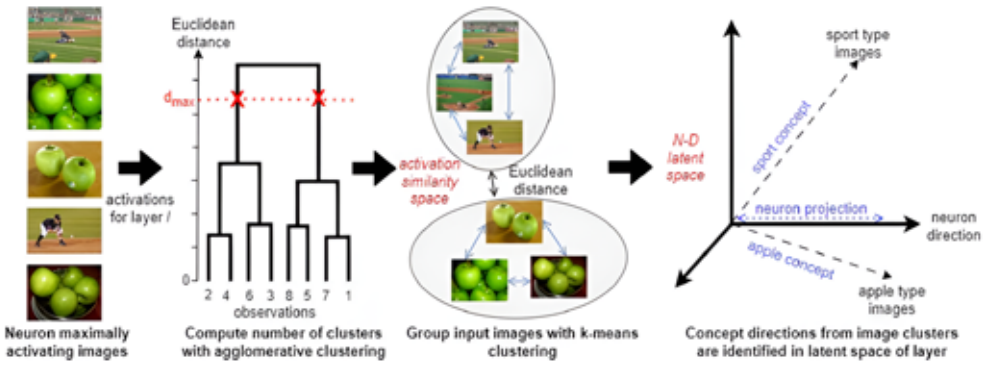
The media sector stands to benefit significantly from advances in mechanistic interpretability. Media companies rely significantly on AI models for a variety of functions, including content recommendation, tailored advertising, and automated content development.
By focusing on specific algorithmic processes such as global average pooling, Euclidean distance measurements, hierarchical clustering, and k-means clustering, the study offers a straightforward way for converting polysemantic neurons into coherent, monosemantic concept vectors. These vectors make media sector applications more efficient and transparent, bringing significant benefit to both service providers and users:

- **Improved Feature Extraction for Content Recommendations:** AI-powered recommendation systems are critical for customising user experiences on media platforms. Media companies might employ mechanistic interpretability to study and enhance these systems, ensuring that their recommendation algorithms are not only accurate but also transparent and fair.
- **Personalised Advertising:** it can help to develop more successful and ethical advertising techniques. Understanding the elements that increase user engagement allows advertisers to create ads that are more relevant and less annoying. Furthermore, it may assist in discovering and reducing biases in the targeting of ads, ensuring a more equal distribution of advertisements.
- **Automated Content Generation:** we are witnessing a fast growing exploitation of generative models, including the ones creating contents such as news and articles. Understanding how these models work can assist content creators ensure the quality and coherence of the generated content. It can also help fine-tune these models to match the editorial standards and values of the media business.

- **AI Safety and Ethics:** given the media industry deploying more and more AI-driven systems, it is critical that they operate safely and ethically. Mechanistic interpretability can help uncover potential dangers and biases in AI models. By offering a greater understanding of how models make decisions, media organisations could establish better safeguards and accountability.
- **User Trust and Transparency:** data privacy and algorithmic transparency are becoming increasingly important, media businesses can employ mechanistic interpretability to boost user trust. Media corporations may improve their reputation and user loyalty by being open about how AI systems work and ensuring that they operate properly.

Future developments on the technology

The future of mechanistic interpretability appears bright, with prospective breakthroughs including the application of similar methodologies to other forms of data, such as text and tabular data. Researchers are also likely to investigate novel ways for identifying concept vectors beyond neurons, which could improve the granularity and accuracy of model interpretations. Furthermore, as AI models get more sophisticated, creating scalable and efficient interpretability approaches will be critical. These breakthroughs will not only improve our understanding of AI models, but will also ensure their safe and ethical application in a variety of industries.



#5

FLCrypt – Secure Federated Learning

Partner organisations involved

FhG-IDMT

→ **Request access to the library here**

A few words about this technology

Federated Learning brings a lot of benefits: it is no longer needed to bring potentially critical data to a central server for training (which might be undesirable or plainly not allowed by law) and instead trains the model in a decentralized way. Only the trained models, not the raw training data will be distributed.

Who can benefit

Everyone with an elevated concern about data privacy can benefit: FICrypt adds another layer of security on top of Federated Learning and might lower the barrier of entrance to deploy custom Federated Learning systems by offering additional protection to sensitive training data.

