AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

www.ai4media.eu

@ai4mediaproject

AI4Media Technological Highlights

Discover AI4Media's key research outcomes on Content-centered 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 illustrates AI4Media contributions in AI for multimedia content production and analysis, pushing the boundaries of deep learning and media applications.

More information on these and other research outcomes of WP5 "Content-centered AI" can be found in relevant public deliverables available here.

Overall, this booklet underscores AI4Media's commitment to advancing AI technologies that significantly impact the media industry, fostering innovation and improving the way media content is created, managed, and consumed.

For each contribution, the booklet specifies the target user groups, outlines the impact and added value to the media industry, and suggests potential future directions.

The views and opinions set out in this document 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.

#1

SR-UNet: A Network for Fast Video Quality and Resolution Improvement

Partner organisations involved

UNIFI

→ Code

A few words about this technology

SR-UNet is a novel technique for super resolution and compression artefact removal in videos.

This technology can be used in a streaming context where the content is generated live, e.g. in video calls, and how it can be optimised when video to be streamed are prepared in advance.

The network can be used as a final post processing, to optimise the visual appearance of a frame before showing it to the end-user in a video player. Thus, it can be applied without any change to existing video coding and transmission pipelines.

Who can benefit

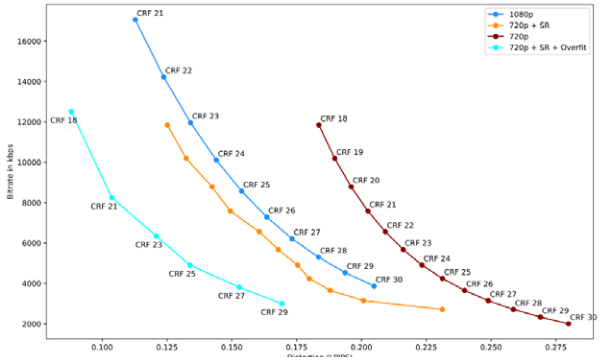
Technical personnel in industry and academia requiring fast enhancement of video streams in terms of super resolution and compression artefact mitigation.

Impact and added value for the media industry

SR-UNet can be used to perform super resolution and compression artefact removal in videos. The network can be used to improve the visual quality of videos compressed with H.265 codec, or to reduce the required bandwidth while maintaining a specified visual quality. The effectiveness has been shown using both subjective and objective metrics, considering both signal-based scores (like VMAF) or perceptual ones like LPIPS.

Super-Resolution and Artefact Removal

Super-resolution refers to the process of increasing the resolution of an image or video beyond its original quality. It involves reconstructing high-resolution images from low-resolution inputs, effectively enhancing detail, clarity, and overall visual quality. Artefact removal, on the other hand, involves eliminating unwanted distortions or imperfections that may arise during image compression, transmission, or editing processes. Together, these technologies can dramatically improve the visual fidelity of media content.



#2

SMEMO: Social MEmory MOdule for trajectory forecasting

Partner organisations involved
UNIFI

→ Code

A few words about this technology

Social Memory Trajectory Forecasting allows to forecast and model pedestrian paths, can be used for automated cinematography to plan drone motion.

Effective modelling of human interactions is of utmost importance when forecasting behaviours such as future trajectories. Each individual, with its motion, influences surrounding agents since everyone obeys to social non-written rules such as collision avoidance or group following.

Social Memory Trajectory Forecasting is a new technique that allows modelling such interactions, which constantly evolve through time, by looking at the problem from an algorithmic point of view, i.e., as a data manipulation task. It involves a neural network based on an end-to-end trainable working memory, which acts as an external storage where information about each agent can be continuously written, updated and recalled. This method is capable of learning explainable cause-effect relationships between motions of different agents, obtaining state-of-the-art results on multiple trajectory forecasting datasets.

