

2025 Activity Report

RESEARCH CENTRE: Inria Centre at Rennes University

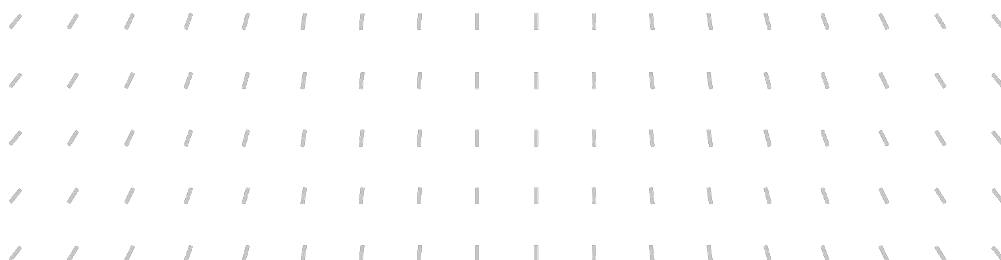
IN PARTNERSHIP WITH: Institut national des sciences appliquées de Rennes, CNRS,
Université de Rennes

Team

LINKMEDIA

Creating and exploiting explicit links between
multimedia fragments

In collaboration with Institut de recherche en informatique et systèmes aléatoires
(IRISA)



Team LINKMEDIA

Creation of the Team: 2014 July 01

Each year, Inria research teams publish an Activity Report presenting their work and results over the reporting period. These reports follow a common structure, with some optional sections depending on the specific team. They typically begin by outlining the overall objectives and research programme, including the main research themes, goals, and methodological approaches. They also describe the application domains targeted by the team, highlighting the scientific or societal contexts in which their work is situated. The reports then present the highlights of the year, covering major scientific achievements, software developments, or teaching contributions. When relevant, they include sections on software, platforms, and open data, detailing the tools developed and how they are shared. A substantial part is dedicated to new results, where scientific contributions are described in detail, often with subsections specifying participants and associated keywords. Finally, the Activity Report addresses funding, contracts, partnerships, and collaborations at various levels, from industrial agreements to international cooperations. It also covers dissemination and teaching activities, such as participation in scientific events, outreach, and supervision. The document concludes with a presentation of scientific production, including major publications and those produced during the year.

Keywords

Computer sciences and digital sciences

- A3.3.2. – Data mining
- A3.3.3. – Big data analysis
- A3.4. – Machine learning and statistics
- A4. – Security and privacy
- A5.3.3. – Pattern recognition
- A5.7. – Audio modeling and processing
- A5.7.1. – Sound
- A5.7.3. – Speech
- A5.8. – Natural language processing
- A9.2. – Machine learning
- A9.2.1. – Supervised learning
- A9.2.2. – Unsupervised learning
- A9.2.8. – Deep learning
- A9.3. – Signal processing
- A9.4. – Natural language processing
- A9.12.1. – Object recognition
- A9.12.3. – Content retrieval

Other research topics and application domains

- B9. – Society and Knowledge
- B9.3. – Medias
- B9.6.10. – Digital humanities
- B9.10. – Privacy

Contents

Team LINKMEDIA	1
1 Team members, visitors, external collaborators	5
2 Overall objectives	5
2.1 Context	5
2.2 Scientific objectives	6
3 Research program	6
3.1 Scientific background	6
3.2 Workplan	7
3.3 Research Direction 1: Extracting and Representing Information	7
3.4 Research Direction 2: Accessing Information	10
4 Application domains	13
4.1 Asset management in the entertainment business	13
4.2 Multimedia Internet	13
4.3 Data journalism	13
5 Social and environmental responsibility	13
5.1 Impact of research results	13
6 Highlights of the year	14
7 Latest software developments, platforms, open data	14
7.1 Latest software developments	14
7.1.1 MADHyS	14
8 New results	14
8.1 Extracting, Representing and Accessing Information	14
8.1.1 Revisiting Transferable Adversarial Images: Systemization, Evaluation, and New Insights	14
8.1.2 Bregman Conditional Random Fields: Sequence Labeling with Parallelizable Inference Algorithms	15
8.1.3 Few-Shot Domain Adaptation for Named-Entity Recognition via Joint Constrained k-Means and Subspace Selection	15
8.1.4 Training LayoutLM from Scratch for Efficient Named-Entity Recognition in the Insurance Domain	15
8.1.5 EuroBERT: Scaling Multilingual Encoders for European Languages	15
8.1.6 Relaxed syntax modeling in Transformers for future-proof license plate recognition	16
8.1.7 CroissantLLM: A Truly Bilingual French-English Language Model	16
8.1.8 Extraction of Contrastive Rules from Syntactic Treebanks: A Case Study in Romance Languages	17
8.1.9 Discrete Latent Structure in Neural Networks	17
8.1.10 Nested Named Entity Recognition as Single-Pass Sequence Labeling	18
9 Bilateral contracts and grants with industry	18
9.1 Bilateral contracts with industry	18
10 Partnerships and cooperations	19
10.1 International initiatives	19
10.2 National initiatives	20

