UMR IRISA

Activity Report 2024

Team LINKMEDIA

Creating and exploiting explicit links between multimedia fragments

Joint team with Centre Inria de l'Université de Rennes

D6 – Signal, Image, Language

Université de Rennes | rennes | INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES | cnrs | Inria | Inserm | CentraleSupélec | IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom | Université Bretagne Sud ubs

Contents

Project-Team LINKMEDIA 1

1 Team members, visitors, external collaborators 2

2 Overall objectives 3
2.1 Context . . . 3
2.2 Scientific objectives . . . 4

3 Research program 4
3.1 Scientific background . . . 4
3.2 Workplan . . . 5
3.3 Research Direction 1: Extracting and Representing Information . . . 5
3.4 Research Direction 2: Accessing Information . . . 8

4 Application domains 11
4.1 Asset management in the entertainment business . . . 11
4.2 Multimedia Internet . . . 11
4.3 Data journalism . . . 11

5 Social and environmental responsibility 11
5.1 Impact of research results . . . 11

6 Highlights of the year 12

7 New software, platforms, open data 12

8 New results 12
8.1 Extracting and Representing Information . . . 12
8.1.1 Decreasing graph complexity with transitive reduction to improve temporal graph classification . . . 12
8.1.2 DINOv2: Learning Robust Visual Features without Supervision . . . 12
8.1.3 Functional invariants to watermark large transformers . . . 13
8.1.4 Recherche de relation à partir d'un seul exemple fondée sur un modèle N-way K-shot : une histoire de distracteurs . . . 13
8.1.5 One-shot relation retrieval in news archives: adapting N-way K-shot relation classification for efficient knowledge extraction . . . 14
8.1.6 Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video . . . 14
8.1.7 AggNet: Learning to aggregate faces for group membership verification . . . 15
8.1.8 REStore: Exploring a Black-Box Defense against DNN Backdoors using Rare Event Simulation . . . 15
8.1.9 Proactive Detection of Voice Cloning with Localized Watermarking . . . 15
8.1.10 A Fast and Sound Tagging Method for Discontinuous Named-Entity Recognition . . . 16
8.1.11 Few-Shot Domain Adaptation for Named-Entity Recognition via Joint Constrained k-Means and Subspace Selection . . . 16
8.1.12 Training LayoutLM from Scratch for Efficient Named-Entity Recognition in the Insurance Domain . . . 16
8.1.13 WaterMax: breaking the LLM watermark detectability-robustness-quality trade-off . . . 16
8.1.14 When does gradient estimation improve black-box adversarial attacks? . . . 17
8.1.15 SWIFT: Semantic Watermarking for Image Forgery Thwarting . . . 17
8.1.16 Distinctive Image Captioning: Leveraging ground truth captions in CLIP guided reinforcement learning . . . 17
8.1.17 Fast Reliability Estimation for Neural Networks with Adversarial Attack-Driven Importance Sampling . . . 18
8.1.18 Watermarking Makes Language Models Radioactive . . . 18

8.1.19 A Double-Edged Sword: The Power of Two in Defending Against DNN Backdoor Attacks . . . 18
8.1.20 A Comprehensive Survey on Backdoor Attacks and Their Defenses in Face Recognition Systems . . . 19
8.1.21 Beyond Internet Images: Evaluating Vision-Language Models for Domain Generalization on Synthetic-to-Real Industrial Datasets . . . 19
8.1.22 HYBRINFOX at CheckThat! 2024 - Task 2: Enriching BERT Models with the Expert System VAGO for Subjectivity Detection . . . 19
8.2 Accessing Information . . . 20
8.2.1 A Multi-Label Dataset of French Fake News: Human and Machine Insights . . . 20
8.2.2 Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification . . . 20

9 Bilateral contracts and grants with industry **21**
9.1 Bilateral contracts with industry . . . 21

10 Partnerships and cooperations **22**
10.1 International initiatives . . . 22
10.1.1 Inria associate team not involved in an IIL or an international program . . . 22
10.1.2 STIC/MATH/CLIMAT AmSud projects . . . 23
10.2 National initiatives . . . 24

11 Dissemination **26**
11.1 Promoting scientific activities . . . 26
11.1.1 Scientific events: organisation . . . 26
11.1.2 Scientific events: selection . . . 26
11.1.3 Journal . . . 27
11.1.4 Invited talks . . . 27
11.1.5 Leadership within the scientific community . . . 27
11.1.6 Scientific expertise . . . 27
11.1.7 Research administration . . . 27
11.2 Teaching - Supervision - Juries . . . 27
11.2.1 Teaching . . . 27
11.2.2 Supervision . . . 28
11.2.3 Juries . . . 29
11.3 Popularization . . . 29
11.3.1 Productions (articles, videos, podcasts, serious games, ...) . . . 29
11.3.2 Participation in Live events . . . 29

12 Scientific production **29**
12.1 Major publications . . . 29
12.2 Publications of the year . . . 30
12.3 Cited publications . . . 33

Project-Team LINKMEDIA

Creation of the Project-Team: 2014 July 01

Keywords

Computer sciences and digital sciences

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A3.4. – Machine learning and statistics

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.8. – Deep learning

A4. – Security and privacy

A5.3.3. – Pattern recognition

A5.4.1. – Object recognition

A5.4.3. – Content retrieval

A5.7. – Audio modeling and processing

A5.7.1. – Sound

A5.7.3. – Speech

A5.8. – Natural language processing

A9.2. – Machine learning

A9.3. – Signal analysis

A9.4. – Natural language processing

Other research topics and application domains

B9. – Society and Knowledge

B9.3. – Medias

B9.6.10. – Digital humanities

B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientists

- Laurent Amsaleg [Team leader, CNRS, Senior Researcher]
- Teddy Furon [INRIA, Senior Researcher]
- Eva Giboulot [INRIA, Researcher, from Oct 2024]
- Guillaume Gravier [CNRS, Senior Researcher]

Faculty Members

- Caio Corro [INSA RENNES, Associate Professor, from Sep 2024]
- Ewa Kijak [UNIV RENNES I, Associate Professor]
- Simon Malinowski [UNIV RENNES I, Associate Professor]
- Pascale Sébillot [INSA RENNES, Professor]

Post-Doctoral Fellows

- Eva Giboulot [INRIA, Post-Doctoral Fellow, until Sep 2024]
- Ryan Webster [INRIA, Post-Doctoral Fellow]

PhD Students

- Adèle Denis [INRAE, from Sep 2024]
- Virgile Dine [INRIA, from Sep 2024]
- Deniz Engin [INRIA, until Feb 2024]
- Gautier Evennou [IMATAG, CIFRE]
- Pierre Fernandez [FACEBOOK, CIFRE]
- Enoal Gesny [INRIA, from Oct 2024]
- Louis Hemadou [SAFRAN, CIFRE]
- Chloé Imadache [INRIA, from Oct 2024]
- Carolina Jeronimo De Almeida [GOUV BRESIL]
- Quentin Le Roux [THALES, CIFRE]
- Hugo Thomas [UNIV RENNES I]
- Karim Tit [INRIA, from Feb 2024 until Mar 2024]
- Karim Tit [THALES, until Jan 2024]
- Shashanka Venkataramanan [INRIA, until May 2024]

Technical Staff

- Morgane Casanova [CNRS, from May 2024 until Oct 2024, Engineer]
- Morgane Casanova [CNRS, Engineer, until Apr 2024]
- Nicolas Fouqué [CNRS, from Mar 2024, Engineer]

Interns and Apprentices

- Enoal Gesny [INRIA, Intern, from Apr 2024 until Sep 2024]
- Chloé Imadache [INRIA, Intern, from May 2024 until Sep 2024]
- Amelie Knecht [UNIV RENNES I, from Sep 2024]

Administrative Assistants

- Aurélie Patier [UNIV RENNES I, until Jul 2024]
- Sabrina Ysope [INRIA, from Jul 2024]

Visiting Scientists

- Isabela Borlido Barcelos [GOUV BRESIL, from Sep 2024 until Sep 2024]
- Caio Corro [SORBONNE UNIVERSITE, from Jul 2024 until Aug 2024]

External Collaborator

- Charly Faure [DGA-MI]

2 Overall objectives

2.1 Context

LINKMEDIA is concerned with the processing of extremely large collections of multimedia material. The material we refer to are collections of documents that are created by humans and intended for humans. It is material that is typically created by media players such as TV channels, radios, newspapers, archivists (BBC, INA, ...), as well as the multimedia material that goes through social-networks. It has images, videos and pathology reports for e-health applications, or that is in relation with e-learning which typically includes a fair amount of texts, graphics, images and videos associating in new ways teachers and students. It also includes material in relation with humanities that study societies through the multimedia material that has been produced across the centuries, from early books and paintings to the latest digitally native multimedia artifacts. Some other multimedia material are out of the scope of LINKMEDIA, such as the ones created by cameras or sensors in the broad areas of video-surveillance or satellite images.

Multimedia collections are rich in contents and potential, that richness being in part within the documents themselves, in part within the relationships between the documents, in part within what humans can discover and understand from the collections before materializing its potential into new applications, new services, new societal discoveries, ... That richness, however, remains today hardly accessible due to the conjunction of several factors originating from the inherent nature of the collections, the complexity of bridging the semantic gap or the current practices and the (limited) technology:

- *Multimodal:* multimedia collections are composed of very diverse material (images, texts, videos, audio, ...), which require sophisticated approaches at analysis time. Scientific contributions from past decades mostly focused on analyzing each media in isolation one from the other, using modality-specific algorithms. However, revealing the full richness of collections calls for jointly taking into account these multiple modalities, as they are obviously semantically connected. Furthermore, involving resources that are external to collections, such as knowledge bases, can only improve gaining insight into the collections. Knowledge bases form, in a way, another type of modality with specific characteristics that also need to be part of the analysis of media collections. Note that determining what a document is about possibly mobilizes a lot of resources, and this is especially costly and time consuming for audio and video. Multimodality is a great source of richness, but causes major difficulties for the algorithms running analysis;

- *Intertwined*: documents do not exist in isolation one from the other. There is more knowledge in a collection than carried by the sum of its individual documents and the relationships between documents also carry a lot of meaningful information. (Hyper)Links are a good support for materializing the relationships between documents, between parts of documents, and having analytic processes creating them automatically is challenging. Creating semantically rich typed links, linking elements at very different granularities is very hard to achieve. Furthermore, in addition to being disconnected, there is often no strong structure into each document, which makes even more difficult their analysis;

- *Collections are very large:* the scale of collections challenges any algorithm that runs analysis tasks, increasing the duration of the analysis processes, impacting quality as more irrelevant multimedia material gets in the way of relevant ones. Overall, scale challenges the complexity of algorithms as well as the quality of the result they produce;

- *Hard to visualize*: It is very difficult to facilitate humans getting insight on collections of multimedia documents because we hardly know how to display them due to their multimodal nature, or due to their number. We also do not know how to well present the complex relationships linking documents together: granularity matters here, as full documents can be linked with small parts from others. Furthermore, visualizing time-varying relationships is not straightforward. Data visualization for multimedia collections remains quite unexplored.

2.2 Scientific objectives

The ambition of LINKMEDIA is to propose **foundations, methods, techniques and tools to help humans make sense of extremely large collections of multimedia material**. Getting useful insight from multimedia is only possible if tools and users interact tightly. Accountability of the analysis processes is paramount in order to allow users understanding their outcome, to understand why some multimedia material was classified this way, why two fragments of documents are now linked. It is key for the acceptance of these tools, or for correcting errors that will exist. Interactions with users, facilitating analytics processes, taking into account the trust in the information and the possible adversarial behaviors are topics LINKMEDIA addresses.

3 Research program

3.1 Scientific background

LINKMEDIA is de facto a multidisciplinary research team in order to gather the multiple skills needed to enable humans to gain insight into extremely large collections of multimedia material. It is *multimedia data* which is at the core of the team and which drives the design of our scientific contributions, backed-up with solid experimental validations. *Multimedia data*, again, is the rationale for selecting problems, applicative fields and partners.

Our activities therefore include studying the following scientific fields:

- multimedia: content-based analysis; multimodal processing and fusion; multimedia applications;

- computer vision: compact description of images; object and event detection;

- machine learning: deep architectures; structured learning; adversarial learning;

- natural language processing: topic segmentation; information extraction;

- information retrieval: high-dimensional indexing; approximate k-nn search; embeddings;

- data mining: time series mining; knowledge extraction.

3.2 Workplan

Overall, LINKMEDIA follows two main directions of research that are (i) extracting and representing information from the documents in collections, from the relationships between the documents and from what user build from these documents, and (ii) facilitating the access to documents and to the information that has been elaborated from their processing.

3.3 Research Direction 1: Extracting and Representing Information

LINKMEDIA follows several research tracks for *extracting* knowledge from the collections and *representing* that knowledge to facilitate users acquiring gradual, long term, constructive insights. Automatically processing documents makes it crucial to consider the accountability of the algorithms, as well as understanding when and why algorithms make errors, and possibly invent techniques that compensate or reduce the impact of errors. It also includes dealing with malicious adversaries carefully manipulating the data in order to compromise the whole knowledge extraction effort. In other words, LINKMEDIA also investigates various aspects related to the *security* of the algorithms analyzing multimedia material for knowledge extraction and representation.