Impact and added value for the media industry

Simply put, everyone in the media industry with problem cases that can be solved by federated learning might profit from FICrypt.

Natural use cases involve handling a lot of potential sensitive customer data, e.g. in the deployment of recommender systems in a federated broadcast association, where the regional parts want to corporate.

Future developments on the technology

Future developments will concentrate on evaluating applicable FHE schemes and their performance implications. In addition, Differential Private Machine Learning will be added to the toolkit.

#6

Differentially Private Graph Learning

Partner organisations involved
IDIAP

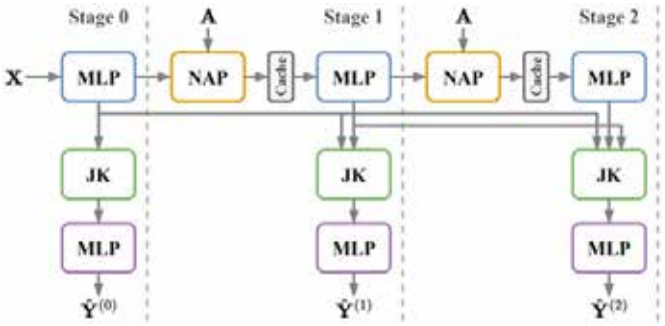
→ **Code**

A few words about this technology

This technology aims to prevent the information leakage of the underlying graph in Graph Neural Networks (GNNs) using Differential Privacy (DP), which is a widely accepted mathematical framework for measuring the privacy guarantees of algorithms that operate on sensitive data. A first version of the technology proposes a novel differentially private GNN based on Aggregation Perturbation (GAP), which adds stochastic noise to statistically obfuscate the presence of a single edge (edge-level privacy) or a single node and all its adjacent edges (node-level privacy). A second version of the technology proposes the application of progressive learning to privacy-enhancing GNNs, showing that it can be used to improve the accuracy-privacy trade-off of DP GNN models without sacrificing privacy.
This approach maintains the representational power of GNNs while limiting the incurred privacy costs.

Who can benefit

GNNs have shown superior performance in solving the problems formulated as a machine learning task over graphs, such as node classification, link prediction, and graph classification, in various media-related areas including social network analysis and recommendation services. The method addresses the case when the graphs used to train such models contain sensitive or personal information, and this information can be leaked through the model's output, when the GNN is released publicly, or when it is offered as a service. For example, a GNN trained over a social network for friendship recommendation may reveal the graph's linkage information through its predictions.



Impact and added value for the media industry

The GAP model, with its focus on differential privacy in Graph Neural Networks, offers a novel approach to various media industry use-cases. In content personalization, GAP could help media platforms deliver tailored content that respects user privacy, thereby enhancing user engagement and trust. For journalists and researchers, GAP's capabilities in social media analytics could provide valuable insights into trends and public opinion while adhering to ethical privacy norms. Lastly, in customer segmentation, media companies could use GAP to better understand their audience and deliver targeted services, all while maintaining strict data privacy standards. Overall, GAP presents a scenario that balances the need for advanced analytics with the imperative of user privacy in the media industry.