11 Dissemination	20
11.1 Promoting scientific activities	21
11.1.1 Scientific events: organisation	21
11.1.2 Scientific events: selection	21
11.1.3 Journal	21
11.1.4 Research administration	21
11.2 Teaching - Supervision - Juries - Educational and pedagogical outreach	21
11.2.1 Teaching	21
11.2.2 Supervision	22
11.2.3 Juries	22
11.2.4 Specific official responsibilities in science outreach structures	22
11.2.5 Participation in Live events	22
12 Scientific production	22
12.1 Major publications	22
12.2 Publications of the year	23
12.3 Cited publications	24

1 Team members, visitors, external collaborators

Research Scientists

- Laurent Amsaleg [Team leader, CNRS, Senior Researcher, HDR]
- Guillaume Gravier [CNRS, Senior Researcher, HDR]

Faculty Members

- Caio Corro [INSA RENNES, Associate Professor]
- Simon Malinowski [UNIV RENNES, Associate Professor, until Feb 2025]
- Pascale Sébillot [INSA RENNES, Professor, HDR]

PhD Students

- Thomas Derrien [CNRS, from Oct 2025]
- Carolina Jeronimo De Almeida [GOUV BRESIL, until Feb 2025]
- Lilas Pastre [ENS RENNES, from Sep 2025]
- Hugo Thomas [INSA RENNES, ATER, from Oct 2025]
- Hugo Thomas [UNIV RENNES, until Sep 2025]

Technical Staff

- Jean-Rémi Bethys [CNRS, Engineer, from Jul 2025]
- Morgane Casanova [CNRS, Engineer]
- Nicolas Fouque [CNRS, Engineer]
- Anne-Charlotte Philippe [CNRS, Engineer, from Feb 2025 until Apr 2025]

Interns and Apprentices

- Rossana Cometa [INRIA, Intern, from Feb 2025 until Jul 2025]
- Thomas Derrien [INRIA, Intern, from Feb 2025 until Aug 2025]
- Amelie Knecht [UNIV RENNES, Apprentice, until Sep 2025]
- Lilas Pastre [CNRS, Intern, from Feb 2025 until Jul 2025]

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 Networks (RNN, such as LSTMs) or convolutional neural network (CNNs) for images that reach excellent performance [39]. 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 [42], or transferring what was learned on few very well annotated samples to allow the precise processing of poorly annotated data [51]. 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 [42] yet as powerful as ad-hoc strategies fitting specific tasks [56]. 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 [46]. 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 [62].
- *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 [66], 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 [47].
- *Multimodal embeddings.* We pursue pioneering work of LINKMEDIA on multimodal embedding, i.e., representing multiple modalities or information sources in a single embedded space [60, 59, 61]. Two main directions are explored: exploiting adversarial architectures (GANs) for embedding via translation from one modality to another, extending initial work in [61] to highly heterogeneous content; combining and constraining word and RDF graph embeddings to facilitate entity linking and explanation of lexical co-occurrences [36].
- *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 [55], Taylor decompositions to observe the individual contributions of each relevance scores or estimating LID [23] 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 [52, 53], relationships between the various objects that are represented in images [29], learning to localizing objects in images with only weak, image-level supervision [55] or fine-grained semantic information in texts [34]. 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 [65], 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 [48, 33] 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 [28]. 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 [32].

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 [22]. 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 [37], 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 [49].
- *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 [45]. 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 [54] 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 [23].

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 [59]. We also investigate the use of ontological knowledge to facilitate information extraction from content [36].
- *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 [43, 44, 40]).