Knowledge is not solely extracted by algorithms, but also by humans as they gradually get insight. This human knowledge can be materialized in computer-friendly formats, allowing algorithms to use this knowledge. For example, humans can create or update ontologies and knowledge bases that are in relation with a particular collection, they can manually label specific data samples to facilitate their disambiguation, they can manually correct errors, etc. In turn, knowledge provided by humans may help algorithms to then better process the data collections, which provides higher quality knowledge to humans, which in turn can provide some better feedback to the system, and so on. This virtuous cycle where algorithms and humans cooperate in order to make the most of multimedia collections requires specific support and techniques, as detailed below.

Machine Learning for Multimedia Material. Many approaches are used to extract relevant information from multimedia material, ranging from very low-level to higher-level descriptions (classes, captions, ...). That diversity of information is produced by algorithms that have varying degrees of supervision. Lately, fully supervised approaches based on deep learning proved to outperform most older techniques. This is particularly true for the latest developments of Recurrent Neural Networkds (RNN, such as LSTMs) or convolutional neural network (CNNs) for images that reach excellent performance [54]. LINKMEDIA contributes to advancing the state of the art in computing representations for multimedia material by investigating the topics listed below. Some of them go beyond the very processing of multimedia material as they also question the fundamentals of machine learning procedures when applied to multimedia.

- *Learning from few samples/weak supervisions.* CNNs and RNNs need large collections of carefully annotated data. They are not fitted for analyzing datasets where few examples per category are available or only cheap image-level labels are provided. LINKMEDIA investigates low-shot, semi-supervised and weakly supervised learning processes: Augmenting scarce training data by automatically propagating labels [57], or transferring what was learned on few very well annotated samples to allow the precise processing of poorly annotated data [66]. Note that this context also applies to the processing of heritage collections (paintings, illuminated manuscripts, ...) that strongly differ from contemporary natural images. Not only annotations are scarce, but the learning processes must cope with material departing from what standard CNNs deal with, as classes such as "planes", "cars", etc, are irrelevant in this case.

- *Ubiquitous Training.* NN (CNNs, LSTMs) are mainstream for producing representations suited for high-quality classification. Their training phase is ubiquitous because the same representations can be used for tasks that go beyond classification, such as retrieval, few-shot, meta- and incremental learning, all boiling down to some form of metric learning. We demonstrated that this ubiquitous training is relatively simpler [57] yet as powerful as ad-hoc strategies fitting specific tasks [71]. We study the properties and the limitations of this ubiquitous training by casting metric learning as a classification problem.

- *Beyond static learning*. Multimedia collections are by nature continuously growing, and ML processes must adapt. It is not conceivable to re-train a full new model at every change, but rather to support continuous training and/or allowing categories to evolve as the time goes by. New classes may be defined from only very few samples, which links this need for dynamicity to the low-shot learning problem discussed here. Furthermore, active learning strategies determining which is the next sample to use to best improve classification must be considered to alleviate the annotation cost and the re-training process [61]. Eventually, the learning process may need to manage an extremely large number of classes, up to millions. In this case, there is a unique opportunity of blending the expertise of LINKMEDIA on large scale indexing and retrieval with deep learning. Base classes can either be "summarized" e.g. as a multi-modal distribution, or their entire training set can be made accessible as an external associative memory [77].

- *Learning and lightweight architectures.* Multimedia is everywhere, it can be captured and processed on the mobile devices of users. It is necessary to study the design of lightweight ML architectures for mobile and embedded vision applications. Inspired by [81], we study the savings from quantizing hyper-parameters, pruning connections or other approximations, observing the trade-off between the footprint of the learning and the quality of the inference. Once strategy of choice is progressive learning which early aborts when confident enough [62].

- *Multimodal embeddings.* We pursue pioneering work of LINKMEDIA on multimodal embedding, i.e., representing multiple modalities or information sources in a single embedded space [75, 74, 76]. Two main directions are explored: exploiting adversarial architectures (GANs) for embedding via translation from one modality to another, extending initial work in [76] to highly heterogeneous content; combining and constraining word and RDF graph embeddings to facilitate entity linking and explanation of lexical co-occurrences [51].

- *Accountability of ML processes.* ML processes achieve excellent results but it is mandatory to verify that accuracy results from having determined an adequate problem representation, and not from being abused by artifacts in the data. LINKMEDIA designs procedures for at least explaining and possibly interpreting and understanding what the models have learned. We consider heat-maps materializing which input (pixels, words) have the most importance in the decisions [70], Taylor decompositions to observe the individual contributions of each relevance scores or estimating LID [38] as a surrogate for accounting for the smoothness of the space.

- *Extracting information.* ML is good at extracting features from multimedia material, facilitating subsequent classification, indexing, or mining procedures. LINKMEDIA designs extraction processes for identifying parts in the images [67, 68], relationships between the various objects that are represented in images [44], learning to localizing objects in images with only weak, image-level supervision [70] or fine-grained semantic information in texts [49]. One technique of choice is to rely on generative adversarial networks (GAN) for learning low-level representations. These representations can e.g. be based on the analysis of density [80], shading, albedo, depth, etc.

- *Learning representations for time evolving multimedia material.* Video and audio are time evolving material, and processing them requests to take their time line into account. In [63, 48] we demonstrated how shapelets can be used to transform time series into time-free high-dimensional vectors, preserving however similarities between time series. Representing time series in a metric space improves clustering, retrieval, indexing, metric learning, semi-supervised learning and many other machine learning related tasks. Research directions include adding localization information to the shapelets, fine-tuning them to best fit the task in which they are used as well as designing hierarchical representations.

Adversarial Machine Learning. Systems based on ML take more and more decisions on our behalf, and maliciously influencing these decisions by crafting adversarial multimedia material is a potential source of dangers: a small amount of carefully crafted noise imperceptibly added to images corrupts classification and/or recognition. This can naturally impact the insight users get on the multimedia collection they work with, leading to taking erroneous decisions for example.

This adversarial phenomenon is not particular to deep learning, and can be observed even when using other ML approaches [43]. Furthermore, it has been demonstrated that adversarial samples generalize very well across classifiers, architectures, training sets. The reasons explaining why such tiny content modifications succeed in producing severe errors are still not well understood.

We are left with little choice: we must gain a better understanding of the weaknesses of ML processes, and in particular of deep learning. We must understand why attacks are possible as well as discover mechanisms protecting ML against adversarial attacks (with a special emphasis on convolutional neural networks). Some initial contributions have started exploring such research directions, mainly focusing on images and computer vision problems. Very little has been done for understanding adversarial ML from a *multimedia* perspective [47].

LINKMEDIA is in a unique position to throw at this problem new perspectives, by experimenting with other modalities, used in isolation one another, as well as experimenting with true multimodal inputs. This is very challenging, and far more complicated and interesting than just observing adversarial ML from a computer vision perspective. No one clearly knows what is at stake with adversarial audio samples, adversarial video sequences, adversarial ASR, adversarial NLP, adversarial OCR, all this being often part of a sophisticated multimedia processing pipeline.

Our ambition is to lead the way for initiating investigations where the full diversity of modalities we are used to work with in multimedia are considered from a perspective of adversarial attacks and defenses, both at learning and test time. In addition to what is described above, and in order to trust the multimedia material we analyze and/or the algorithms that are at play, LINKMEDIA investigates the following topics:

- *Beyond classification.* Most contributions in relation with adversarial ML focus on classification tasks. We started investigating the impact of adversarial techniques on more diverse tasks such as retrieval [37]. This problem is related to the very nature of euclidean spaces where distances and neighborhoods can all be altered. Designing defensive mechanisms is a natural companion work.

- *Detecting false information.* We carry-on with earlier pioneering work of LINKMEDIA on false information detection in social media. Unlike traditional approaches in image forensics [52], we build on our expertise in content-based information retrieval to take advantage of the contextual information available in databases or on the web to identify out-of-context use of text or images which contributed to creating a false information [64].

- *Deep fakes.* Progress in deep ML and GANs allow systems to generate realistic images and are able to craft audio and video of existing people saying or doing things they never said or did [60]. Gaining in sophistication, these machine learning-based "deep fakes" will eventually be almost indistinguishable from real documents, making their detection/rebutting very hard. LINKMEDIA develops deep learning based counter-measures to identify such modern forgeries. We also carry on with making use of external data in a provenance filtering perspective [69] in order to debunk such deep fakes.

- *Distributions, frontiers, smoothness, outliers.* Many factors that can possibly explain the adversarial nature of some samples are in relation with their distribution in space which strongly differs from the distribution of natural, genuine, non adversarial samples. We are investigating the use of various information theoretical tools that facilitate observing distributions, how they differ, how far adversarial samples are from benign manifolds, how smooth is the feature space, etc. In addition, we are designing original adversarial attacks and develop detection and curating mechanisms [38].

Multimedia Knowledge Extraction. Information obtained from collections via computer ran processes is not the only thing that needs to be represented. Humans are in the loop, and they gradually improve their level of understanding of the content and nature of the multimedia collection. Discovering knowledge and getting insight is involving multiple people across a long period of time, and what each understands, concludes and discovers must be recorded and made available to others. Collaboratively inspecting collections is crucial. Ontologies are an often preferred mechanism for modeling what is inside a collection, but this is probably limitative and narrow.

LINKMEDIA is concerned with making use of existing strategies in relation with ontologies and knowledge bases. In addition, LINKMEDIA uses mechanisms allowing to materialize the knowledge gradually acquired by humans and that might be subsequently used either by other humans or by computers in order to better and more precisely analyze collections. This line of work is instantiated at the core of the iCODA project LINKMEDIA coordinates.

We are therefore concerned with:

- *Multimedia analysis and ontologies.* We develop approaches for linking multimedia content to entities in ontologies for text and images, building on results in multimodal embedding to cast entity linking into a nearest neighbor search problem in a high-dimensional joint embedding of content and entities [74]. We also investigate the use of ontological knowledge to facilitate information extraction from content [51].

- *Explainability and accountability in information extraction.* In relation with ontologies and entity linking, we develop innovative approaches to explain statistical relations found in data, in particular lexical or entity co-occurrences in textual data, for example using embeddings constrained with translation properties of RDF knowledge or path-based explanation within RDF graphs. We also work on confidence measures in entity linking and information extraction, studying how the notions of confidence and information source can be accounted for in knowledge basis and used in human-centric collaborative exploration of collections.

- *Dynamic evolution of models for information extraction.* In interactive exploration and information extraction, e.g., on cultural or educational material, knowledge progressively evolves as the process goes on, requiring on-the-fly design of new models for content-based information extractors from very few examples, as well as continuous adaptation of the models. Combining in a seamless way low-shot, active and incremental learning techniques is a key issue that we investigate to enable this dynamic mechanisms on selected applications.

3.4 Research Direction 2: Accessing Information

LINKMEDIA centers its activities on enabling humans to make good use of vast multimedia collections. This material takes all its cultural and economic value, all its artistic wonder when it can be accessed, watched, searched, browsed, visualized, summarized, classified, shared, ... This allows users to fully enjoy the incalculable richness of the collections. It also makes it possible for companies to create business rooted in this multimedia material.

Accessing the multimedia data that is inside a collection is complicated by the various type of data, their volume, their length, etc. But it is even more complicated to access the information that is not materialized in documents, such as the relationships between parts of different documents that however share some similarity. LINKMEDIA in its first four years of existence established itself as one of the leading teams in the field of multimedia analytics, contributing to the establishment of a dedicated community (refer to the various special sessions we organized with MMM, the iCODA and the LIMAH projects, as well as [58, 59, 55]).

Overall, facilitating the access to the multimedia material, to the relevant information and the corresponding knowledge asks for algorithms that efficiently *search* collections in order to identify the elements of collections or of the acquired knowledge that are matching a query, or that efficiently allow *navigating* the collections or the acquired knowledge. Navigation is likely facilitated if techniques are able to handle information and knowledge according to hierarchical perspectives, that is, allow to reveal data according to various levels of details. Aggregating or *summarizing* multimedia elements is not trivial.

Three topics are therefore in relation with this second research direction. LINKMEDIA tackles the issues in relation to searching, to navigating and to summarizing multimedia information. Information needs when discovering the content of a multimedia collection can be conveniently mapped to the exploration-search axis, as first proposed by Zahálka and Worring in [79], and illustrated by Figure 1 where expert users typically work near the right end because their tasks involve precise queries probing search engines. In contrast, lay-users start near the exploration end of the axis. Overall, users may alternate searches and explorations by going back and forth along the axis. The underlying model and system must therefore be highly dynamic, support interactions with the users and propose means for

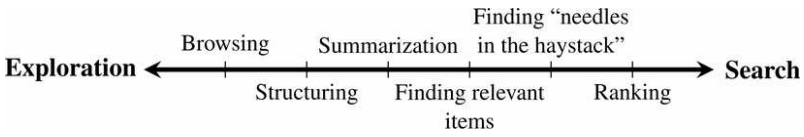

Figure 1: Exploration-search axis with example tasks

easy refinements. LINKMEDIA contributes to advancing the state of the art in searching operations, in navigating operations (also referred to as browsing), and in summarizing operations.