#7

Knowledge-driven Active Learning

Partner organisations involved

3IA-UCA

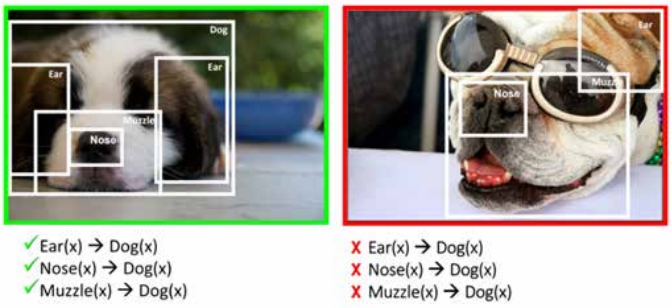
→ **Code**

A few words about this technology

The deployment of Deep Learning (DL) models is still precluded in those contexts where the amount of supervised data is limited. To answer this issue, active learning strategies aim at minimising the amount of labelled data required to train a DL model. Most active strategies are based on uncertain sample selection, and even often restricted to samples lying close to the decision boundary. These techniques are theoretically sound, but an understanding of the selected samples based on their content is not straightforward, further driving non-experts to consider DL as a black-box.
For the first time, here we propose to take into consideration "common sense" domain-knowledge and enable non-experts in AI to train a model with easy to define and easy to explain (for a domain-expert) newly selected samples. In our Knowledge-driven Active Learning (KAL) framework, rule-based knowledge (easy to explain and provided by the domain-expert) is converted into logic constraints and their violation is checked as a natural guide for sample selection.
We show that even simple relationships among data and output classes offer a way to spot predictions for which the model needs supervision. To the best of our knowledge, our approach is the first explainable active learning strategy for deep networks. We empirically show that KAL (i) outperforms many active learning strategies, particularly in those contexts where domain knowledge is rich, (ii) it discovers data distribution lying far from the initial training data, (iii) it ensures domain experts that the provided knowledge is acquired by the model, (iv) it is suitable for regression and object recognition tasks unlike uncertainty-based strategies, and (v) its computational demand is low.

Who can benefit

Any person who wants to build a retrieval model exploiting a specific knowledge, can benefit from our active learning strategy, either to train iteratively from scratch a new model or to fine-tune a pretrained one. The new unlabelled samples to be annotated by a domain expert, then to be added to the training set for the next learning iteration, are selected thanks to easy to define and easy to explain logic rules based on domain



With our method, between the two unlabelled images above, the right one will be chosen to be added to the training set and retrain the model. Indeed, even though Ear, Nose, and Muzzle are detected, our model in its current state does not predict Dog for the image. The contradictions between detected parts and the prior knowledge on what is a dog, make this image interesting to train on.

expertise. For instance, a journalist who is looking for specific pictures illustrating a given event, can easily select unlabelled pictures, wrongly retrieved by the current model as relevant, while clearly displaying irrelevant anachronistic events, or persons, or objects, or inappropriate location, etc. Then, it is easy to define a logic rule: if classified as relevant while containing this anachronistic event/person/object, then annotate as non-relevant and add it to the training set. This should be easier for a journalist than to "understand" why, on the basis of pixels, one image would be more useful than another for improving the model as much as possible in the next learning iteration.

Impact and added value for the media industry

It is easy to exploit existing knowledge in order to classify, to retrieve, unlabelled samples, label them, add them to the training set, retrain, and benefit from a better model.
Any decision model for any kind of multimodal data (image, audio, text, video, etc) can be improved with our knowledge-driven active learning.

Future developments on the technology

Extending these results to more complex input data such as graphs, and more complex relations.

#8

Exploring Fairness of an AI-based Deepfake Detection Service

Partner organisations involved
IBM, CERTH

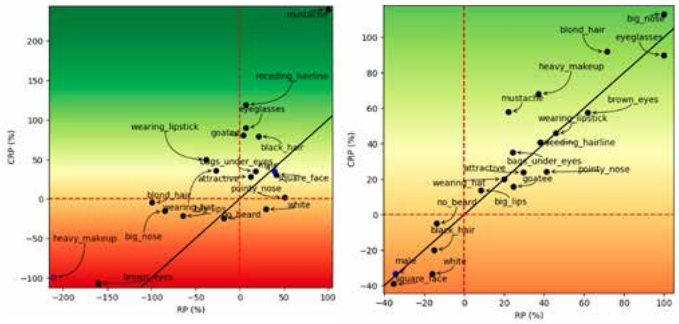
→ **Video**

A few words about this technology

Deepfake detectors attempt to identify fake images and videos generated by deep learning methods and are essential for mitigating the spread of disinformation. But how can we be sure these detectors are fair to all? This work addressed a gap in evaluations of deepfake detectors by assessing fairness with respect to a variety of protected and non-protected attributes found in two standard deepfake datasets. Using IBM's open-source AIF360 toolkit, a number of biases were identified and interpreted using state-of-the-art fairness metrics. Additionally, as deep fake detectors are targets for adversarial attack, this work also investigated Robustness Bias, identifying which protected groups were more or less susceptible to successful adversarial attacks, in both defenceless and defended deepfake detectors. The attacks were conducted using IBM's open-source Adversarial Robustness Toolbox (ART), and simulated scenarios in which malicious actors had full access to the deep fake detector (a white-box attack, which is less likely) and scenarios where malicious actors had only API access (black-box attack, which is more likely for deployed detectors).