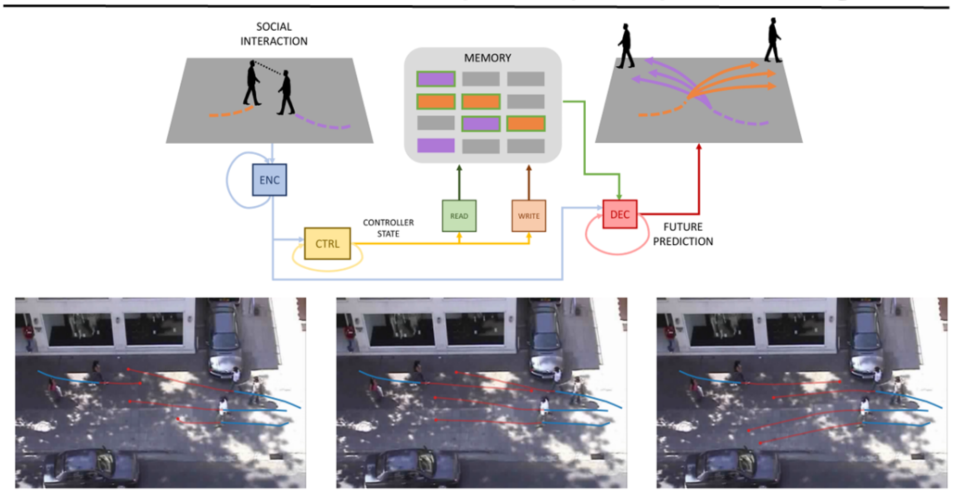
Who can benefit

People working in automated cinematography methodologies require precise pedestrian trajectory forecasting in scenarios for which it is important to take into account social rules.

Impact and added value for the media industry

Trajectory forecasting is a crucial component in automated cinematography, significantly enhancing the ability to autonomously capture dynamic scenes with optimal framing and composition. This technique involves predicting the future positions of moving subjects, such as actors or vehicles, based on their current trajectories and behavioural patterns. By integrating trajectory forecasting into automated cinematography systems, filmmakers and videographers can achieve smoother, more professional-looking shots with minimal manual intervention.

SMEMO: Social Memory for Trajectory Forecasting



Social Multi-modal Prediction

#3

Expressive piano performance rendering

Partner organisations involved

IRCAM

→ Method

A few words about this technology

This model deals with the expressivity of piano performances. Synthesising digital music from note sequences of raw scores, the rendering may sound mechanical. Based on unaligned examples of raw scores and human performances, this model reproduces the characteristics that makes a piano performance more human and expressive. It acts on the tempo changes, the note positions in time, and the nuances.

Who can benefit

Musicians, and more specifically composers, are the first beneficiaries of this module. Currently, when composers edit a score in a dedicated software, the rendering plays the piece in a "mechanical" way, with perfectly aligned notes in time, a fixed tempo, and without nuance. This module makes expressive rendering possible as a human does, by inserting changes that are not written.

Additionally, this module is a preliminary step for other deep learning tasks related to the performance modelling. Hence this work is also intended for researchers and developers to derive other tasks, such as: performance analysis.

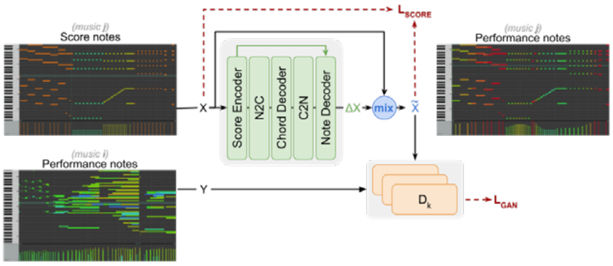
Impact and added value for the media industry

Currently, the performance modelling makes possible the rendering of expressivity in order to play music scores as human musicians do. This is an interesting application for composers who want to listen to their music during the edition, and before musicians play it.
Moreover, such modelling will make possible analyses of music performances, for example to characterise playing styles of pianists. This may have a beneficial impact on pedagogy for beginner musicians.

For the purpose of music creation, rather than generating music signals directly from a textual prompt, with poor control and without any transparency, we aim at modelling each step of the music creation separately, for better control and transparency, with the possibility to help instrumentalists and composers rather than replacing them.

Future developments on the technology

The current module is in a preliminary state and it needs further improvements. First, connected with the DDSP-Piano module (also developed in the context of AI4Media, T5.2), we plan to train the networks directly on audio recordings rather than on symbolic performance. Indeed, the amount of expressive performances in the symbolic domain is limited, whereas thousands of piano recordings are available in the audio domain. Then, analyses of performance will be developed, for example in order to profile musician styles.