Searching. Search engines must run similarity searches very efficiently. High-dimensional indexing techniques therefore play a central role. Yet, recent contributions in ML suggest to revisit indexing in order to adapt to the specific properties of modern features describing contents.

- *Advanced scalable indexing.* High-dimensional indexing is one of the foundations of LINKMEDIA. Modern features extracted from the multimedia material with the most recent ML techniques shall be indexed as well. This, however, poses a series of difficulties due to the dimensionality of these features, their possible sparsity, the complex metrics in use, the task in which they are involved (instance search, k-nn, class prototype identification, manifold search [57], time series retrieval, ...). Furthermore, truly large datasets require involving sketching [41], secondary storage and/or distribution [40, 39], alleviating the explosion of the number of features to consider due to their local nature or other innovative methods [56], all introducing complexities. Last, indexing multimodal embedded spaces poses a new series of challenges.

- *Improving quality.* Scalable indexing techniques are approximate, and what they return typically includes a fair amount of false positives. LINKMEDIA works on improving the quality of the results returned by indexing techniques. Approaches taking into account neighborhoods [50], manifold structures instead of pure distance based similarities [57] must be extended to cope with advanced indexing in order to enhance quality. This includes feature selection based on intrinsic dimensionality estimation [38].

- *Dynamic indexing.* Feature collections grow, and it is not an option to fully reindex from scratch an updated collection. This trivially applies to the features directly extracted from the media items, but also to the base class prototypes that can evolve due to the non-static nature of learning processes. LINKMEDIA will continue investigating what is at stake when designing dynamic indexing strategies.

Navigating. Navigating a multimedia collection is very central to its understanding. It differs from searching as navigation is not driven by any specific query. Rather, it is mostly driven by the relationships that various documents have one another. Relationships are supported by the links between documents and/or parts of documents. Links rely on semantic similarity, depicting the fact that two documents share information on the same topic. But other aspects than semantics are also at stake, e.g., time with the dates of creation of the documents or geography with mentions or appearance in documents of some geographical landmarks or with geo-tagged data.

In multimedia collections, links can be either implicit or explicit, the latter being much easier to use for navigation. An example of an implicit link can be the name of someone existing in several different news articles; we, as humans, create a mental link between them. In some cases, the computer misses such configurations, leaving such links implicit. Implicit links are subject to human interpretation, hence they are sometimes hard to identify for any automatic analysis process. Implicit links not being materialized, they can therefore hardly be used for navigation or faceted search. Explicit links can typically be seen as hyperlinks, established either by content providers or, more aligned with LINKMEDIA, automatically determined from content analysis. Entity linking (linking content to an entity referenced in a knowledge base) is a good example of the creation of explicit links. Semantic similarity links, as investigated in the LIMAH project and as considered in the search and hyperlinking task at MediaEval and TRECVid, are

also prototypical links that can be made explicit for navigation. Pursuing work, we investigate two main issues:

- *Improving multimodal content-based linking.* We exploit achievements in entity linking to go beyond lexical or lexico-visual similarity and to provide semantic links that are easy to interpret for humans; carrying on, we work on link characterization, in search of mechanisms addressing link explainability (i.e., what is the nature of the link), for instance using attention models so as to focus on the common parts of two documents or using natural language generation; a final topic that we address is that of linking textual content to external data sources in the field of journalism, e.g., leveraging topic models and cue phrases along with a short description of the external sources.

- *Dynamicity and user-adaptation.* One difficulty for explicit link creation is that links are often suited for one particular usage but not for another, thus requiring creating new links for each intended use; whereas link creation cannot be done online because of its computational cost, the alternative is to generate (almost) all possible links and provide users with selection mechanisms enabling personalization and user-adaptation in the exploration process; we design such strategies and investigate their impact on exploration tasks in search of a good trade-off between performance (few high-quality links) and genericity.

Summarizing. Multimedia collections contain far too much information to allow any easy comprehension. It is mandatory to have facilities to aggregate and summarize a large body on information into a compact, concise and meaningful representation facilitating getting insight. Current technology suggests that multimedia content aggregation and story-telling are two complementary ways to provide users with such higher-level views. Yet, very few studies already investigated these issues. Recently, video or image captioning [78, 73] have been seen as a way to summarize visual content, opening the door to state-of-the-art multi-document text summarization [53] with text as a pivot modality. Automatic story-telling has been addressed for highly specific types of content, namely TV series [45] and news [65, 72], but still need a leap forward to be mostly automated, e.g., using constraint-based approaches for summarization [42, 72].

Furthermore, not only the original multimedia material has to be summarized, but the knowledge acquired from its analysis is also to summarize. It is important to be able to produce high-level views of the relationships between documents, emphasizing some structural distinguishing qualities. Graphs establishing such relationships need to be constructed at various level of granularity, providing some support for summarizing structural traits.

Summarizing multimedia information poses several scientific challenges that are:

- *Choosing the most relevant multimedia aggregation type*: Taking a multimedia collection into account, a same piece of information can be present in several modalities. The issue of selecting the most suitable one to express a given concept has thus to be considered together with the way to mix the various modalities into an acceptable production. Standard summarization algorithms have to be revisited so that they can handle continuous representation spaces, allowing them to benefit from the various modalities [46].

- *Expressing user's preferences*: Different users may appreciate quite different forms of multimedia summaries, and convenient ways to express their preferences have to be proposed. We for example focus on the opportunities offered by the constraint-based framework.

- *Evaluating multimedia summaries*: Finding criteria to characterize what a good summary is remains challenging, e.g., how to measure the global relevance of a multimodal summary and how to compare information between and across two modalities. We tackle this issue particularly via a collaboration with A. Smeaton at DCU, comparing the automatic measures we will develop to human judgments obtained by crowd-sourcing.

- *Taking into account structuring and dynamicity*: Typed links between multimedia fragments, and hierarchical topical structures of documents obtained via work previously developed within the team are two types of knowledge which have seldom been considered as long as summarization is concerned. Knowing that the event present in a document is causally related to another event

described in another document can however modify the ways summarization algorithms have to consider information. Moreover the question of producing coarse-to-fine grain summaries exploiting the topical structure of documents is still an open issue. Summarizing dynamic collections is also challenging and it is one of the questions we consider.

4 Application domains

4.1 Asset management in the entertainment business

Media asset management—archiving, describing and retrieving multimedia content—has turned into a key factor and a huge business for content and service providers. Most content providers, with television channels at the forefront, rely on multimedia asset management systems to annotate, describe, archive and search for content. So do archivists such as the Institut National de l'Audiovisuel, the bibliothèque Nationale de France, the Nederlands Instituut voor Beeld en Geluid or the British Broadcast Corporation, as well as media monitoring companies, such as Yacast in France. Protecting copyrighted content is another aspect of media asset management.

4.2 Multimedia Internet

One of the most visible application domains of linked multimedia content is that of multimedia portals on the Internet. Search engines now offer many features for image and video search. Video sharing sites also feature search engines as well as recommendation capabilities. All news sites provide multimedia content with links between related items. News sites also implement content aggregation, enriching proprietary content with user-generated content and reactions from social networks. Most public search engines and Internet service providers offer news aggregation portals. This also concerns TV on-demand and replay services as well as social TV services and multi-screen applications. Enriching multimedia content, with explicit links targeting either multimedia material or knowledge databases is central here.

4.3 Data journalism

Data journalism forms an application domain where most of the technology developed by LINKMEDIA can be used. On the one hand, data journalists often need to inspect multiple heterogeneous information sources, some being well structured, some other being fully unstructured. They need to access (possibly their own) archives with either searching or navigational means. To gradually construct insight, they need collaborative multimedia analytics processes as well as elements of trust in the information they use as foundations for their investigations. Trust in the information, watching for adversarial and/or (deep) fake material, accountability are all crucial here.

5 Social and environmental responsibility

5.1 Impact of research results

The SYNAPSES Labcom The year 2024 is marked by close collaboration with a major French media organization. The Linkmedia Ouest-France team is launching Synapses, the first "joint laboratory" with a press organization to develop AI for journalism. Supported by the French National Research Agency (ANR), it comes after thirty years of partnership, and targets the analysis of photo archives, the processing of historical texts and the visualization of complex data. Synapses combines "AI and data sovereignty" to exploit a unique heritage of 105 million documents. This partnership highlights the sharing of scientific knowledge, but also our respective sensitivities to the societal impact of AI in order to work on better information for diverse audiences.

6 Highlights of the year

The Linkmedia team split in September 2024, giving birth to the Artishau team, lead by Teddy Furon. Artishau is composed of many (old) members of Linkmedia plus a few members from the Wide team.

Because that split is very recent, the current activity report merges the elements from Artishau and the ones from Linkmedia, leading to the writing of a single report for both teams.

In order to materialize the split, a few elements are important to note:

- Eva Giboulot got a permanent INRIA research position (CRCN) at the time of that split, and she is now a member of Artishau. Before that, she was a post-doc inside Linkmedia.
- 4 students (Denis, Dine, Gesny, Imadache) started their PhD at the time of that split.
- Publications of Artishau have been added to the ones of Linkmedia.

Artishau will have its own activity report in 2025.

7 New software, platforms, open data

8 New results

8.1 Extracting and Representing Information

8.1.1 Decreasing graph complexity with transitive reduction to improve temporal graph classification

Participants: Carolina Jerônimo, Zenilton Patrocínio *(PUC Minas - Pontifícia Universidade Católica de Minas Gerais)*, Simon Malinowski, Guillaume Gravier, Silvio Guimarães *(PUC Minas - Pontifícia Universidade Católica de Minas Gerais)*.

Domains such as bioinformatics, social network analysis, and computer vision, describe relations between entities and cannot be interpreted as vectors or fixed grids. Instead, they are naturally represented by graphs. Often this kind of data evolves over time in a dynamic world, respecting a temporal order being known as temporal graphs. The latter became a challenge since subgraph patterns are very difficult to find and the distance between those patterns may change irregularly over time. While state-of-the-art methods are primarily designed for static graphs and may not capture temporal information, recent works have proposed mapping temporal graphs to static graphs to allow for the use of conventional static kernels approaches. This work presents a new method for temporal graph classification based on transitive reduction, which explores new kernels and graph neural networks for temporal graph classification [11]. We compare the transitive reduction impact on the map to static graphs in terms of accuracy and computational efficiency across different classification tasks. Experimental results demonstrate the effectiveness of the proposed mapping method in improving the accuracy of supervised classification for temporal graphs while maintaining reasonable computational efficiency.

8.1.2 DINOv2: Learning Robust Visual Features without Supervision

Participants: Maxime Oquab *(Meta AI)*, Timothée Darcet *(Meta AI, Thoth)*, Théo Moutakanni *(CentraleSupélec, Meta AI, Université Paris-Saclay)*, Huy Vo *(Meta AI)*, Marc Szafraniec *(Meta AI)*, Vasil Khalidov *(Meta AI)*, Pierre Fernandez *(Meta AI)*, Daniel Haziza *(Meta AI)*, Francisco Massa *(Meta AI)*, Alaaeldin El-Nouby *(Meta AI)*, Mahmoud Assran *(Meta AI)*, Nicolas Ballas *(Meta AI)*, Wojciech Galuba *(Meta AI)*, Russell Howes *(Meta AI)*, Po-Yao Huang *(Meta AI)*, Shang-Wen Li *(Meta AI)*, Ishan Misra *(Meta AI)*, Michael Rabbat *(Meta AI)*, Vasu Sharma *(Meta AI)*, Gabriel Synnaeve *(Meta AI)*, Hu Xu *(Meta AI)*, Hervé Jegou *(Meta AI)*, Julien Mairal *(Thoth)*, Patrick Labatut *(Meta AI)*, Armand Joulin *(Meta AI)*, Piotr Bojanowski *(Meta AI)*.

The recent breakthroughs in natural language processing for model pretraining on large quantities of data have opened the way for similar foundation models in computer vision. These models could greatly simplify the use of images in any system by producing all-purpose visual features, i.e., features that work across image distributions and tasks without finetuning. This work shows that existing pretraining methods, especially self-supervised methods, can produce such features if trained on enough curated data from diverse sources. We revisit existing approaches and combine different techniques to scale our pretraining in terms of data and model size [12]. Most of the technical contributions aim at accelerating and stabilizing the training at scale. In terms of data, we propose an automatic pipeline to build a dedicated, diverse, and curated image dataset instead of uncurated data, as typically done in the self-supervised literature. In terms of models, we train a ViT model with 1B parameters and distill it into a series of smaller models that surpass the best available all-purpose features, OpenCLIP on most of the benchmarks at image and pixel levels.

8.1.3 Functional invariants to watermark large transformers

Participants: Pierre Fernandez *(Meta)*, Guillaume Couairon *(Meta)*, Teddy Furon, Matthijs Douze *(Meta)*.

The rapid growth of transformer-based models increases the concerns about their integrity and ownership insurance. Watermarking addresses this issue by embedding a unique identifier into the model, while preserving its performance. However, most existing approaches require to optimize the weights to imprint the watermark signal, which is not suitable at scale due to the computational cost. This paper explores watermarks with virtually no computational cost [19]. It is applicable to a non-blind white-box setting (assuming access to both the original and watermarked networks). They generate functionally equivalent copies by leveraging the models' invariance, via operations like dimension permutations or scaling/unscaling. This enables to watermark models without any change in their outputs and remains stealthy. Experiments demonstrate the effectiveness of the approach and its robustness against various model transformations (fine-tuning, quantization, pruning), making it a practical solution to protect the integrity of large models.