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

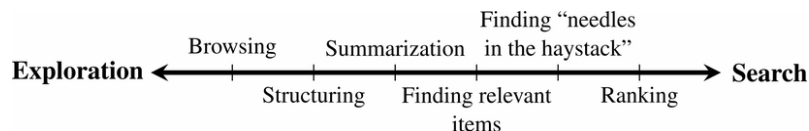


Figure 1: Exploration-search axis with example tasks

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 [64], 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 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 [42], time series retrieval, . . .). Furthermore, truly large datasets require involving sketching [26], secondary storage and/or distribution [25, 24], alleviating the explosion of the number of features to consider due to their local nature or other innovative methods [41], 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 [35], manifold structures instead of pure distance based similarities [42] must be extended to cope with advanced indexing in order to enhance quality. This includes feature selection based on intrinsic dimensionality estimation [23].
- *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 of 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 [63, 58] have been seen as a way to summarize visual content, opening the door to state-of-the-art multi-document text summarization [38] with text as a pivot modality. Automatic story-telling has been addressed for highly specific types of content, namely TV series [30] and news [50, 57], but still need a leap forward to be mostly automated, e.g., using constraint-based approaches for summarization [27, 57].

Furthermore, not only the original multimedia material has to be summarized, but the knowledge acquired from its analysis is also to be summarized. 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 [31].
- *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 2025 is marked by close collaboration with a major French media organization. The Linkmedia Ouest-France team is running 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 ends on December 31, 2025.

7 Latest software developments, platforms, open data

7.1 Latest software developments

7.1.1 MADHyS

Name: MULTI-LEVEL AGGREGATIONS FOR DYNAMIC HYPERGRAPHS STORYLINES

Keywords: Data Exploration, Data visualization

Functional Description: Visualization of large numbers of interdependent data sets, with relationships that evolve over time. Need to simultaneously visualize information at different scales, both detailed and broad, in order to understand a phenomenon in its entirety.

Contact: Nicolas Fouque

Participants: Laurent Amsaleg, Vanessa Pena Araya, Anastasia Bezerianos

8 New results

8.1 Extracting, Representing and Accessing Information

8.1.1 Revisiting Transferable Adversarial Images: Systemization, Evaluation, and New Insights

Participants: Zhengyu Zhao (*Xjtu - Xi'an Jiaotong University*), Hanwei Zhang (*Institute of Intelligent Software, Guangzhou – Saarland University, Saarbrücken*), Renjue Li (*School of Artificial Intelligence - Nanjing*), Ronan Sindre (*M2P2 - Laboratoire de Mécanique, Modélisation et Procédés Propres*), Laurent Amsaleg, Michael Backes (*CISPA - Helmholtz Center for Information Security, Saarbrücken*), Qi Li (*THU - Tsinghua University, Beijing*), Qian Wang (*Artificial Intelligence Institute of Wuhan University, Wuhan City*), Chao Shen (*Xjtu - Xi'an Jiaotong University*).

Transferable adversarial images raise critical security concerns for computer vision systems in real-world, blackbox attack scenarios. Although many transfer attacks have been proposed, existing research lacks a systematic and comprehensive evaluation. In this paper [12], we systemize transfer attacks into five categories around the general machine learning pipeline and provide the first comprehensive evaluation, with 23 representative attacks against 11 representative defenses, including the recent, transfer-oriented defense and the real-world Google Cloud Vision. In particular, we identify two main problems of existing evaluations: (1) for attack transferability, lack of intra-category analyses with fair hyperparameter settings, and (2) for attack stealthiness, lack of diverse measures. Our evaluation results validate that these problems have indeed caused misleading conclusions and missing points, and addressing them leads to new, consensuschallenging insights, such as (1) an early attack, DI, even outperforms all similar follow-up ones, (2) the state-of-the-art (whitebox) defense, DiffPure, is even vulnerable to (black-box) transfer attacks, and (3) even under the same L_p constraint, different attacks yield dramatically different stealthiness results regarding diverse imperceptibility metrics,

finer-grained measures, and a user study. We hope that our analyses will serve as guidance on properly evaluating transferable adversarial images and advance the design of attacks and defenses.

8.1.2 Bregman Conditional Random Fields: Sequence Labeling with Parallelizable Inference Algorithms

Participants: Caio Corro, Mathieu Lacroix (*LIPN - Laboratoire d'Informatique de Paris-Nord*), Joseph Le Roux (*LIPN - Laboratoire d'Informatique de Paris-Nord*).