#4

DDSP-Piano Synthesizer

Partner organisations involved
IRCAM

→ **Code**

A few words about this technology

A polyphonic differentiable model for piano sound synthesis, conditioned on Musical Instrument Digital Interface (MIDI) inputs. The model architecture is motivated by high-level acoustic modelling knowledge of the instrument which, in tandem with the sound structure priors inherent to the DDSP components, makes for a lightweight, interpretable and realistic sounding piano model.

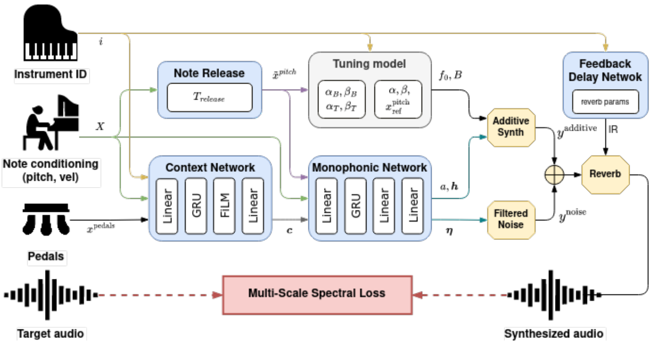
Who can benefit

Being a synthesiser, runnable in real-time, DDSP-Piano can be used for musicians who want to produce piano sounds of their composition, with MIDI devices, MIDI files, or a Digital Audio Workstation. But DDSP-Piano has been initially designed for researchers and developers. Thanks to its differentiable implementation, it is planned to insert it into deep neural networks, in order to train other tasks, such as automatic transcription, or performance analysis/rendering, in the audio domain and without aligned data.

Impact and added value for the media industry

Although DDSP-Piano can be used as a piano synthesiser, similar softwares, e.g. pianoteq, are already available for the music industry, offering significant realism for musicians.

Nevertheless, its differentiable architecture enables the continuation of further research works in order to achieve relevant results for the music sector, as mentioned previously: automatic transcription, piano source separation denoising, performance analysis/rendering. These tasks are very useful to analyse music catalogues of the industry. Additionally, such performance analyses make it possible to help beginner pianists by an analysis of their playing profile, a task which is related to active learning.



Future developments on the technology

Some improvements are needed: implement a denoising module in its architecture, and improve the estimation of parameters related to the inharmonicity and the detuning. Then, it is planned to insert DDSP-Piano into deep neural architectures, in order to train other deep learning models, such as: automatic piano transcription, and possibly piano source separation. Finally, we can derive a latent space to encode piano sound characteristics which will enable the exploration of different timbres in order to "sculpt" the sound as the pianist wants, using high-level controllers.

#5

Few-shot learning service

Partner organisations involved

JR and UPB

→ Code

A few words about this technology

We provide a service to train a few-shot object detector, i.e., adding support for new classes to an object detector, using the two-stage finetuning paradigm. Few-shot means that the method is capable of learning a new class from a limited number of samples, typically 10-30, or even less. The training backend supports efficient ensemble learning, which finetunes a set of models in parallel in order to improve the performance without requiring additional data.

In order to enable domain experts to perform the training themselves, we developed a service to trigger the training process from a set of images, together with annotations for the new classes. In order to streamline the annotation process, integration with a lightweight image annotation tool is provided. After training, the updated object detector is automatically deployed to TorchServe, enabling it to be tested directly on other data.

The training backend, the pipeline for running the training pipeline and the inference server are deployed as a single Docker container. A web user interface to run the entire pipeline from a browser is provided.

Who can benefit

This technology mostly benefits media professionals who need to annotate visual content (photos, videos) with specific objects, which may not be covered by object detectors trained on standard datasets. This includes for example journalists and editors who need to find content with objects of interest for a specific production, archivists handling specific requests, or preparing a topical collection, or investigative journalists who need to analyse collections of visual content.