8.1.4 Recherche de relation à partir d'un seul exemple fondée sur un modèle N-way K-shot : une histoire de distracteurs

Participants: Hugo Thomas, Guillaume Gravier, Pascale Sébillot.

La recherche de relation à partir d'un exemple consiste à trouver dans un corpus toutes les occurrences d'un type de relation liant deux entités dans une phrase, nommé type cible et caractérisé à l'aide d'un seul exemple. Nous empruntons le scénario d'entraînement et évaluation N-way K-shot à la tâche de classification de relations rares qui prédit le type de relation liant deux entités à partir de peu d'exemples

d'entraînement, et l'adaptons à la recherche de relation avec un exemple. Lors de l'évaluation, un modèle entraîné pour la classification de relations en N-way K-shot est utilisé, dans lequel K vaut un pour le type cible, une des N classes (du N-way) représente le type cible, et les N-1 classes restantes sont des distracteurs modélisant la classe de rejet. Les résultats sur FewRel et TACREV démontrent l'efficacité de notre approche malgré la difficulté de la tâche. L'étude de l'évolution des performances en fonction du nombre de distracteurs et des stratégies de leur choix met en avant une bonne configuration globale, à savoir un nombre élevé de distracteurs à une distance intermédiaire du type de relation cible dans l'espace latent appris par le modèle. Le diagnostic a posteriori de notre méthode révèle l'existence de configurations optimales pour chaque type cible que nos analyses actuelles échouent à caractériser, ouvrant la voie à de futurs travaux [32].

8.1.5 One-shot relation retrieval in news archives: adapting N-way K-shot relation classification for efficient knowledge extraction

Participants: Hugo Thomas, Guillaume Gravier, Pascale Sébillot.

One-shot relation retrieval is the knowledge extraction task that consists in searching in a textual dataset for all occurrences of a relation of interest, named the source relation, characterized by a single example–a relation being a link between a pair of entities in an utterance. Performing this task on large datasets requires an intelligent system to automate the process, for instance when exploring news archives for press review or business intelligence. We propose a framework that leverages the representation learning capabilities of N-way K-shot models for few-shot relation classification and extends these models to enable one-shot retrieval with a rejection class [28]. At evaluation time, one-shot relation retrieval is performed in a N-way K-shot setting where 1 of the N ways (or relations) is the source relation and the N-1 others are distractors, i.e., relations modeling a rejection class. We benchmark this framework and investigate the influence of the number and the choice of distractors on the standard TACREV and FewRel datasets. Experimental results demonstrate the effectiveness of our approach to address this highly challenging task, however with high variability primarily induced by the type of the source relation. Experiments also highlight a sound strategy for the choice of distractors-a large number of distractors at an intermediate distance from the embedding of the source relation in the latent space learned by the model-, which provides a competing trade-off between recall and precision. This strategy is globally optimal but can however be surpassed on certain source relations by others, depending on the characteristics of the source relation, paving the way for future work. We finally show the substantial benefit of two-shot retrieval over one-shot retrieval, which sheds light on the design of actual intelligent applications leveraging one-or few-shot relation retrieval.

8.1.6 Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video

Participants: Shashanka Venkataramanan, Mamshad Rizve *(UFC)*, João Carreira *(Google DeepMind)*, Yuki Asano *(UvA)*, Yannis Avrithis *(IARAI)*.

Self-supervised learning has unlocked the potential of scaling up pretraining to billions of images, since annotation is unnecessary. But are we making the best use of data? How more economical can we be? In this work, we attempt to answer this question by making two contributions. First, we investigate first-person videos and introduce a "Walking Tours" dataset. These videos are high-resolution, hourslong, captured in a single uninterrupted take, depicting a large number of objects and actions with natural scene transitions. They are unlabeled and uncurated, thus realistic for self-supervision and comparable with human learning. Second, we introduce a novel self-supervised image pretraining method tailored for learning from continuous videos. Existing methods typically adapt image-based pretraining approaches to incorporate more frames. Instead, we advocate a "tracking to learn to recognize" approach. Our method called DORA, leads to attention maps that Discover and tRAck objects over time in an end-to-end manner, using transformer cross-attention [31]. We derive multiple views from the tracks and use them

in a classical self-supervised distillation loss. Using our novel approach, a single Walking Tours video remarkably becomes a strong competitor to ImageNet for several image and video downstream tasks.

8.1.7 AggNet: Learning to aggregate faces for group membership verification

Participants: Marzieh Gheisari *(IBENS)*, Javad Amirian *(ISIR)*, Teddy Furon, Laurent Amsaleg.

In certain applications of face recognition, our goal is to verify whether an individual belongs to a particular group while keeping their identity undisclosed. Existing methods have suggested a process of quantizing pre-computed face descriptors into discrete embeddings and aggregating them into a single representation for the group. However, this mechanism is only optimized for a given closed set of individuals and requires relearning the group representations from scratch whenever the groups change. In this paper, we introduce a deep architecture that simultaneously learns face descriptors and the aggregation mechanism to enhance overall performance [10]. Our system can be utilized for new groups comprising individuals who have never been encountered before, and it easily handles new memberships or the termination of existing memberships. Through experiments conducted on multiple extensive, real-world face datasets, we demonstrate that our proposed method achieves superior verification performance compared to other baseline approaches.

8.1.8 REStore: Exploring a Black-Box Defense against DNN Backdoors using Rare Event Simulation

Participants: Quentin Le Roux, Kassem Kallas, Teddy Furon.

Backdoor attacks pose a significant threat to deep neural networks as they allow an adversary to inject a malicious behavior in a victim model during training. This paper addresses the challenge of defending against backdoor attacks in a blackbox setting where the defender has a limited access to a suspicious model. In this paper, we introduce Importance Splitting, a Sequential Monte-Carlo method previously used in neural network robustness certification, as an off-the-shelf tool for defending against backdoors [25]. We demonstrate that a black-box defender can leverage rare event simulation to assess the presence of a backdoor, reconstruct its trigger, and finally purify test-time input data in real-time. So-called REStore, our input purification defense proves effective in black-box scenarios because it uses triggers recovered with a query access to a model (only observing its logit, probit, or top-1 label outputs). We test our method on MNIST, CIFAR-10, and CASIA-Webface. We believe we are the first to demonstrate that backdoors may be considered under the lens of rare event simulation. Moreover, REStore is the first one-stage black-box input purification defense that approaches the performance of more complex comparables. REStore avoids gradient estimation, model reconstruction, or the vulnerable training of additional models.

8.1.9 Proactive Detection of Voice Cloning with Localized Watermarking

Participants: Robin San Roman *(FAIR, MultiSpeech)*, Pierre Fernandez *(Meta)*, Hady Elsahar *(FAIR)*, Alexandre Défossez *(Kyutai)*, Teddy Furon, Tuan Tran *(FAIR)*.

In the rapidly evolving field of speech generative models, there is a pressing need to ensure audio authenticity against the risks of voice cloning. We present AudioSeal, the first audio watermarking technique designed specifically for localized detection of AI-generated speech [26]. AudioSeal employs a generator / detector architecture trained jointly with a localization loss to enable localized watermark detection up to the sample level, and a novel perceptual loss inspired by auditory masking, that enables AudioSeal to achieve better imperceptibility. AudioSeal achieves state-of-the-art performance in terms

of robustness to real life audio manipulations and imperceptibility based on automatic and human evaluation metrics. Additionally, AudioSeal is designed with a fast, single-pass detector, that significantly surpasses existing models in speed, achieving detection up to two orders of magnitude faster, making it ideal for large-scale and real-time applications. Code is available at audioseal.

8.1.10 A Fast and Sound Tagging Method for Discontinuous Named-Entity Recognition

Participant: Caio Corro.

We introduce a novel tagging scheme for discontinuous named entity recognition based on an explicit description of the inner structure of discontinuous mentions [16]. We rely on a weighted finite state automaton for both marginal and maximum a posteriori inference. As such, our method is sound in the sense that (1) well-formedness of predicted tag sequences is ensured via the automaton structure and (2) there is an unambiguous mapping between well-formed sequences of tags and (discontinuous) mentions. We evaluate our approach on three English datasets in the biomedical domain, and report comparable results to state-of-the-art while having a way simpler and faster model.

8.1.11 Few-Shot Domain Adaptation for Named-Entity Recognition via Joint Constrained k-Means and Subspace Selection

Participants: Ayoub Hammal *(STL LISN)*, Benno Uthayasooriyar *(LMBA, SCOR SE)*, Caio Corro.

Named-entity recognition (NER) is a task that typically requires large annotated datasets, which limits its applicability across domains with varying entity definitions. This paper addresses few-shot NER, aiming to transfer knowledge to new domains with minimal supervision [22]. Unlike previous approaches that rely solely on limited annotated data, we propose a weakly supervised algorithm that combines small labeled datasets with large amounts of unlabeled data. Our method extends the k-means algorithm with label supervision, cluster size constraints and domain-specific discriminative subspace selection. This unified framework achieves state-of-the-art results in few-shot NER on several English datasets.

8.1.12 Training LayoutLM from Scratch for Efficient Named-Entity Recognition in the Insurance Domain

Participants: Benno Uthayasooriyar *(LMBA, SCOR SE)*, Antoine Ly *(SCOR SE)*, Franck Vermet *(LMBA)*, Caio Corro.

Generic pre-trained neural networks may struggle to produce good results in specialized domains like finance and insurance. This is due to a domain mismatch between training data and downstream tasks, as in-domain data are often scarce due to privacy constraints. In this work, we compare different pre-training strategies for LAYOUTLM [30]. We show that using domain-relevant documents improves results on a named-entity recognition (NER) problem using a novel dataset of anonymized insurance-related financial documents called PAYSLIPS. Moreover, we show that we can achieve competitive results using a smaller and faster model.

8.1.13 WaterMax: breaking the LLM watermark detectability-robustness-quality trade-off

Participants: Eva Giboulot, Teddy Furon.

Watermarking is a technical means to dissuade malfeasant usage of Large Language Models. This paper proposes a novel watermarking scheme, so-called WaterMax, that enjoys high detectability while sustaining the quality of the generated text of the original LLM [21]. Its new design leaves the LLM untouched (no modification of the weights, logits, temperature, or sampling technique). WaterMax balances robustness and complexity contrary to the watermarking techniques of the literature inherently provoking a trade-off between quality and robustness. Its performance is both theoretically proven and experimentally validated. It outperforms all the SotA techniques under the most complete benchmark suite.

8.1.14 When does gradient estimation improve black-box adversarial attacks?

Participants: Enoal Gesny, Eva Giboulot, Teddy Furon.

The recent black-box adversarial attack SurFree demonstrated its high effectiveness resorting to a purely geometric construction. The method drastically reduced the number of queries necessary to craft low-distortion adversarial examples compared to the preceding art which relied on costly gradient estimation. Recently, CGBA proposed to reintroduce gradient information to SurFree. Despite promising empirical results, no theoretical study of the method was provided. This paper fills this gap by providing a comprehensive analysis of the performance of SurFree and CGBA [20]. Notably, we express conditions under which using the gradient information is guaranteed to improve upon SurFree performance. We also provide the theoretical distortion of each attack at a given iteration, demonstrating the convergence of CGBA to the optimal adversarial image. Finally, we study the optimal query allocation schedule for CGBA. The accompanying code is to be found at Use-of-gradient-for-black-box-attacks.

8.1.15 SWIFT: Semantic Watermarking for Image Forgery Thwarting

Participants: Gautier Evennou *(IMATAG)*, Vivien Chappelier *(IMATAG)*, Ewa Kijak, Teddy Furon.

This paper proposes a novel approach towards image authentication and tampering detection by using watermarking as a communication channel for semantic information [17]. We modify the HiDDeN deep-learning watermarking architecture to embed and extract high-dimensional real vectors representing image captions. Our method improves significantly robustness on both malign and benign edits. We also introduce a local confidence metric correlated with Message Recovery Rate, enhancing the method's practical applicability. This approach bridges the gap between traditional watermarking and passive forensic methods, offering a robust solution for image integrity verification. The code is available at swift_watermarking.

8.1.16 Distinctive Image Captioning: Leveraging ground truth captions in CLIP guided reinforcement learning

Participants: Antoine Chaffin *(IMATAG)*, Ewa Kijak, Vincent Claveau.

Training image captioning models using teacher forcing results in very generic samples, whereas more distinctive captions can be very useful in retrieval applications or to produce alternative texts describing images for accessibility. Reinforcement Learning (RL) allows to use cross-modal retrieval similarity score between the generated caption and the input image as reward to guide the training, leading to more distinctive captions. Recent studies show that pre-trained cross-modal retrieval models can be used to provide this reward, completely eliminating the need for reference captions. However, we argue in this paper that Ground Truth (GT) captions can still be useful in this RL framework. We propose a new image

captioning model training strategy that makes use of GT captions in different ways [15]. Firstly, they can be used to train a simple MLP discriminator that serves as a regularization to prevent reward hacking and ensures the fluency of generated captions, resulting in a textual GAN setup extended for multimodal inputs. Secondly, they can serve as strong baselines when added to the pool of captions used to compute the proposed contrastive reward to reduce the variance of gradient estimate. Thirdly, they can serve as additional trajectories in the RL strategy, resulting in a teacher forcing loss weighted by the similarity of the GT to the image. This objective acts as an additional learning signal grounded to the distribution of the GT captions. Experiments on MS COCO demonstrate the interest of the proposed training strategy to.