We propose a novel discriminative model for sequence labeling called Bregman conditional random fields (BCRF). Contrary to standard linear-chain conditional random fields, BCRF allows fast parallelizable inference algorithms based on iterative Bregman projections. In this paper, we show how such models can be learned using Fenchel-Young losses, including extension for learning from partial labels [14]. Experimentally, our approach delivers comparable results to CRF while being faster, and achieves better results in highly constrained settings compared to mean field, another parallelizable alternative.

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

Participants: Ayoub Hammal (*STL - Sciences et Technologies des Langues - LISN*), Benno Uthayasooryar (*LMBA - Laboratoire de Mathématiques de Bretagne Atlantique, SCOR SE, Paris*), 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 [15]. 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 kmeans algorithm with label supervision, cluster size constraints and domain-specific discriminative subspace selection. This unified framework achieves state-of-the-art results in fewshot NER on several English datasets.

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

Participants: Benno Uthayasooryar (*LMBA - Laboratoire de Mathématiques de Bretagne Atlantique, SCOR SE, Paris*), Antoine Ly (*SCOR SE, Paris*), Franck Vermet (*LMBA - Laboratoire de Mathématiques de Bretagne Atlantique*), 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 [19]. 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 PAYSLLIPS. Moreover, we show that we can achieve competitive results using a smaller and faster model.

8.1.5 EuroBERT: Scaling Multilingual Encoders for European Languages

Participants: Nicolas Boizard (*MICS - Mathématiques et Informatique pour la Complexité et les Systèmes*), Hippolyte Gisserot-Boukhlef (*MICS - Mathématiques et Informatique pour la Complexité et les Systèmes*), Duarte M. Alves (*Instituto Superior Técnico*), André F T Martins (*Instituto Superior Técnico*), Ayoub Hammal (*Université Paris-Saclay*), Caio Corro, Céline Hudelot (*MICS - Mathématiques et Informatique pour la Complexité et les Systèmes*), Emmanuel Malherbe (*Artefact, Paris*), Etienne Malaboeuf (*CINES*), Fanny Jourdan (*IRT Saint Exupéry - Institut de Recherche Technologique*), Gabriel Hautreux (*CINES*), João Alves (*Unbabel*), Kevin El-Haddad (*ISIA - Institut Supérieur d'Informatique et d'Automatique*), Manuel Faysse (*Illuin Technology, Centrale Supélec*), Maxime Peyrard (*GETALP - Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole*), Nuno M Guerreiro (*MICS - Mathématiques et Informatique pour la Complexité et les Systèmes*), Patrick Fernandes (*Instituto Superior Técnico*), Ricardo Rei (*Unbabel*), Pierre Colombo (*MICS - Mathématiques et Informatique pour la Complexité et les Systèmes*).

General-purpose multilingual vector representations, used in retrieval, regression, and classification, are traditionally obtained from bidirectional encoder models. Despite their wide applicability, encoders have been recently overshadowed by advances in generative decoder-only models. However, many innovations driving this progress are not inherently tied to decoders. In this paper, we revisit the development of multilingual encoders through the lens of these advances, and introduce EuroBERT, a family of multilingual encoders covering European and widely spoken global languages [13]. Our models outperform existing alternatives across a diverse range of tasks, spanning multilingual capabilities, mathematics, and coding, and natively support sequences of up to 8,192 tokens. We also examine the design decisions behind EuroBERT, offering insights into our dataset composition and training pipeline. We publicly release the EuroBERT models, including intermediate training checkpoints, together with our training framework.

8.1.6 Relaxed syntax modeling in Transformers for future-proof license plate recognition

Participants: Florent Meyer (*ANTAI*), Laurent Guichard (*ANTAI*), Denis Coquenat (*SHADOC*), Guillaume Gravier, Yann Souillard (*SHADOC*), Bertrand Couasnon (*SHADOC*).

Effective license plate recognition systems are required to be resilient to constant change, as new license plates are released into traffic daily. While Transformer-based networks excel in their recognition at first sight, we observe significant performance drop over time which proves them unsuitable for tense production environments. Indeed, such systems obtain state-of-the-art results on plates whose syntax is seen during training. Yet, we show they perform similarly to random guessing on future plates where legible characters are wrongly recognized due to a shift in their syntax. After highlighting the flows of positional and contextual information in Transformer encoder-decoders, we identify several causes for their over-reliance on past syntax. Following, we devise architectural cut-offs and replacements which we integrate into SaLT, an attempt at a Syntax-Less Transformer for syntax-agnostic modeling of license plate representations. Experiments on both real and synthetic datasets show that our approach reaches top accuracy on past syntax and most importantly nearly maintains performance on future license plates. We further demonstrate the robustness of our architecture enhancements by way of various ablations [17].