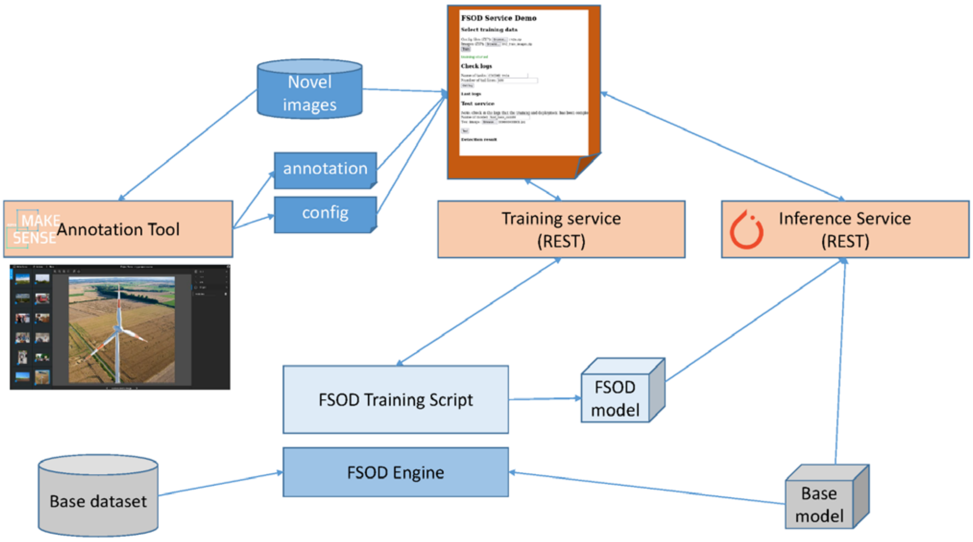
Impact and added value for the media industry

Up to now, training new classes for object detectors requires running a pipeline of training and deployment scripts. While this enables controlling the training process and running it on an organisations' own infrastructure, adding support for new classes to an object detector requires involving a developer to set up and run the pipeline. Some cloud service providers offer at least some support to achieve this in a low-code or no-code fashion. However, this means that (i) a base model trained on data not under control of the media organisation might be used, and (ii) the training data for new classes as well as the data to which the new detector is applied need to be shared with the service provider.

This technology enables professionals in the media industry to train object detectors without having deep technical knowledge. Detectors can be trained for specific objects of interest in a production context, and applied to automatically annotate a content set. The deployment of the services can be done with a few simple steps, and the process can be controlled via a web user interface. Due to the service-based approach, the control of the workflow can also be easily integrated into existing tools, such as digital asset management systems.

As it is possible to run the entire process on local infrastructure, this approach is also applicable when the underlying content is confidential, for example, in applications in investigative journalism.

This approach democratises the customisation of AI-based object detection technology to the needs of professionals in media production and archiving processes.



Future developments on the technology

Follow-up work addresses the efficient mining of training data from weakly annotated video content (e.g., with programme-level annotations, while only few frames contain the object of interest), as well as fast class-incremental training, in order to enable interactive use of the few-shot learning service.

#6

VISIONE System

Partner organisations involved

CNR

→ Code

A few words about this technology

VISIONE is an Interactive Video Retrieval System which leverages various technologies to make Video Retrieval effective and efficient. The system provides free text search, spatial colour and object search, and visual-semantic similarity search. More in details, VISIONE uses solutions based on Large Multimodal models to allow effective free text search on non-annotated videos while object detection and recognition is used to allow searching for specific objects in the scene, and visual-semantic similarity search is used to search for similar video segments. An index relying on specialised text encodings based on the Surrogate Text Representations (STRs) approach allows efficient searching on very large datasets.