8.1.17 Fast Reliability Estimation for Neural Networks with Adversarial Attack-Driven Importance Sampling

Participants: Karim Tit *(uni.lu)*, Teddy Furon.

This paper introduces a novel approach to evaluate the reliability of Neural Networks (NNs) by integrating adversarial attacks with Importance Sampling (IS), enhancing the assessment's precision and efficiency [29]. Leveraging adversarial attacks to guide IS, our method efficiently identifies vulnerable input regions, offering a more directed alternative to traditional Monte Carlo methods. While comparing our approach with classical reliability techniques like FORM and SORM, and with classical rare event simulation methods such as Cross-Entropy IS, we acknowledge its reliance on the effectiveness of adversarial attacks and its inability to handle very high-dimensional data such as ImageNet. Despite these challenges, our comprehensive empirical validations on the datasets the MNIST and CIFAR10 demonstrate the method's capability to accurately estimate NN reliability for a variety of models. Our research not only presents an innovative strategy for reliability assessment in NNs but also sets the stage for further work exploiting the connection between adversarial robustness and the field of statistical reliability engineering.

8.1.18 Watermarking Makes Language Models Radioactive

Participants: Tom Sander *(Meta AI Research, X)*, Pierre Fernandez *(Meta AI Research)*, Alain Durmus *(Centre Borelli)*, Matthijs Douze *(Meta AI Research)*, Teddy Furon.

We investigate the radioactivity of text generated by large language models (LLM), i.e., whether it is possible to detect that such synthetic input was used to train a subsequent LLM. Current methods like membership inference or active IP protection either work only in settings where the suspected text is known or do not provide reliable statistical guarantees. We discover that, on the contrary, it is possible to reliably determine if a language model was trained on synthetic data if that data is output by a watermarked LLM. Our new methods, specialized for radioactivity, detects with a provable confidence weak residuals of the watermark signal in the fine-tuned LLM [27]. We link the radioactivity contamination level to the following properties: the watermark robustness, its proportion in the training set, and the fine-tuning process. For instance, if the suspect model is open-weight, we demonstrate that training on watermarked instructions can be detected with high confidence (p-value $< 10^{-5}$) even when as little as 5% of training text is watermarked.

8.1.19 A Double-Edged Sword: The Power of Two in Defending Against DNN Backdoor Attacks

Participants: Quentin Le Roux *(THALES)*, Kassem Kallas, Teddy Furon.

Backdoor attacks on deep neural networks work by injecting them with a malicious behavior during training. Such behavior can then be activated at test-time using cleverly-crafted triggers. Defending

against backdoors is key in machine learning security in order to safeguard the trust between model providers and users. This paper demonstrates the open problem of back-door defense performance against a representative selection of backdoor attacks, with a main focus on input purification (a valuable defense category in black-box contexts where all DNN inputs are preprocessed in the hope of erasing a potential trigger). We show that current defenses are adversary-aware and dataset-dependent. They typically focus on patch-based attacks and simpler image classification datasets. This brittleness when using stand-alone defenses highlights the cat-and-mouse game currently affecting the backdoor literature. In this context, we propose a two-defense strategy using existing methods as a palliative solution while waiting for future developments [24].

8.1.20 A Comprehensive Survey on Backdoor Attacks and Their Defenses in Face Recognition Systems

Participants: Quentin Le Roux *(THALES)*, Eric Bourbao *(THALES)*, Yannick Teglia, Kassem Kallas.

Deep learning has significantly transformed face recognition, enabling the deployment of large-scale, state-of-the-art solutions worldwide. However, the widespread adoption of deep neural networks (DNNs) and the rise of Machine Learning as a Service emphasize the need for secure DNNs. This paper revisits the face recognition threat model in the context of DNN ubiquity and the common practice of outsourcing their training and hosting to third-parties. Here, we identify backdoor attacks as a significant threat to modern DNN-based face recognition systems (FRS). Backdoor attacks involve an attacker manipulating a DNN's training or deployment, injecting it with a stealthy and malicious behavior. Once the DNN has entered its inference stage, the attacker may activate the backdoor and compromise the DNN's intended functionality. Given the critical nature of this threat to DNN-based FRS, our paper comprehensively surveys the literature of backdoor attacks and defenses previously demonstrated on FRS DNNs [13]. As a last point, we highlight potential vulnerabilities and unexplored areas in FRS security.

8.1.21 Beyond Internet Images: Evaluating Vision-Language Models for Domain Generalization on Synthetic-to-Real Industrial Datasets

Participants: Louis Hemadou, Helena Vorobieva *(Safran Tech)*, Ewa Kijak, Frédéric Jurie *(GREYC, UNICAEN)*.

Vision Language Foundation Models (VLFMs) have shown impressive generalization capabilities, making them suitable for Domain Generalization (DG) tasks, such as training on synthetic images and testing on real data. However, existing evaluations predominantly use academic benchmarks constructed from internet images, akin to the datasets used for training VLFMs. This paper assesses the performance of VLFM-based DG algorithms on two synthetic-to-real classification datasets, Rareplanes-tiles and Aerial Vehicles, designed to emulate industrial contexts [33]. Our findings reveal that while VLFMs excel on academic benchmarks, outperforming randomly initialized networks, their advantage is significantly diminished on these industrial-like datasets. This study underscores the importance of evaluating models on diverse, representative data to understand their real-world applicability and limitations.

8.1.22 HYBRINFOX at CheckThat! 2024 - Task 2: Enriching BERT Models with the Expert System VAGO for Subjectivity Detection

Participants: Morgane Casanova, Julien Chanson *(Mondeca)*, Benjamin Icard *(DEC, ENS-PSL, PSL, SMA)*, Géraud Faye *(MICS, Airbus Defence and Space)*, Guillaume Gadek *(Airbus Defence and Space)*, Guillaume Gravier, Paul Égré *(IJN, ENS-PSL)*.

This paper presents the HYBRINFOX method used to solve Task 2 of Subjectivity detection of the CLEF 2024 CheckThat! competition [14]. The specificity of the method is to use a hybrid system, combining a RoBERTa model, fine-tuned for subjectivity detection, a frozen sentence-BERT (sBERT) model to capture semantics, and several scores calculated by the English version of the expert system VAGO, developed independently of this task to measure vagueness and subjectivity in texts based on the lexicon. In English, the HYBRINFOX method ranked 1st with a macro F1 score of 0.7442 on the evaluation data. For the other languages, the method used a translation step into English, producing more mixed results (ranking 1st in Multilingual and 2nd in Italian over the baseline, but under the baseline in Bulgarian, German, and Arabic). We explain the principles of our hybrid approach, and outline ways in which the method could be improved for other languages besides English.

8.2 Accessing Information

8.2.1 A Multi-Label Dataset of French Fake News: Human and Machine Insights

Participants: Benjamin Icard *(DEC, ENS-PSL, PSL, SMA)*, François Maine *(SMA, Freedom Partners)*, Morgane Casanova, Géraud Faye *(MICS, Airbus Defence and Space)*, Julien Chanson *(Mondeca)*, Guillaume Gadek *(Airbus Defence and Space)*, Ghislain Atemezing *(Mondeca)*, François Bancilhon *(Observatoire des Médias)*, Paul Égré *(IJN, ENS-PSL)*.

We present in [23] a corpus of 100 documents, OBSINFOX, selected from 17 sources of French press considered unreliable by expert agencies, annotated using 11 labels by 8 annotators. By collecting more labels than usual, by more annotators than is typically done, we can identify features that humans consider as characteristic of fake news, and compare them to the predictions of automated classifiers. We present a topic and genre analysis using Gate Cloud, indicative of the prevalence of satire-like text in the corpus. We then use the subjectivity analyzer VAGO, and a neural version of it, to clarify the link between ascriptions of the label Subjective and ascriptions of the label Fake News. The annotated dataset is available online at the following url: OBSINFOX

8.2.2 Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification

Participants: Géraud Faye *(MICS, Airbus Defence and Space)*, Benjamin Icard *(DEC, ENS-PSL, PSL, SMA)*, Morgane Casanova, Julien Chanson *(Mondeca)*, François Maine *(SMA, Freedom Partners)*, François Bancilhon *(Observatoire des Médias)*, Guillaume Gadek *(Airbus Defence and Space)*, Guillaume Gravier, Paul Égré *(IJN, ENS-PSL)*.

This paper investigates the language of propaganda and its stylistic features [18]. It presents the PPN dataset, standing for Propagandist Pseudo-News, a multisource, multilingual, multimodal dataset composed of news articles extracted from websites identified as propaganda sources by expert agencies. A limited sample from this set was randomly mixed with papers from the regular French press, and their URL masked, to conduct an annotation-experiment by humans, using 11 distinct labels. The results show that human annotators were able to reliably discriminate between the two types of press across each of the labels. We propose different NLP techniques to identify the cues used by the annotators, and to compare them with machine classification. They include the analyzer VAGO to measure discourse vagueness and subjectivity, a TF-IDF to serve as a baseline, and four different classifiers: two RoBERTa-based models, CATS using syntax, and one XGBoost combining syntactic and semantic features.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

CIFRE PhD: Certification of Deep Neural Networks

Participants: Teddy Furon, Quentin Le Roux.

Duration: 3 years, started in November 2022
Partner:THALES

This is a CIFRE PhD thesis project aiming at assessing the security of already trained Deep Neural Networks, especially in the context of face recognition.

CIFRE PhD: Watermarking and deep learning

Participants: Teddy Furon, Pierre Fernandez.

Duration: 3 years, started in May 2022
Partner: META AI

This is a CIFRE PhD thesis project aiming at watermarking deep learning models analyzing or generating images or at using deep learning to watermark images.

CIFRE PhD: Domain generalization exploiting synthetic data

Participants: Ewa Kijak, Louis Hemadou.

Duration: 3 years, started in Nov. 2022
Partner: SAFRAN

This is a CIFRE PhD thesis project aiming at exploiting synthetic data to be able to perform transfer learning in presence of very few or inexistent real data in the context of image detection or classification tasks.

CIFRE PhD: Detection and explanation of semantic manipulations in multimedia content

Participants: Ewa Kijak, Gautier Evennou.

Duration: 3 years, started in Sep. 2023
Partner: IMATAG

This is a CIFRE PhD thesis project aiming at detecting and explaining semantic manipulations in multimedia content, in the context of misinformation.

CIFRE PhD: Machine learning for identification of factors impacting the quality of service of urban buses

Participants: Simon Malinowski, Guillaume Gravier, Erwan Vincent.

Duration: 3 years, started in Feb. 2022
Partner: KEOLIS

This is a CIFRE PhD thesis project aiming at identifying factors that have an impact on the quality of service of urban buses, and at predicting inter-arrival times in order to better understand the urban bus network.

CIFRE PhD: Introduction of rejection capabilities and externalized language models in deep learning systems for text reading under adverse conditions

Participants: Guillaume Gravier.

Duration: 3 years, started in June 2023
Partner: ANTAI

The thesis, in conjunction with the team SHADOC at IRISA, studies deep models for license plate recognition capable of balancing end-to-end training with separate language model training and adaptation.

Telegramme-CNRS bilateral contract: NLP for computational journalism

Participants: Laurent Amsaleg, Pascale Sébillot, Christian Raymond *(Insa Rennes)*, Nicolas Fouqué.

Duration: 2 years, started in Jan 2022

The project aims at developing a wide range of text-mining and classification tools with the French press group Le Télégramme. In particular, we aim at discovering cues of success in the already published news articles and then exploit them to propose new angles of coverage of newsworthy events to the journalists.

DGA-Inria collaboration

Participants: Teddy Furon, Charly Faure *(DGA-MI)*, Virgile Dine.