8.1.7 CroissantLLM: A Truly Bilingual French-English Language Model

Participants: Manuel Faysse (*MICS - Mathématiques et Informatique pour la Complexité et les Systèmes*), Patrick Fernandes (*Instituto de Telecomunicações, Lisboa, Portugal*), Nuno M Guerreiro (*MICS - Mathématiques et Informatique pour la Complexité et les Systèmes*), Antonio Loison (*Illuin Technology*), Duarte M. Alves (*Instituto Superior Técnico*), Caio Corro, Nicolas Boizard (*MICS - Mathématiques et Informatique pour la Complexité et les Systèmes*), Ricardo Rei (*INESC-ID - Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa*), Pedro Raphaël Martins (*LTSI*), Antoni Casademunt (*Imperial College London*), François Yvon (*MLIA - Machine Learning and Information Access*), André Martins (*Instituto de Telecomunicações, Lisboa, Portugal*), Gautier Viaud (*Illuin Technology*), Céline Hudelot (*MICS - Mathématiques et Informatique pour la Complexité et les Systèmes*), Pierre Colombo (*MICS - Mathématiques et Informatique pour la Complexité et les Systèmes*).

We introduce CroissantLLM [10], a 1.3B language model pretrained on a set of 3T English and French tokens, to bring to the research and industrial community a high-performance, fully open-sourced bilingual model that runs swiftly on consumer-grade local hardware. To that end, we pioneer the approach of training an intrinsically bilingual model with a 1:1 English-to-French pretraining data ratio, a custom tokenizer, and bilingual finetuning datasets. We release the training dataset, notably containing a French split with manually curated, high-quality, and varied data sources. To assess performance outside of English, we craft a novel benchmark, FrenchBench, consisting of an array of classification and generation tasks, covering various orthogonal aspects of model performance in the French Language. Additionally, rooted in transparency and to foster further Large Language Model research, we release codebases, and dozens of checkpoints across various model sizes, training data distributions, and training steps, as well as fine-tuned Chat models, and strong translation models. We evaluate our model through the FMTI framework, and validate 81 % of the transparency criteria, far beyond the scores of even most open initiatives. This work enriches the NLP landscape, breaking away from previous English-centric work in order to strengthen our understanding of multilinguality in language models.

8.1.8 Extraction of Contrastive Rules from Syntactic Treebanks: A Case Study in Romance Languages

Participants: Santiago Herrera (*MoDyCo - Modèles, Dynamiques, Corpus*), Ioana-Madalina Silai (*Université Paris Nanterre - Département Sciences du Langage*), Caio Corro, Bruno Guillaume (*SEMAGRAMME*), Sylvain Kahane (*MoDyCo - Modèles, Dynamiques, Corpus*).

In this paper, we develop a data-driven contrastive framework to extract common and distinctive linguistic descriptions from syntactic treebanks [16]. The extracted contrastive rules are defined by a statistically significant difference in frequency and precision, and classified as common and distinctive rules across the set of treebanks. We illustrate our method by working on object word order using Universal Dependencies (UD) treebanks in 6 Romance languages: Brazilian Portuguese, Catalan, French, Italian, Romanian and Spanish. We discuss the limitations faced due to inconsistent annotation and the feasibility of conducting contrastive studies using the UD collection.

8.1.9 Discrete Latent Structure in Neural Networks

Participants: Vlad Niculae (*Informatics Institute Amsterdam*), Caio Corro, Nikita Nangia (*NYU - New York University, New York*), Tsvetomila Mihaylova (*IST / Técnico Lisboa - Instituto Superior Técnico, Universidade de Lisboa, Lisboa*), André Martins (*IST / Técnico Lisboa - Instituto Superior Técnico, Universidade de Lisboa, Unbabel*).

Many types of data from fields including natural language processing, computer vision, and bioinformatics, are well represented by discrete, compositional structures such as trees, sequences, or matchings. Latent structure models are a powerful tool for learning to extract such representations, offering a way to incorporate structural bias, discover insight about the data, and interpret decisions. However, effective training is challenging, as neural networks are typically designed for continuous computation. This text explores three broad strategies for learning with discrete latent structure: continuous relaxation, surrogate gradients, and probabilistic estimation. Our presentation relies on consistent notations for a wide range of models. As such, we reveal many new connections between latent structure learning strategies, showing how most consist of the same small set of fundamental building blocks, but use them differently, leading to substantially different applicability and properties [11].