Who can benefit

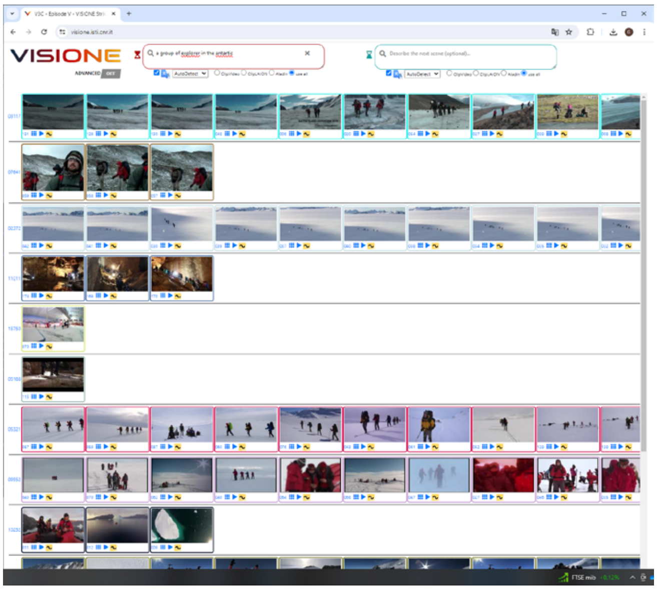
The VISIONE system might be useful to organisations that need to manage/store/retrieve large amounts of non-annotated visual data. It allows users to easily search for material of interest using natural language queries that are matched against real content of media.

Impact and added value for the media industry

The VISIONE system provides the media industry with a new valuable opportunity for managing large amounts of visual material. Thanks to its capabilities that allow users to search for non-annotated visual material, using natural language queries, it is possible to handle large visual archives that are poorly annotated or to retrieve visual documents focussing on details that were not taken into consideration at annotation time. In addition, it allows managing, organising, and searching the continuous flow of material provided, for instance, by content producers (freelance photographers, video makers, etc.) to media agencies, and allows potential users of this material (for instance journalists) to easily search and filter fresh material of their interest.

Future developments on the technology

We aim at leveraging solutions based on relevance feedback and on the use of innovative user interfaces to further improve the performance of the system.



a group of explorer in the antartic

Describe the next scene (optional)...

#7

MaskCon: Masked Contrastive Learning for Coarse-Labelled Dataset

Partner organisations involved

QMUL

→ Code

A few words about this technology

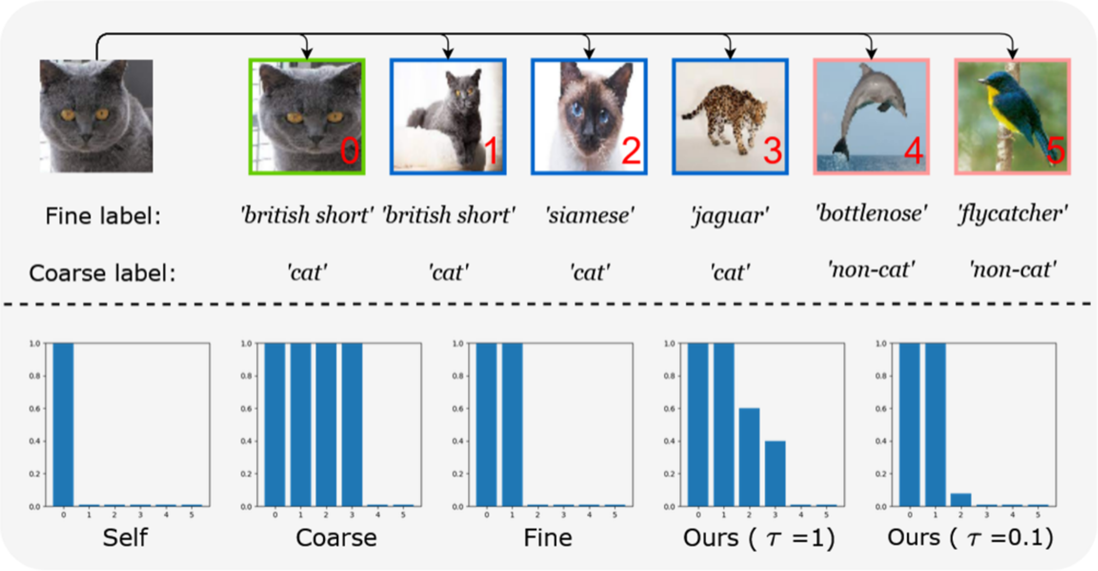
Deep learning has achieved great success in recent years with the aid of advanced neural network structures and large-scale human-annotated datasets. However, it is often costly and difficult to accurately and efficiently annotate large-scale datasets, especially for some specialised domains where fine-grained labels are required. In this setting, coarse labels are much easier to acquire as they do not require expert knowledge. In this work, we propose a contrastive learning method, called Masked Contrastive learning~(MaskCon) to address the under-explored problem setting, where we learn with a coarse-labelled dataset in order to address a finer labelling problem. More specifically, within the contrastive learning framework, for each sample our method generates soft-labels against other samples and another augmented view of the sample in question. By contrast to self-supervised contrastive learning where only the sample's augmentations are considered hard positives, and in supervised contrastive learning where only samples with the same coarse labels are considered hard positives, we propose soft labels based on sample distances that are masked by the coarse labels.
This allows us to utilise both inter-sample relations and coarse labels. Our method can obtain as special cases many existing state-of-the-art works and that it provides tighter bounds on the generalisation error.
Experimentally, our method achieves significant improvement over the current state-of-the-art in various datasets, including CIFAR10, CIFAR100, ImageNet-1K, Stanford Online Products and Stanford Cars196 datasets.