Duration: 3 years, started in Oct. 2024

The project aims at developing algorithms to make computer unlearn. From a model trained over a training dataset, we aim at deriving a second model ignoring some training samples, or some classes of samples without retraining it from scratch.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Inria associate team not involved in an IIL or an international program

LOGIC

Title: Learning on graph-based hierarchical methods for image and multimedia data

Duration: 2020 ->

Coordinator: Silvio Jamil Guimaraes (sjamil@pucminas.br)

Partners:

- Pontifícia Universidade Católica de Minas Gerais Belo Horizonte (Brésil)

Inria contact: Simon Malinowski

Summary: The main goal of this project is related to learning graph-based hierarchical methods to be applied on image and multimedia data. Regarding image data, we aim at advancing in the state-of-the-art on hierarchy of partitions taking into account aspects of efficiency, quality, and interactivity, as well as the use of hierarchical information to help the information extraction process. Research on graph-based multimedia label/information propagation will be developed within this project along two main lines of research : - construction of multimedia graphs where links should depict semantic proximity between documents or fragments of documents - how different graph structures can be used to propagate information (usually tags or labels) from one document to another and across modalities

10.1.2 STIC/MATH/CLIMAT AmSud projects

GIMMD

Title: Graph-based analysis and understanding of image, video and multimedia data

Program: STIC-AmSud

Duration: January 2, 2024 – December 31, 2025

Local supervisor: Simon Malinowski

Partners:

- Guimarães (Brésil)
- Randall (Uruguay)

Inria contact: Simon Malinowski

Summary: Graphs can be seen as a way of representing relationships between elements, which can be pixels in image analysis, voxels in video analysis, people in contact networks, or even weather stations for data capture. Understanding the relationships between elements, called vertices, as well as identifying groups of elements that have similar characteristics make the use of graphs a powerful tool to solve real problems through their representation (or modeling) in graphs. Still, methods of analyzing images and videos, and even social networks, which use hierarchical representations, aim to explore the visual representation as a space-scale oriented by regions, that is, a set of representations based, for example, on graphs, with different levels of detail, in which representation at finer levels are nested to obtain coarser levels, thus producing a hierarchy of partitions. This type of data structure has been successfully applied in medical imaging, object detection and video captioning, as well as community identification in social networks. Despite the various approaches to computing partition hierarchies, developing efficient and effective methods is not an easy task, due to the semantic information needed to perform the segmentation. In fact, the state-of-the-art in graph partitioning methods are highly dependent on using good gradients, when there is differentiability between elements, to produce good results. Models based on optimal paths in trees represent an excellent direction to consider any problems produced by hierarchies, since any errors in the delineation of the borders of the regions can be corrected. These methods can eventually be transformed, without loss of quality, into hierarchical methods, incorporating new properties thanks to the use of hierarchy. In addition, with the advances of deep learning, it becomes essential to explore semantic relationships through graphs for the annotation of pseudo labels in order to train deep neural networks in addition to estimating saliences through networks to assist in the graphbased segmentation. The main objective of this study is both to advance the state of the art in partition hierarchy, considering aspects of efficiency, quality, hierarchical transformations and interactivity, as well as to explore the relationships of graphs and neural networks in image/video applications like inpainting, video captioning, for instances. Finally, we will explore

methods of semi-supervised segmentation through the (semi) automatic location of markers. The results of these studies will be used to resolve various applications such as identi cation of cancer-susceptible cells in medical images, labeling regions in images and videos, identifying superpixels and supervoxels, inpainting, predicting solar irradiation in regions of interest, among others. We will build upon existing research and skills at LIGM, IRISA, UNICAMP, PUC Minas and UDELAR to develop collaborative work exploiting complementarity of these institutions.

10.2 National initiatives

Chaire Security of AI for Defense Applications (SAIDA)

Participants: Teddy Furon, Laurent Amsaleg, Mathias Rousset *(SIMSMART)*, Quentin Le Roux, Karim Tit.

Duration: 4 years, started Sept 2020
ANR-20-CHIA-0011-01

SAIDA targets the AID "Fiabilité de l'intelligence artificielle, vulnérabilités et contre-mesures" chair. It aims at establishing the fundamental principles for designing reliable and secure AI systems: a reliable AI maintains its good performance even under uncertainties; a secure AI resists attacks in hostile environments. Reliability and security are challenged at training and at test time. SAIDA therefore studies core issues in relation with poisoning training data, stealing the parameters of the model or inferring sensitive training from information leaks. Additionally, SAIDA targets uncovering the fundamentals of attacks and defenses engaging AI at test time. Three converging research directions make SAIDA: 1) theoretical investigations grounded in statistics and applied mathematics to discover the underpinnings of reliability and security, 2) connects adversarial sampling and Information Forensics and Security, 3) protecting the training data and the AI system. SAIDA thus combines theoretical investigations with more applied and heuristic studies to guarantee the applicability of the findings as well as the ability to cope with real world settings.

ANR MEERQAT: MultimEdia Entity Representation and Question Answering Tasks

Participants: Laurent Amsaleg, Yannis Avrithis, Ewa Kijak, Shashanka Venkataramanan.

Duration: 3.5 year, started in April 2020
Partners: Inria project-teams Linkmedia, CEA LIST, LIMSI, IRIT.

The overall goal of the project is to tackle the problem of ambiguities of visual and textual content by learning then combining their representations. As a final use case, we propose to solve a Multimedia Question Answering task, that requires to rely on three different sources of information to answer a (textual) question with regard to visual data as well as an external knowledge base containing millions of unique entities, each being represetd by textual and visual content as well as some links to other entities. An important work will deal with the representation of entities into a common tri-modal space, in which one should determine the content to associate to an entity to adequately represent it. The challenge consists in defining a representation that is compact (for performance) while still expressive enough to reflect the potential links between the entity and a variety of others.

MinArm: EVE4

Participants: Teddy Furon, Eva Giboulot.

Duration: 3 year, started in April 2022
Partners: MinArm, CRIStAL Lille, LIRMM, Univ. Troyes, Univ. Paris Saclay

Teaching and technology survey on steganography and steganalysis in the real world.

ASTRID: HybrInfox

Participants: Vincent Claveau, Guillaume Gravier, Morgane Casanova.

Duration: 20 months, started Jan. 2022

This ANR-AID funded project aims at building exploring how hybridation of symbolic and deep learning NLP tools. These hybrid tools are expected to be used to detect some types of disinformation; in particular, these NLP tools target vagueness (non precise) or subjective (opinion rather than factual) discourses.

Labcom SYNAPSES

Participants: Laurent Amsaleg, Guillaume Gravier, Pascale Sébillot, Michel Le Nouy *(Ouest-France)*, Morgane Casanova.

Duration: 54 months, started Jan. 2024

In spring 2024, the French ANR accepted to financially support the SYNAPSES Laboratoire commun with *Ouest-France*. It is administratively managed by the CNRS. For 5 years, starting in spring 2024, we will work closely with *Ouest-France* on a rather applied research program with the goal to eventually transfer some technological solutions to their development teams. The support from ANR amounts will be used to hire two engineers who will prepare proof-of-concept prototypes demonstrating the power of DL technologies applied to a subset of their photo stock and of their news archives. CIFRE PhDs as well as PhDs funded by academia will be enrolled to explore open issues. Note that the consortium agreement signed for SYNAPSES includes chapters clarifying the intellectual property and PGDR issues.

ANR AGAPE

Participants: Laurent Amsaleg, Guillaume Gravier, Pascale Sébillot.

Duration: 48 months, started Jan. 2025

That ANR (ANR-24-CE38-7253), accepted during the summer of 2024, is coordinated by the Lastig laboratory of the IGN. It includes Linkmedia, Ilda from INRIA, the LIRIS, the National Archives, France TV and the University G. Eiffel. AGAPE aims to aggregate and process multimedia content related to cultural and natural heritage, leveraging open data policies and the vast information available online. The project focuses on visual-based documents, such as images, videos, 3D point clouds, and text descriptions. Its first goal is to conduct innovative research on multimodal analysis to link and structure this diverse content. The second objective is to integrate the structured data into a 3D environment, offering new ways of visualizing, navigating, and interacting with it. AGAPE seeks to create an open-source, interoperable, and reproducible framework encapsulated in a digital twin dedicated to heritage. This framework will be validated and applied in various fields, supporting archivists in enriching collections, historians in studying substandard housing, and journalists in engaging the public through media. A Ph.D. thesis for Linkmedia will be funded by AGAPE.

PEPR Cybersecurity COMPROMIS project

Participants: Teddy Furon, Eva Giboulot, Ewa Kijak, Enoal Gesny.

Duration: 4.5 years, started Apr. 2024

The COMPROMIS project is based on a modern vision of multimedia data protection, with deep learning at its heart. This project defends the idea that the protection of multimedia data must necessarily be associated with the security of the tools that analyse this data, i.e. these days Artificial Intelligence (AI). The observation is simple: the protection of multimedia data is undoubtedly the area of cybersecurity that has benefited most from AI, but it has neglected to check the level of security of this new tool. AI has become one of the weak links in the protection of multimedia data. The scientific hurdles thus concern both the classic applications of multimedia data protection and the emerging field of deep learning.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- Laurent Amsaleg was the general co-chair of CBMI 2024
- Teddy Furon was the general chair of ESSAI 2024, European Symposium on Security of Artificial Intelligence
- Ewa Kijak was technical program chair for CBMI 2024, the 21st International Conference on Content-based Multimedia Indexing

11.1.2 Scientific events: selection

Member of the conference program committees

- Caio Corro was an area chair for EMNLP 2024
- Laurent Amsaleg was a senior area chair for ACM Multimedia 2024
- Laurent Amsaleg was a PC member of ICMR, ICME, MMM, SISAP, CBMI
- Pascale Sébillot was a PC member for Conférence nationale en intelligence artificielle CNIA 2024

Reviewer

- Caio Corro was a reviewer for Coling 2025
- Teddy Furon was a reviewer for CVPR 2025, ICLR 2025, ICML 2025, NeurIPS 2024
- Pascale Sébillot was a reviewer for LREC-Coling 2024
- Eva Giboulot was a reviewer for IEEE WIFS 2024, ACM AiSec 2024, IEEE ICASSP 2025, ACM IH&MMSec 2024

11.1.3 Journal

Reviewer - reviewing activities

- Teddy Furon was a reviewer for IEEE Transactions on Information Forensics and Security, IEEE Transactions on Dependable and Secure Computing, Transactions on Machine Learning Research
- Eva Giboulot was a reviewer for IEEE Transactions on Information Forensics and Security

11.1.4 Invited talks

- Teddy Furon was an invited speaker at 'Trustworthy Machine Learning' (Sorbonne Center for AI), Salon VivaTech Paris, and Atelier Inria - UK AI Safety Institute, Summer school 'Cyber in Normandy', Winter school of the CyberSchool
- Ewa Kijak gave an invited talk for the scientific seminar of the ENS Rennes computer science department
- Ewa Kijak was an invited speaker for the French Ministry of Defence's scientific and strategic day on LLMs and generative AIs
- Caio Corro gave an invited talk at INRIA Paris: "Named-Entity Recognition: Resurrecting Old School Machine Learning in the Era of Deep Learning", Dec. 2024

11.1.5 Leadership within the scientific community

- Guillaume Gravier is a member of the scientific board of the GDR Traitement automatique des langues
- Pascale Sébillot is a member of the board of the GDR Traitement automatique des langues

11.1.6 Scientific expertise

- Caio Corro reviewed grants for UTTER's Financial Support for Third Parties call (collaborative Research and Innovation project funded under Horizon Europe)
- Teddy Furon reviewed grant proposal for Region Normandy

11.1.7 Research administration

- Guillaume Gravier is director of IRISA (UMR 6074)
- Pascale Sébillot is deputy director of IRISA

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

Participants: Eva Giboulot, Ewa Kijak, Laurent Amsaleg, Guillaume Gravier, Pascale Sébillot.

- Master: Laurent Amsaleg, Bases de données avancées, 25h, M2, INSA Rennes, France
- Master: Eva Giboulot, Rare Event Simulations, 40h, INSA Rennes, France
- Licence: Guillaume Gravier, Natural language processing, 12h, L3, INSA Rennes
- Licence: Guillaume Gravier, Markov models, 6h, L3, INSA Rennes
- Master: Guillaume Gravier, Natural Language Processing, 6h, M1, INSA Rennes

- Master: Guillaume Gravier, Natural Language Processing, 51h, M2, ENSAI
- Master: Pascale Sébillot, Natural Language Processing, 4h, M1, INSA Rennes, France
- Master: Pascale Sébillot, Databases, 18h, M1, DIGISPORT graduate school (EUR), France
- Licence: Pascale Sébillot, Natural Language Processing, 6h, L3, INSA Rennes, France
- Licence: Caio Corro, Databases, 34h, L2, INSA Rennes, France
- Licence: Caio Corro, Probabilities, 26h, L3, INSA Rennes, France
- Ewa Kijak is head of the Image engineering track (M1-M2) of ESIR, Univ. Rennes
- Master: Ewa Kijak, Information retrieval and Multimodal applications, 24h, M2, ESIR
- Master: Ewa Kijak, Deep Learning for Vision, 12h, M2, ESIR
- Master: Ewa Kijak, Supervised machine learning, 20h, M1R, ENS Rennes
- Master: Ewa Kijak, Machine learning, 12h, M1, ESIR
- Master: Ewa Kijak, Image processing, 45h, M1, ESIR, Univ. Rennes

11.2.2 Supervision

- Ph.D. Duc Hau Nguyen, Making AI understandable for humans: the plausibility of attention-based mechanisms in natural language processing. Oct. 11, 2024. With Guillaume Gravier and Pascale Sébillot.
- Ph.D. Shashanka Venkataramanan, Metric learning for instance- and category-level visual representations. Jul. 1, 2024. With Yannis Avrithis and Ewa Kijak.
- Ph.D. Karim Tit, Reliability of Deep Learning with rare event simulation : theory and practice. Apr. 22, 2024. With Matthias Rousset and Teddy Furon
- PhD in progress: Gautier Evennou, Detection and explanation of semantic manipulations in multimedia content. Started in Sep. 2023, Ewa Kijak
- PhD in progress: Louis Hemadou, Domain generalization exploiting synthetic data. Started Nov. 2022, Ewa Kijak
- PhD in progress: Hugo Thomas, Zero-shot and few-shot relation extraction in press archives. Started Sept. 2022, Guillaume Gravier and Pascale Sébillot
- PhD in progress: Ahmed Abdourahman, AI-driven character simulation based on Multi-Agents Interaction Imitation Learning. Started Dec. 2023, Ewa Kijak and Franck Multon (MIMETIC Team at IRISA)
- PhD in progress: Adèle Denis, IA-based automated detection and behavior analysis among piglets. Started Sep. 2024, Ewa Kijak, Caroline Clouard (INRAE) and Céline Tallet (INRAE)
- PhD in progress: Virgile Dine, Machine Unlearning. Started Oct. 2024, Teddy Furon
- PhD in progress: Enoal Gesny, Watermarking of Generative AI. Started Nov. 2024, Eva Giboulot and Teddy Furon
- PhD in progress: Chloé Imadache, Security of Deep Learning based Watermarking. Started Dec. 2024, Eva Giboulot and Teddy Furon