Who can benefit

The users of deepfake detectors stand to directly benefit from this work as it provides an improved understanding of detector performance, their limitations and scenarios in which they should not be solely relied upon. This mainly encompasses those working in the media domain, such as journalists, who need to validate the authenticity of images and videos circulated to them. Through improved evaluations of deepfake detectors, such as our evaluation regarding bias and adversarial attacks, these users can be made more aware of the groups that are more susceptible to failures in the detectors they are working with, and implement mechanisms to ensure no group is unfairly treated. Subsequently, the general population also benefit from this work, as it reduces the likelihood of unfair treatment at the hands of media outlets, or other institutions deploying deepfake detectors e.g. social media platforms (when uploading content), banks (when applying for loans), recruiters (when applying for jobs) and so on.



Impact and added value for the media industry

The fairness evaluation of the MeVer Deep fake Detection Service is relevant to UC1 as it tackles disinformation detection and, specifically in this evaluation, the performance of deepfake detection across different groups. The results of such evaluations which study how such a deep fake detector can be advantageous or disadvantageous to certain groups of people over others also makes this work relevant to UC4 "AI for Social Sciences and Humanities" as the impact of deploying such an AI system "in-the-wild" without concern for fair treatment of subjects, may introduce biases and scenarios in which certain people are treated unfairly or are discriminated against. As a result, this work is also relevant to UC2 "AI for News", as a tool which can help journalists discern authentic content from deep fakes, must also ideally be fair and unbiased, which this work strives to achieve.

Future developments on the technology

Future research developments in this space include identifying improved bias mitigation techniques for deep fake detectors which enhance fair treatment across all groups without reducing their performance detecting deepfake material as well as identifying techniques for defending against adversarial attack which facilitate the preservation/enhancement of fairness metrics when attempting to remove the adversarial perturbations added by malicious actors. Future work should also build toward scalable frameworks which can evaluate the intersection of bias and robustness on a wider range of deepfake detectors in an explainable and easily shareable form.

#9

AI4Media Benchmarking Platform

Partner organisations involved

UPB

→ **Platform**

A few words about this technology

The AI4Media benchmarking platform provides support for benchmarking organisers and AI developers by helping in the creation, maintenance and management of benchmarking competitions. Thus, this platform can be integrated in organisations that wish to ensure ethical considerations like data protection, fairness and robustness in performance assessment, and reproducibility for AI models. The platform is deployed as an online Evaluation-as-a-Service (EaaS) tool, providing functionality for creating and deploying benchmarking competitions, offering common data, data splits, metrics for evaluating the performance of AI models, concept definitions, as well as a complete history and logging of the submitted runs, automatic scripts for detecting errors at submission time, and integration with 3rd party cloud providers for hosting and running competition-related scripts and services, as well as distributed computing for running the AI models in a reproducible environment. Furthermore, we wish to encourage the development of environmentally friendly AI models and provide methods of integrating metrics that can measure the computational complexity of the submitted AI models.

Who can benefit

The AI4Media benchmarking platform is geared towards integration in research organisations and companies that are interested in benchmarking the performance of AI models for a series of tasks in a fair and reproducible way.
These entities can either organise their benchmarking competition for internal use (choosing an optimal AI model for deployment in a product or for internal use) or as a publicly available platform, in open benchmarking tasks, available to the worldwide research community.

Impact and added value for the media industry

We summarise the impact of the platform with three key points and differentiators, as follows: follows: (i) providing a comprehensive and easy-to-implement API collection that can aid competition organisers in deploying computational complexity-related metrics, while also providing an implemented time complexity metric that organisers can either use as-is or use as an implementation example for their own metrics; (ii) using both API integration and containerization-based integration for submitting participant methods, thus offering several options for competition organisers to check and implement reproducibility for the proposed AI models; (iii) offering an important EU-based benchmarking platform, that can be focused towards common AI goals for the European Community.