Who can benefit

This method could be useful to users in industry or in academia that work with fine-grained labelled data. MaskCon’s methodology allows users to train models with coarse-grained data, such that they are better positioned to classify and retrieve samples based on fine-grained semantics, for which labels are not plentiful.

Impact and added value for the media industry

The proposed Masked Contrastive learning (MaskCon) method has the potential to impact the media industry by addressing the challenges associated with fine-grained data annotation. The media industry heavily relies on accurate and detailed data for various applications, including content recommendation, audience analysis, and personalised advertising. However, acquiring finely labelled datasets is often expensive and time-consuming due to the need for expert annotation. This makes relying on standard supervised methods for training deep learning models challenging, as such models require large amounts of fine-grained labelled data, and, when trained, are restricted by the training data and may face challenges in terms of generalising to unseen fine-grained labels. MaskCon offers a solution, as it can leverage coarse-labelled data to learn representations that can effectively learn finer-grained distinctions, and therefore can be used in fine-grained tasks with improved results. This is particularly valuable for media companies that may have large volumes of data with basic labels but lack the resources for detailed annotations. By using MaskCon, these companies can improve their data's granularity without the prohibitive costs of manual fine-grained labelling.



Future developments on the technology

Extensive experiments demonstrate MaskCon's effectiveness across several datasets. In the future, we will investigate ways to overcome the limitations imposed by standard contrastive learning, such as the fact that it results in colour-invariant representations, which can be detrimental in scenarios where colour is a distinguishing feature between fine-grained classes.

#8

DivClust: Controlling Diversity in Deep Clustering

Partner organisations involved

QMUL

→ **Code**

A few words about this technology

Clustering has been a major research topic in the field of machine learning, one to which Deep Learning has recently been applied with significant success. However, an aspect of clustering that is not addressed by existing deep clustering methods, is that of efficiently producing multiple, diverse partitionings for a given dataset. This is particularly important, as a diverse set of base clusterings are necessary for consensus clustering, which has been found to produce better and more robust results than relying on a single clustering. This gap is addressed by DivClust,
a diversity controlling loss that can be incorporated into existing deep clustering frameworks to produce multiple clusterings with the desired degree of diversity. DivClust a) effectively controls diversity across frameworks and datasets with very small additional computational cost, b) learns sets of clusterings that include solutions that significantly outperform single-clustering baselines, and c) using off-the-shelf consensus clustering algorithms, DivClust produces consensus clustering solutions that consistently outperform single-clustering outcomes, effectively improving the performance of the base deep clustering framework.