11.2.3 Juries

- Laurent Amsaleg was a reviewer for the PhD. of Huiyu Li, Univ. Nice, Nov. 2024.
- Laurent Amsaleg was the president of the PhD jury of Tom Bachard, Univ. Rennes, Nov. 2024.
- Teddy Furon was the president of the PhD jury of Etienne Levecque, Univ. Lille, Nov. 2024.
- Ewa Kijak was a reviewer for the PhD. of Emile Blettery, Univ. Gustave Eiffel, Jan. 2024.
- Ewa Kijak was a jury member for the PhD. of Mireille El Assal, Univ. Lille, Fev. 2024.
- Ewa Kijak was a jury member for the PhD. of Guillaume Jeanneret, Univ. Caen, Sep. 2024.
- Ewa Kijak was a jury member for the PhD. of Tom Bachard, Univ. Rennes, Nov. 2024.
- Ewa Kijak was a jury member for the PhD. of Paul Berg, Univ. Bretagne Sud, Dec. 2024.
- Pascale Sébillot was a jury member for the PhD. of Kim-Anh Nguyen, Sorbonne-Univ., Apr. 2024
- Pascale Sébillot was a reviewer for the PhD. of Evan Dufraisse, Univ. Lorraine, Sept. 2024
- Pascale Sébillot was the president of the PhD. jury of Hui-Syuan Yeh, Univ. Paris-Saclay, Dec. 2024
- Caio Corro was a jury member for the PhD. of Nathan Godey, Inria, Dec. 2024.

11.3 Popularization

11.3.1 Productions (articles, videos, podcasts, serious games, ...)

- Teddy Furon was involved in the writing of the policy paper "Cybersecurity specific to AI" from the Inria Program Agency

11.3.2 Participation in Live events

- Laurent Amsaleg was involved into the "Chiche" program with 6 classes at the Lycée Saint joseph, Bruz.
- Laurent Amsaleg was invited at a panel during the "50 ans du Club de la presse de Bretagne"
- Teddy Furon was a speaker at 'Math and Art' Festival of Saint-Brieuc
- Teddy Furon was involved into the "Chiche" program with 6 classes at Lycée Lasalle (Verrières en Anjou) and Lycée Rabelais (Saint-Brieuc)

12 Scientific production

12.1 Major publications

[1] L. Amsaleg, J. Bailey, A. Barbe, S. Erfani, T. Furon, M. Houle, M. Radovanovic and N. X. Vinh. 'High Intrinsic Dimensionality Facilitates Adversarial Attack: Theoretical Evidence'. In: *IEEE Transactions on Information Forensics and Security* 16 (Sept. 2020), pp. 1–12. DOI: 10.1109/TIFS.2020.3023274. URL: https://hal.archives-ouvertes.fr/hal-02938099.

[2] B. Bonnet, T. Furon and P. Bas. 'Generating Adversarial Images in Quantized Domains'. In: *IEEE Transactions on Information Forensics and Security* (2022). DOI: 10.1109/TIFS.2021.3138616. URL: https://hal.archives-ouvertes.fr/hal-03467692.

[3] A. Chaffin, V. Claveau and E. Kijak. 'PPL-MCTS: Constrained Textual Generation Through Discriminator-Guided Decoding'. In: CtrlGen 2021 - Workshop on Controllable Generative Modeling in Language and Vision at NeurIPS 2021. Proceedings of the CtrlGen workshop. virtual, United States, 13th Dec. 2021, pp. 1–19. URL: https://hal.archives-ouvertes.fr/hal-03494695.

[4] P. Fernandez, A. Chaffin, K. Tit, V. Chappelier and T. Furon. 'Three bricks to consolidate watermarks for large language models'. In: *Proceedings of IEEE WIFS*. WIFS 2023 - IEEE International Workshop on Information Forensics and Security. Nuremberg, Germany: IEEE, Dec. 2023, pp. 1–9. URL: https://inria.hal.science/hal-04361015.

[5] P. Fernandez, G. Couairon, H. Jégou, M. Douze and T. Furon. 'The Stable Signature: Rooting Watermarks in Latent Diffusion Models'. In: *2023 IEEE International Conference on Computer Vision (ICCV)*. ICCV 2023 - International Conference on Computer Vision. 2023 IEEE International Conference on Computer Vision. Paris, France, Oct. 2023. URL: https://hal.science/hal-04176523.

[6] A. Iscen, G. Tolias, Y. Avrithis, T. Furon and O. Chum. 'Efficient Diffusion on Region Manifolds: Recovering Small Objects with Compact CNN Representations'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, United States, July 2017. URL: https://hal.inria.fr/hal-01505470.

[7] T. Maho, T. Furon and E. L. Merrer. 'SurFree: a fast surrogate-free black-box attack'. In: CVPR 2021 - Conference on Computer Vision and Pattern Recognition. Proc. of IEEE Conference on Computer Vision and Pattern Recognition, CVPR. Virtual, France, 19th June 2021, pp. 10430–10439. URL: https://hal.archives-ouvertes.fr/hal-03177639.

[8] S. Venkataramanan, E. Kijak, L. Amsaleg and Y. Avrithis. 'AlignMixup: Improving Representations By Interpolating Aligned Features'. In: CVPR 2022 - IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, United States: IEEE, June 2022, pp. 1–13. URL: https://hal.inria.fr/hal-03620779.

[9] V. Vukotić, C. Raymond and G. Gravier. 'A Crossmodal Approach to Multimodal Fusion in Video Hyperlinking'. In: *IEEE MultiMedia* 25.2 (2018), pp. 11–23. DOI: 10.1109/MMUL.2018.023121161. URL: https://hal.inria.fr/hal-01848539.

12.2 Publications of the year

International journals

[10] M. Gheisari, J. Amirian, T. Furon and L. Amsaleg. 'AggNet: Learning to aggregate faces for group membership verification'. In: *Signal Processing: Image Communication* 132 (7th Dec. 2024), p. 117237. DOI: 10.1016/j.image.2024.117237. URL: https://inria.hal.science/hal-04875732 (cit. on p. 15).

[11] C. Jerônimo, Z. Patrocínio, S. Malinowski, G. Gravier and S. J. Ferzoli Guimarães. 'Decreasing graph complexity with transitive reduction to improve temporal graph classification'. In: *International Journal of Data Science and Analytics* (2024). DOI: 10.1007/s41060-024-00632-8. URL: https://hal.science/hal-04770287 (cit. on p. 12).

[12] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin and P. Bojanowski. 'DINOv2: Learning Robust Visual Features without Supervision'. In: *Transactions on Machine Learning Research Journal* (2024). DOI: 10.48550/arxiv.2304.07193. URL: https://hal.science/hal-04376640 (cit. on p. 13).

[13] Q. L. Roux, E. Bourbao, Y. Teglia and K. Kallas. 'A Comprehensive Survey on Backdoor Attacks and Their Defenses in Face Recognition Systems'. In: *IEEE Access* 12 (2024), pp. 47433–47468. DOI: 10.1109/ACCESS.2024.3382584. URL: https://hal.science/hal-04850549 (cit. on p. 19).

International peer-reviewed conferences

[14] M. Casanova, J. Chanson, B. Icard, G. Faye, G. Gadek, G. Gravier and P. Égré. 'HYBRINFOX at CheckThat! 2024 - Task 2: Enriching BERT Models with the Expert System VAGO for Subjectivity Detection'. In: *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF 2024 CheckThat!)* CLEF 2024 - Conference and Labs of the Evaluation Forum. Grenoble, France, 2024, pp. 1–9. DOI: 10.48550/arXiv.2407.03770. URL: https://hal.science/hal-04871189 (cit. on p. 20).

[15] A. Chaffin, E. Kijak and V. Claveau. 'Distinctive image captioning: leveraging ground truth captions in clip guided reinforcement learning'. In: *Proceedings of 2024 IEEE International Conference on Image Processing*. ICIP 2024 - IEEE International Conference on Image Processing. Abu Dhabi, United Arab Emirates, 27th Oct. 2024. URL: https://hal.science/hal-04889761 (cit. on p. 18).

[16] C. Corro. 'A Fast and Sound Tagging Method for Discontinuous Named-Entity Recognition'. In: EMNLP 2024 - Conference on Empirical Methods in Natural Language Processing. Miami, United States: Association for Computational Linguistics, 12th Nov. 2024, pp. 19506–19518. DOI: 10.18653/v1/2024.emnlp-main.1087. URL: https://hal.science/hal-04870964 (cit. on p. 16).

[17] G. Evennou, V. Chappelier, E. Kijak and T. Furon. 'SWIFT: Semantic Watermarking for Image Forgery Thwarting'. In: *Proc. of IEEE WIFS*. WIFS 2024 - 16th IEEE International Workshop on Information Forensics and Security. Roma, Italy: IEEE, Dec. 2024, pp. 1–6. URL: https://hal.science/hal-04728070 (cit. on p. 17).

[18] G. Faye, B. Icard, M. Casanova, J. Chanson, F. Maine, F. Bancilhon, G. Gadek, G. Gravier and P. Égré. 'Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification'. In: *Proceedings of the EACL Workshop on Understanding Implicit and Underspecified Language (UnImplicit 2024)*. EACL Workshop on Understanding Implicit and Underspecified Language (UnImplicit 2024). Malte, Malta, 2024. URL: https://hal.science/hal-04443096 (cit. on p. 20).

[19] P. Fernandez, G. Couairon, T. Furon and M. Douze. 'Functional invariants to watermark large transformers'. In: *Proceedings of ICASSP'24*. ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, South Korea, Apr. 2024, pp. 1–5. URL: https://inria.hal.science/hal-04361026 (cit. on p. 13).

[20] E. Gesny, E. Giboulot and T. Furon. 'When does gradient estimation improve black-box adversarial attacks?' In: *Proceedings of IEEE WIFS 2024*. WIFS 2024 -16th IEEE International Workshop on Information Forensics and Security. Roma, Italy: IEEE, Dec. 2024, pp. 1–6. URL: https://hal.science/hal-04728275 (cit. on p. 17).

[21] E. Giboulot and T. Furon. 'WaterMax: breaking the LLM watermark detectability-robustness-quality trade-off'. In: *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. NeurIPS 2024 - 38th Conference on Neural Information Processing Systems. Vancouver, Canada, Dec. 2024, pp. 1–34. URL: https://hal.science/hal-04766606 (cit. on p. 17).

[22] A. Hammal, B. Uthayasooriyar and C. Corro. 'Few-Shot Domain Adaptation for Named-Entity Recognition via Joint Constrained k-Means and Subspace Selection'. In: Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025). Abu DHABI, France, 19th Jan. 2025. URL: https://hal.science/hal-04877776 (cit. on p. 16).

[23] B. Icard, F. Maine, M. Casanova, G. Faye, J. Chanson, G. Gadek, G. Atemezing, F. Bancilhon and P. Égré. 'A Multi-Label Dataset of French Fake News: Human and Machine Insights'. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. LREC-COLING 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. Torino, Italy: ELRA; ICCL, May 2024, pp. 812–818. URL: https://hal.science/hal-04521059 (cit. on p. 20).

[24] Q. Le Roux, K. Kallas and T. Furon. 'A Double-Edged Sword: The Power of Two in Defending Against DNN Backdoor Attacks'. In: EUSIPCO 2024 - 32nd IEEE European Signal Processing Conference. Lyon, France: IEEE, 2024, pp. 2007–2011. DOI: 10.23919/EUSIPCO63174.2024.10715340. URL: https://hal.science/hal-04850574 (cit. on p. 19).

[25] Q. Le Roux, K. Kallas and T. Furon. 'REStore: Exploring a Black-Box Defense against DNN Backdoors using Rare Event Simulation'. In: SaTML 2024 - 2nd IEEE Conference on Secure and Trustworthy Machine Learning. Vol. Proceedings of the 2nd IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). Toronto, Canada: IEEE, 2024, pp. 1–22. URL: https://hal.science/hal-04485197 (cit. on p. 15).

[26] R. S. Roman, P. Fernandez, H. Elsahar, A. Défossez, T. Furon and T. Tran. 'Proactive Detection of Voice Cloning with Localized Watermarking'. In: *Proceedings of the 41st International Conference on Machine Learning*. ICML 2024 - 41st International Conference on Machine Learning. Vol. 235. Vienna, Austria, July 2024, pp. 1–17. URL: https://hal.science/hal-04610152 (cit. on p. 15).