8.1.10 Nested Named Entity Recognition as Single-Pass Sequence Labeling

Participants: Alberto Muñoz-Ortiz (*Universidade da Coruña*), David Vilares (*Universidade da Coruña*), Caio Corro, Carlos Gómez-Rodríguez (*Universidade da Coruña*).

In this paper [18] we cast nested named entity recognition (NNER) as a sequence labeling task by leveraging prior work that linearizes constituency structures, effectively reducing the complexity of this structured prediction problem to straightforward token classification. By combining these constituency linearizations with pretrained encoders, our method captures nested entities while performing exactly n tagging actions. Our approach achieves competitive performance compared to less efficient systems, and it can be trained using any off-the-shelf sequence labeling library.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

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.

10 Partnerships and cooperations

10.1 International initiatives

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 saliencies 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 identification 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

Astrid Maturation: TrustedNews

Participants: Guillaume Gravier, Morgane Casanova, Laurent Amsaleg.

Duration: 36 months, started Nov. 2025

This ANR-AID funded project aims to automatically assess the reliability of online content (for both civilian and military users) by identifying biases, manipulative discourse, or hostile narratives, and classifying texts based on their nature—facts, opinions, or argumentation. Using a hybrid approach that combines symbolic AI and neural networks, it delivers transparent analysis to guide users without replacing fact-checking.

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, Thomas Derrien, 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. The Ph.D. of Thomas Derrien explores the issues in relation with multimodal entity linking.

11 Dissemination

Laurent Amsaleg Caio Corro Guillaume Gravier Pascale Sebillot

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees

- Laurent Amsaleg was the PhD Symposium chair of SISAP 2025.

11.1.2 Scientific events: selection

Member of the conference program committees

- Laurent Amsaleg was a senior area chair of ACM Multimedia 2025.
- Laurent Amsaleg was PC member of ICMR, ICME, MMM, SISAP, CBMI.
- Pascale Sébillot was a PC member for CNIA, TALN
- Caio Corro was an Area Chair for ACL 2025 and EMNLP 2025
- Caio Corro was a PC member for TALN 2025

Reviewer

- Caio Corro was a reviewer for UncertaiNLP2025

11.1.3 Journal

Member of the editorial boards

- Caio Corro is a member of the editorial board of TAL

Reviewer - reviewing activities

- Caio Corro was a reviewer for TMLR

11.1.4 Research administration

- Guillaume Gravier was director of IRISA (UMR 6074) till December 2025
- Pascale Sébillot was deputy director of IRISA till December 2025

11.2 Teaching - Supervision - Juries - Educational and pedagogical outreach

11.2.1 Teaching

- Master: Laurent Amsaleg, Bases de données avancées, 25h, M2, INSA Rennes, France
- Master: Guillaume Gravier, Natural Language Processing, 8h, M1, INSA Rennes
- Licence: Guillaume Gravier, Natural language processing, 8h, L3, INSA Rennes
- 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, Machine Learning, 10h, L3, INSA Rennes, France

11.2.2 Supervision

- PhD in progress: Hugo Thomas, Zero-shot and few-shot relation extraction in press archives. Started Sept. 2022, Guillaume Gravier and Pascale Sébillot
- Ph.D. in progress: Thomas Derrien, Liage d'entités multimodal. Started Oct. 2025, Laurent Amsaleg and Pascale Sébillot
- Ph.D. in progress: Lilas Pastré, À la conquête de l'Ouest-France. Started Oct. 2025, Laurent Amsaleg and Christian Le Bart (sciencePo) and Olivier Trédan (Univ. Rennes)
- Ph.D. in progress: Ayoub Hammal, Language modeling under distribution shifts. Started Nov. 2024, Caio Corro and Pierre Zweigenbaum (LISN)
- Ph.D. in progress: Carolina Jeronimo De Almeida, Machine learning for temporal graphs. Started Sept. 2022, Silvio Guimarães (PUC Minas, Brésil), Guillaume Gravier, Simon Malinowski (équipe MALT)
- Ph.D. finished in Nov. 2025: Benno Uthayasooriyar, Insurance Document Understanding with Transformers based Language Models, Caio Corro, Franck Vermet (Université de Bretagne Occidentale) and Antoine Ly (entreprise SCOR)