Who can benefit

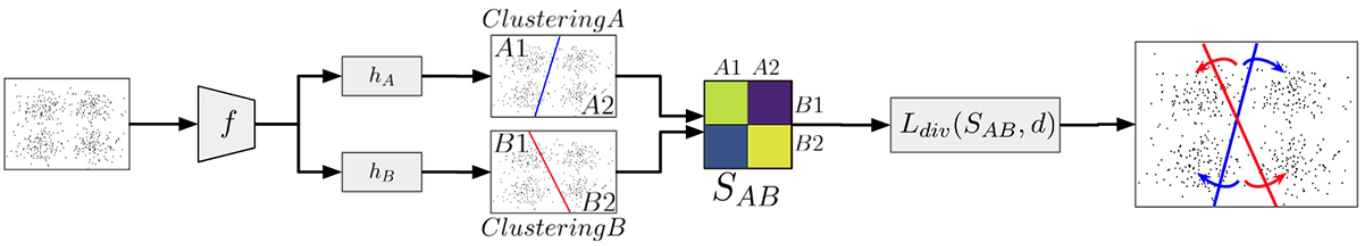
This method could be useful to users in industry or in academia that use clustering to analyse in the absence of labels. The developed method DivClust allows users to not only obtain clustering results of improved quality, but also to efficiently explore diverse clusterings for a given set of data, which may reveal different properties.

Impact and added value for the media industry

DivClust could have several applications in the media industry to address the limitations of existing deep clustering methods, particularly in producing multiple, diverse partitionings for datasets. This capability may be significant for media companies, which deal with vast and heterogeneous data sources, requiring nuanced and varied clustering to optimise their operations. DivClust can produce, for a given set of data, improved single-clustering solutions in terms of robustness and reliability, as well as diverse clusterings that capture different attributes of the data and partition them accordingly. This can be incorporated in various applications such as content recommendation systems, where meaningful clustering of users and content is critical, audience segmentation and analysis, and advertising. Furthermore, the ability of DivClust to integrate with existing deep clustering frameworks with minimal computational overhead is a significant advantage. Media companies can adopt this method without extensive resource allocation, ensuring a cost-effective implementation.

Future developments on the technology

The experiments conducted with several datasets and baseline deep clustering frameworks demonstrate the effectiveness of DivClust in controlling inter-clustering diversity and producing improved single-clustering solutions. In the future, we will seek to improve it by developing effective methods for applying it to web-scale datasets, for determining the optimal degree of diversity for a given dataset, and for semantically evaluating the quality of individual clusterings produced by DivClust.



#9

Playable Video Generation

Partner organisations involved

UNITN

→ Code

A few words about this technology

In Playable Video Generation (PVG), we aim at allowing a user to control the generated video by selecting a discrete action at every time step as when playing a video game. The difficulty of the task lies both in learning semantically consistent actions and in generating realistic videos conditioned on the user input. We propose a novel framework for PVG that is trained in a self-supervised manner on a large dataset of unlabelled videos. We employ an encoder-decoder architecture where the predicted action labels act as bottlenecks. The network is constrained to learn a rich action space using, as main driving loss, a reconstruction loss on the generated video. The effectiveness of the proposed approach was demonstrated on several datasets with wide environmental variety.

Who can benefit

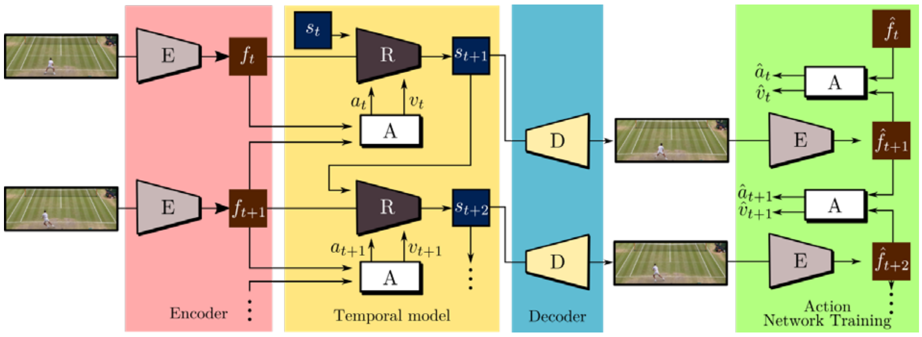
The tool could be useful to all the users (both professionals and amateurs) that want to produce videos based on a real game but employing real-video footage. The tool is dedicated to tennis but with appropriate knowledge can be extended to other types of games.