[27] T. Sander, P. Fernandez, A. Durmus, M. Douze and T. Furon. 'Watermarking Makes Language Models Radioactive'. In: *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. NeurIPS 2024 - 38th Conference on Neural Information Processing Systems. Vol. Spotlight. Vancouver, Canada, Dec. 2024, pp. 1–35. URL: https://hal.science/hal-04766621 (cit. on p. 18).

[28] H. Thomas, G. Gravier and P. Sébillot. 'One-shot relation retrieval in news archives: adapting N-way K-shot relation classification for efficient knowledge extraction'. In: KES 2024 - 28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. Seville, Spain, 2024, pp. 1060–1069. URL: https://hal.science/hal-04708239 (cit. on p. 14).

[29] K. Tit and T. Furon. 'Fast Reliability Estimation for Neural Networks with Adversarial Attack-Driven Importance Sampling'. In: UAI 2024 - 40th Conference on Uncertainty in Artificial Intelligence. Barcelona, Spain, 2024. URL: https://hal.science/hal-04565441 (cit. on p. 18).

[30] B. Uthayasooriyar, A. Ly, F. Vermet and C. Corro. 'Training LayoutLM from Scratch for Efficient Named-Entity Recognition in the Insurance Domain'. In: *Proceeedings of the COLING 2025 Workshop on Financial Technology and Natural Language Processing (FinNLP), Financial Narrative Processing (FNP), and on Large Language Models for Finance and Legal (LLMFinLegal)*. COLING 2025 Workshop on Financial Technology and Natural Language Processing (FinNLP), Financial Narrative Processing (FNP), and on Large Language Models for Finance and Legal (LLMFinLegal). Abu Dabi, United Arab Emirates, 19th Dec. 2024. URL: https://hal.science/hal-04877824 (cit. on p. 16).

[31] S. Venkataramanan, M. N. Rizve, J. Carreira, Y. Asano and Y. Avrithis. 'Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video'. In: ICLR 2024 - Twelfth International Conference on Learning Representations. Vienna, Austria, 2024, pp. 1–21. URL: https://inria.hal.science/hal-04407117 (cit. on p. 14).

National peer-reviewed Conferences

[32] H. Thomas, G. Gravier and P. Sébillot. 'Recherche de relation à partir d'un seul exemple fondée sur un modèle N-way K-shot : une histoire de distracteurs'. In: *35èmes Journées d'Études sur la Parole (JEP 2024)*. 35èmes Journées d'Études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024). Vol. 1 : articles longs et prises de position. Toulouse, France: ATALA & AFPC, 2024, pp. 157–168. URL: https://inria.hal.science/hal-04623015 (cit. on p. 14).

Conferences without proceedings

[33] L. Hémadou, H. Vorobieva, E. Kijak and F. Jurie. 'Beyond Internet Images: Evaluating Vision-Language Models for Domain Generalization on Synthetic-to-Real Industrial Datasets'. In: Synthetic Data for Computer Vision Workshop @ CVPR 2024. Seattle, Washington, United States, June 2024. URL: https://hal.science/hal-04889782 (cit. on p. 19).

Doctoral dissertations and habilitation theses

[34] D. Engin. 'Video question answering with limited supervision'. Université de Rennes, 11th June 2024. URL: https://theses.hal.science/tel-04694856.

[35] S. Venkataramanan. 'Metric learning for instance and category-level visual representation'. Université de Rennes, 1st July 2024. URL: https://theses.hal.science/tel-04711670.

Reports & preprints

[36] R. S. Roman, P. Fernandez, A. Deleforge, Y. Adi and R. Serizel. *Latent Watermarking of Audio Generative Models*. 2024. DOI: 10.48550/arXiv.2409.02915. URL: https://hal.science/hal-04716743.

12.3 Cited publications

[37] L. Amsaleg, J. E. Bailey, D. Barbe, S. Erfani, M. E. Houle, V. Nguyen and M. Radovanović. 'The Vulnerability of Learning to Adversarial Perturbation Increases with Intrinsic Dimensionality'. In: *WIFS*. 2017 (cit. on p. 7).

[38] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K.-I. Kawarabayashi and M. Nett. 'Estimating Local Intrinsic Dimensionality'. In: *KDD*. 2015 (cit. on pp. 6, 7, 9).

[39] L. Amsaleg, G. Þ. Guðmundsson, B. Þ. Jónsson and M. J. Franklin. 'Prototyping a Web-Scale Multimedia Retrieval Service Using Spark'. In: *ACM TOMCCAP* 14.3s (2018) (cit. on p. 9).

[40] L. Amsaleg, B. Þ. Jónsson and H. Lejsek. 'Scalability of the NV-tree: Three Experiments'. In: *SISAP*. 2018 (cit. on p. 9).

[41] R. Balu, T. Furon and L. Amsaleg. 'Sketching techniques for very large matrix factorization'. In: *ECIR*. 2016 (cit. on p. 9).

[42] S. Berrani, H. Boukadida and P. Gros. 'Constraint Satisfaction Programming for Video Summarization'. In: *ISM*. 2013 (cit. on p. 10).

[43] B. Biggio and F. Roli. 'Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning'. In: *Pattern Recognition* (2018) (cit. on p. 7).

[44] P. Bosilj. 'Image indexing and retrieval using component trees'. Theses. Université de Bretagne Sud, 2016 (cit. on p. 6).

[45] X. Bost. 'A storytelling machine? : Automatic video summarization: the case of TV series'. PhD thesis. University of Avignon, France, 2016 (cit. on p. 10).

[46] M. Budnik, M. Demirdelen and G. Gravier. 'A Study on Multimodal Video Hyperlinking with Visual Aggregation'. In: *ICME*. 2018 (cit. on p. 10).

[47] N. Carlini and D. A. Wagner. 'Audio Adversarial Examples: Targeted Attacks on Speech-to-Text'. In: *CoRR* abs/1801.01944 (2018). arXiv: 1801.01944 (cit. on p. 7).

[48] R. Carlini Sperandio, S. Malinowski, L. Amsaleg and R. Tavenard. 'Time Series Retrieval using DTW-Preserving Shapelets'. In: *SISAP*. 2018 (cit. on p. 6).

[49] V. Claveau, L. E. S. Oliveira, G. Bouzillé, M. Cuggia, C. M. Cabral Moro and N. Grabar. 'Numerical eligibility criteria in clinical protocols: annotation, automatic detection and interpretation'. In: *AIME*. 2017 (cit. on p. 6).

[50] A. Delvinioti, H. Jégou, L. Amsaleg and M. E. Houle. 'Image Retrieval with Reciprocal and shared Nearest Neighbors'. In: *VISAPP*. 2014 (cit. on p. 9).

[51] C. B. El Vaigh, F. Goasdoué, G. Gravier and P. Sébillot. 'Using Knowledge Base Semantics in Context-Aware Entity Linking'. In: *DocEng 2019 - 19th ACM Symposium on Document Engineering*. Berlin, Germany: ACM, Sept. 2019, pp. 1–10. DOI: 10.1007/978-3-030-27520-4_8. URL: https://hal.inria.fr/hal-02171981 (cit. on pp. 6, 8).

[52] H. Farid. *Photo Forensics*. The MIT Press, 2016 (cit. on p. 7).

[53] M. Gambhir and V. Gupta. 'Recent automatic text summarization techniques: a survey'. In: *Artif. Intell. Rev.* 47.1 (2017) (cit. on p. 10).

[54] I. Goodfellow, Y. Bengio and A. Courville. *Deep Learning*. MIT Press, 2016 (cit. on p. 5).

[55] G. Gravier, M. Ragot, L. Amsaleg, R. Bois, G. Jadi, E. Jamet, L. Monceaux and P. Sébillot. 'Shaping-Up Multimedia Analytics: Needs and Expectations of Media Professionals'. In: *MMM, Special Session Perspectives on Multimedia Analytics*. 2016 (cit. on p. 8).

[56] A. Iscen, L. Amsaleg and T. Furon. 'Scaling Group Testing Similarity Search'. In: *ICMR*. 2016 (cit. on p. 9).

[57] A. Iscen, G. Tolias, Y. Avrithis and O. Chum. 'Mining on Manifolds: Metric Learning without Labels'. In: *CVPR*. 2018 (cit. on pp. 5, 9).

[58] B. Þ. Jónsson, G. Tómasson, H. Sigurþórsson, Á. Eríksdóttir, L. Amsaleg and M. K. Larusdottir. 'A Multi-Dimensional Data Model for Personal Photo Browsing'. In: *MMM*. 2015 (cit. on p. 8).

[59] B. Þ. Jónsson, M. Worring, J. Zahálka, S. Rudinac and L. Amsaleg. 'Ten Research Questions for Scalable Multimedia Analytics'. In: *MMM, Special Session Perspectives on Multimedia Analytics*. 2016 (cit. on p. 8).

[60] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, N. Nießner, P. Pérez, C. Richardt, M. Zollhöfer and C. Theobalt. 'Deep Video Portraits'. In: *ACM TOG* (2018) (cit. on p. 7).

[61] M. Laroze, R. Dambreville, C. Friguet, E. Kijak and S. Lefèvre. 'Active Learning to Assist Annotation of Aerial Images in Environmental Surveys'. In: *CBMI*. 2018 (cit. on p. 6).

[62] S. Leroux, P. Molchanov, P. Simoens, B. Dhoedt, T. Breuel and J. Kautz. 'IamNN: Iterative and Adaptive Mobile Neural Network for Efficient Image Classification'. In: *CoRR* abs/1804.10123 (2018). arXiv: 1804.10123 (cit. on p. 6).

[63] A. Lods, S. Malinowski, R. Tavenard and L. Amsaleg. 'Learning DTW-Preserving Shapelets'. In: *IDA*. 2017 (cit. on p. 6).

[64] C. Maigrot, E. Kijak and V. Claveau. 'Context-Aware Forgery Localization in Social-Media Images: A Feature-Based Approach Evaluation'. In: *ICIP*. 2018 (cit. on p. 7).

[65] D. Shahaf and C. Guestrin. 'Connecting the dots between news articles'. In: *KDD*. 2010 (cit. on p. 10).

[66] M. Shi, H. Caesar and V. Ferrari. 'Weakly Supervised Object Localization Using Things and Stuff Transfer'. In: *ICCV*. 2017 (cit. on p. 5).

[67] R. Sicre, Y. Avrithis, E. Kijak and F. Jurie. 'Unsupervised part learning for visual recognition'. In: *CVPR*. 2017 (cit. on p. 6).

[68] R. Sicre and H. Jégou. 'Memory Vectors for Particular Object Retrieval with Multiple Queries'. In: *ICMR*. 2015 (cit. on p. 6).

[69] A. da Silva Pinto, D. Moreira, A. Bharati, J. Brogan, K. W. Bowyer, P. J. Flynn, W. J. Scheirer and A. Rocha. 'Provenance filtering for multimedia phylogeny'. In: *ICIP*. 2017 (cit. on p. 7).

[70] O. Siméoni, A. Iscen, G. Tolias, Y. Avrithis and O. Chum. 'Unsupervised Object Discovery for Instance Recognition'. In: *WACV*. 2018 (cit. on p. 6).

[71] H. O. Song, Y. Xiang, S. Jegelka and S. Savarese. 'Deep Metric Learning via Lifted Structured Feature Embedding'. In: *CVPR*. 2016 (cit. on p. 5).

[72] C. Tsai, M. L. Alexander, N. Okwara and J. R. Kender. 'Highly Efficient Multimedia Event Recounting from User Semantic Preferences'. In: *ICMR*. 2014 (cit. on p. 10).

[73] O. Vinyals, A. Toshev, S. Bengio and D. Erhan. 'Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge'. In: *TPAMI* 39.4 (2017) (cit. on p. 10).

[74] V. Vukotić. 'Deep Neural Architectures for Automatic Representation Learning from Multimedia Multimodal Data'. Theses. INSA de Rennes, 2017 (cit. on pp. 6, 8).

[75] V. Vukotić, C. Raymond and G. Gravier. 'Bidirectional Joint Representation Learning with Symmetrical Deep Neural Networks for Multimodal and Crossmodal Applications'. In: *ICMR*. 2016 (cit. on p. 6).

[76] V. Vukotić, C. Raymond and G. Gravier. 'Generative Adversarial Networks for Multimodal Representation Learning in Video Hyperlinking'. In: *ICMR*. 2017 (cit. on p. 6).

[77] J. Weston, S. Chopra and A. Bordes. 'Memory Networks'. In: *CoRR* abs/1410.3916 (2014). arXiv: 1410.3916 (cit. on p. 6).

[78] H. Yu, J. Wang, Z. Huang, Y. Yang and W. Xu. 'Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks'. In: *CVPR*. 2016 (cit. on p. 10).

[79] J. Zahálka and M. Worring. 'Towards interactive, intelligent, and integrated multimedia analytics'. In: *VAST*. 2014 (cit. on p. 8).

[80] L. Zhang, M. Shi and Q. Chen. 'Crowd Counting via Scale-Adaptive Convolutional Neural Network'. In: *WACV*. 2018 (cit. on p. 6).

[81] X. Zhang, X. Zhou, M. Lin and J. Sun. 'ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices'. In: *CoRR* abs/1707.01083 (2017). arXiv: 1707.01083 (cit. on p. 6).