11.2.3 Juries

- Pascale Sébillot was the president of the PhD. jury of Darun Cao, Univ. Bretagne Sud, Feb. 2025
- Pascale Sébillot was a jury member for the PhD. of Oumaima El Khezzari, Nantes Univ., Feb. 2025
- Pascale Sébillot was a reviewer for the PhD. of Marco Naguib, Univ. Paris-Saclay, Sept. 2025
- Pascale Sébillot was the president of the PhD. jury of Élise Lincker, Conservatoire national des arts et métiers, Dec. 2025
- Caio Corro was a jury member for the Ph.D. of Junjie Yang, Institut polytechnique de Paris, July 2025
- Caio Corro was a jury member for the Ph.D. of Santiago Herrera, Université de Nanterre, September 2025

11.2.4 Specific official responsibilities in science outreach structures

- Caio Corro is a member of the Comité de Rédaction of Bulletins de l'AFIA

11.2.5 Participation in Live events

- Laurent Amsaleg ran two webinars inside the premises of Ouest-France, presenting the main concepts of IA and detailing the synapses project.

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](https://doi.org/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](https://doi.org/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. Toliás, 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](https://doi.org/10.1109/MMUL.2018.023121161). URL: <https://hal.inria.fr/hal-01848539>.

12.2 Publications of the year

International journals

- [10] M. Faysse, P. Fernandes, N. Guerreiro, A. Loison, D. Alves, C. Corro, N. Boizard, J. Alves, R. Rei, P. R. Martins, A. Casademunt, F. Yvon, A. Martins, G. Viaud, C. Hudelot and P. Colombo. ‘CroissantLLM: A Truly Bilingual French-English Language Model’. In: *Transactions of Machine Learning Research* (12th Mar. 2025). URL: <https://hal.science/hal-04574908> (cit. on p. 17).
- [11] V. Niculae, C. Corro, N. Nangia, T. Mihaylova and A. Martins. ‘Discrete Latent Structure in Neural Networks’. In: *Foundations and Trends in Signal Processing*. Foundations and Trends in Signal Processing 19.2 (2nd June 2025), pp. 99–211. DOI: [10.1561/2000000134](https://doi.org/10.1561/2000000134). URL: <https://hal.science/hal-05457940> (cit. on p. 18).
- [12] Z. Zhao, H. Zhang, R. Li, R. Sicre, L. Amsaleg, M. Backes, Q. Li, Q. Wang and C. Shen. ‘Revisiting Transferable Adversarial Images: Systemization, Evaluation, and New Insights’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (16th Sept. 2025), pp. 1–16. DOI: [10.1109/TPAMI.2025.3610085](https://doi.org/10.1109/TPAMI.2025.3610085). URL: <https://inria.hal.science/hal-05267252> (cit. on p. 14).

International peer-reviewed conferences

- [13] N. Boizard, H. Gisserot-Boukhlef, D. M. Alves, A. F. T. Martins, A. Hammal, C. Corro, C. Hudelot, E. Malherbe, E. Malaboeuf, F. Jourdan, G. Hautreux, J. Alves, K. El-Haddad, M. Faysse, M. Peyrard, N. M. Guerreiro, P. Fernandes, R. Rei and P. Colombo. ‘EuroBERT: Scaling Multilingual Encoders for European Languages’. In: *Second Conference on Language Modeling*. COLM 2025 - Second Conference on Language Modeling. Montreal, Canada, 10th Mar. 2025, pp. 1–28. URL: <https://hal.science/hal-05226285> (cit. on p. 16).