Impact and added value for the media industry

Humans at a very early age can identify key objects and understand how each object can interact with its environment. This ability is particularly notable when watching videos of sports or games. We can understand and anticipate actions in videos despite never being given an explicit list of plausible actions. We develop this skill in an unsupervised manner as we see actions live and on the screen. We can further analyse the technique with which an action is performed as well as the “amount” of action, i.e. how much to the left. Furthermore, we can reason about what happens if the player took a different action and how this would change the video. From this observation, PVG, aims to learn a set of distinct actions from real-world video clips in an unsupervised manner in order to offer the user the possibility to interactively generate new videos. At test time, the user provides a discrete action label at every time step and can see in real-time its impact on the generated video, similarly to video games. Introducing this novel problem gives the media industry actors the way toward methods that can automatically simulate real-world environments and provide a gaming-like experience.

Future developments on the technology

We evaluated our method on varied datasets and showed state of-the-art performance. Our experiments indicate that we can learn a rich set of actions that offer the user a gaming-like experience to control the generated video. As future work, we plan to extend our method to multi-agent environments.



#10

Diff-Vectors: Contrastive Vectorial Representations for Text Classification and their Application to Authorship Analysis

Partner organisations involved
CNR

→ Code

A few words about this technology

We have investigated the effects on authorship identification tasks (including authorship verification, closed-set authorship attribution, and closed-set and open-set same-author verification) of a fundamental shift in how to conceive the vectorial representations of texts that are given as input to a supervised learner for text classification purposes.
In "classic" authorship analysis, a feature vector represents a text, the value of a feature represents (an increasing function of) the relative frequency of the feature in the text, and the class label represents the author of the text.
We have instead investigated the situation in which a feature vector represents an unordered pair of texts, the value of a feature represents the absolute difference in the relative frequencies (or increasing functions thereof) of the feature in the two texts, and the class label indicates whether the two texts are from the same author or not.

This latter (learner-independent) type of representation has been occasionally used before, but has never been studied systematically. We argue that it is advantageous, and that, in some cases (e.g., authorship verification), it provides a much larger quantity of information to the training process than the standard representation. The experiments that we have carried out on several publicly available datasets show that feature vectors representing pairs of texts (that we here call Diff-Vectors) bring about systematic improvements in the effectiveness of authorship identification tasks, and especially so when training data are scarce (as it is often the case in real-life authorship identification scenarios).
Our experiments tackled same-author verification, authorship verification, and closed-set authorship attribution; while DVs are naturally geared for solving the 1st, we also provide two novel methods for solving the 2nd and 3rd that use a solver for the 1st as a building block.

Who can benefit

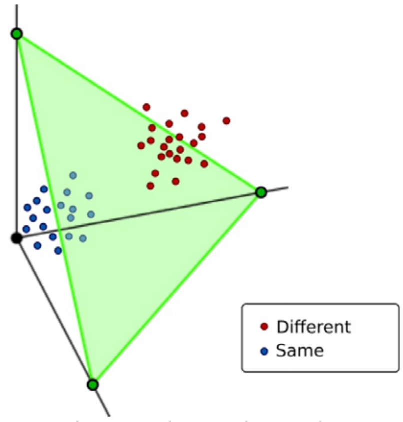
All applications of authorship analysis can benefit from these developments, since the tasks on which benefits have been reported (authorship attribution, authorship verification, same-author verification) are extremely general, and popular among practitioners.

Impact and added value for the media industry

While the experiments we have carried out concern authorship analysis, the methods we have discussed are obviously applicable to other text classification tasks.
In the near future, we plan to test them on the task of classification by topic (e.g., classifying news articles according to classes such as "Home News", "Sports", "Lifestyles", etc.), so as to check whether the advantages that DV-based methods have shown in authorship analysis tasks can also be enjoyed in contexts in which the dimension according to which texts are classified is not authorship.

Future developments on the technology

We also plan to test whether training data from different datasets can be merged and used for training the same-different verifier, since the "Same" and "Different" classes are general and dataset-independent.



AI4media

ARTIFICIAL INTELLIGENCE FOR
THE MEDIA AND SOCIETY

© Copyright 2024 AI4Media Consortium

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the AI4Media Consortium. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced. All rights reserved.

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