Future developments on the technology

Future developments for this platform will be geared towards expanding the key differentiators and advantages that this platform brings to the AI benchmarking landscape.
Concretely, we will look into expanding the possibilities of integrating GPU processing power into the platform either via more 3rd party cloud infrastructure options, or via integrating the infrastructure that the implementing organisation has available, as well as further development on the computational complexity metrics.

#10

Integrated Framework for Multi-granular Explanation of Video Summarization

Partner organisations involved
CERTH

→ **Paper**

A few words about this technology

Our framework integrates methods for producing explanations about the suggestions of a video summarization technology, both at the fragment level (indicating which video fragments influenced the decisions of the summariser the most) and the more fine-grained visual object level (highlighting which visual objects were the most influential for the summarizer). It was built by extending our previous work on this field to investigate the use of a model-agnostic, perturbation-based approach for fragment-level explanation of the video summarization results, and introducing a new method that combines the results of video panoptic segmentation with an adaptation of a perturbation-based explanation approach to produce object-level explanations. Its performance has been evaluated using a state-of-the-art summarization method (CA-SUM) and two benchmarking datasets (SumMe, TVSum). The findings of the conducted evaluations demonstrated the ability of our framework to spot the most and least influential fragments and visual objects of the video for the summarizer, and to provide a comprehensive set of visual-based explanations about the output of the summarization method.

Who can benefit

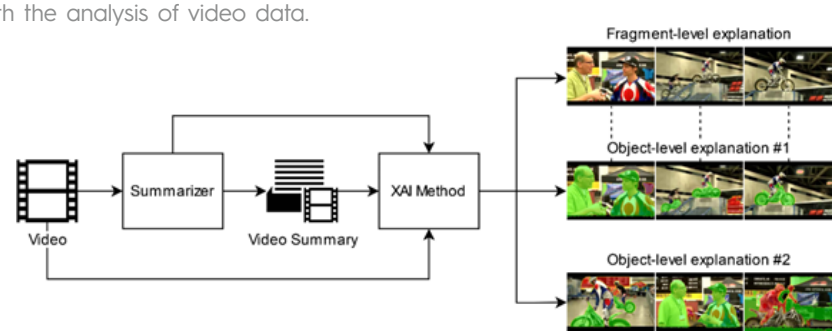
The provision of explanations about the suggestions of a summarization technology would benefit professional video editors in the Media industry, as it allows a level of understanding about the functionality of this technology, thus increasing the editors' trust in it and facilitating content curation. Moreover, our framework can benefit researchers working on the fields of video summarization and explainable AI, as it contributes towards obtaining a better understanding of the working mechanism and the output of network architectures dealing with the analysis of video data.

Impact and added value for the media industry

The current practice in the Media industry for producing a video summary requires a professional video editor to watch the entire content and decide about the parts of it that should be included in the summary. This is a laborious task and can be very time-consuming in the case of long videos or when different summaries of the same video should be prepared for distribution via multiple video sharing platforms (e.g., YouTube, Vimeo, TikTok) and social networks (e.g., Facebook, Twitter, Instagram) with different specifications about the optimal or maximum video duration. The use of video summarization technologies by Media organisations can drastically reduce the needed resources for video summarization in terms of both time and human effort, and facilitate indexing, browsing, retrieval and promotion of their media assets. Nevertheless, the outcomes of these technologies still need to be curated by a video editor, to make sure that all the necessary parts of the video were included in the summary. This content production step can be further facilitated if the video editor is provided with explanations about the suggestions made by the used video summarization technology.

Future developments on the technology

In terms of future developments, we plan to evaluate the performance of our framework using additional state-of-the-art methods for video summarization. Moreover, we aim to leverage advanced vision-language models (e.g., CLIP and BLIP-2) and extend our framework to provide a textual description of the produced visual-based explanations, thus making it more user-friendly for media professionals.



AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

Our Consortium

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