- [14] C. Corro, M. Lacroix and J. L. Roux. ‘Bregman Conditional Random Fields: Sequence Labeling with Parallelizable Inference Algorithms’. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. ACL 2025 - 63rd Annual Meeting of the Association for Computational Linguistics. Vol. 1. Vienne, Austria: Association for Computational Linguistics, 27th July 2025, pp. 29557–29574. DOI: [10.18653/v1/2025.acl-long.1430](https://doi.org/10.18653/v1/2025.acl-long.1430). URL: <https://hal.science/hal-05360479> (cit. on p. 15).
- [15] A. Hammal, B. Uthayasooryar and C. Corro. ‘Few-Shot Domain Adaptation for Named-Entity Recognition via Joint Constrained k-Means and Subspace Selection’. In: COLING 2025 - 31st International Conference on Computational Linguistics. Abu Dhabi, United Arab Emirates, 2025, pp. 9902–9916. DOI: [10.48550/arxiv.2412.00426](https://doi.org/10.48550/arxiv.2412.00426). URL: <https://hal.science/hal-04877776> (cit. on p. 15).
- [16] S. Herrera, I.-M. Silai, C. Corro, B. Guillaume and S. Kahane. ‘Extraction of Contrastive Rules from Syntactic Treebanks: A Case Study in Romance Languages’. In: QUASY 2025 - Third Workshop on Quantitative Syntax. Ljubljana, Slovenia, Aug. 2025, pp. 26–38. URL: <https://inria.hal.science/hal-05404380> (cit. on p. 17).
- [17] F. Meyer, L. Guichard, D. Coquenot, G. Gravier, Y. Soullard and B. Coïasnon. ‘Relaxed syntax modeling in Transformers for future-proof license plate recognition’. In: *International Conference on Document Analysis and Recognition (ICDAR) 2025*. ICDAR 2025 - International Conference on Document Analysis and Recognition. Wuhan, China, 2025, pp. 154–171. DOI: [10.1007/978-3-032-04627-7_9](https://doi.org/10.1007/978-3-032-04627-7_9). URL: <https://hal.science/hal-05147482> (cit. on p. 16).
- [18] A. Muñoz-Ortiz, D. Vilares, C. Corro and C. Gómez-Rodríguez. ‘Nested Named Entity Recognition as Single-Pass Sequence Labeling’. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. EMNLP 2025 - Conference on Empirical Methods in Natural Language Processing. Suzhou, China, Nov. 2025, pp. 9993–10002. DOI: [10.18653/v1/2025.findings-emnlp.530](https://doi.org/10.18653/v1/2025.findings-emnlp.530). URL: <https://hal.science/hal-05457961> (cit. on p. 18).
- [19] B. Uthayasooryar, A. Ly, F. Vermet and C. Corro. ‘Training LayoutLM from Scratch for Efficient Named-Entity Recognition in the Insurance Domain’. In: *Proceedings 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 - 31st International Conference on Computational Linguistics. Abu Dhabi, United Arab Emirates, 2025, pp. 1–9. URL: <https://hal.science/hal-04877824> (cit. on p. 15).

Edition (books, proceedings, special issue of a journal)

- [20] *18th International Conference on Similarity Search and Applications: 18th International Conference, SISAP 2025, Reykjavik, Iceland, October 1–3, 2025, Proceedings*. SISAP 2025 - 18th International Conference on Similarity Search and Applications. Vol. 16134. Lecture Notes in Computer Science. Reykjavik, Iceland: Springer Nature Switzerland, 2025. DOI: [10.1007/978-3-032-06069-3](https://doi.org/10.1007/978-3-032-06069-3). URL: <https://inria.hal.science/hal-05315331>.
- [21] J.-D. Kant, G. Bonnet and D. Longin, eds. *IA & économie*. Association Française pour l’Intelligence Artificielle 129 (July 2025). URL: <https://hal.science/hal-05289899>.

12.3 Cited publications

- [22] 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. 9).
- [23] 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. 8, 10, 11).
- [24] 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. 11).
- [25] L. Amsaleg, B. Þ. Jónsson and H. Lejsek. ‘Scalability of the NV-tree: Three Experiments’. In: *SISAP*. 2018 (cit. on p. 11).

- [26] R. Balu, T. Furon and L. Amsaleg. ‘Sketching techniques for very large matrix factorization’. In: *ECIR*. 2016 (cit. on p. 11).
- [27] S. Berrani, H. Boukadida and P. Gros. ‘Constraint Satisfaction Programming for Video Summarization’. In: *ISM*. 2013 (cit. on p. 12).
- [28] B. Biggio and F. Roli. ‘Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning’. In: *Pattern Recognition* (2018) (cit. on p. 9).
- [29] P. Bosilj. ‘Image indexing and retrieval using component trees’. Theses. Université de Bretagne Sud, 2016 (cit. on p. 8).
- [30] X. Bost. ‘A storytelling machine? : Automatic video summarization: the case of TV series’. PhD thesis. University of Avignon, France, 2016 (cit. on p. 12).
- [31] M. Budnik, M. Demirdelen and G. Gravier. ‘A Study on Multimodal Video Hyperlinking with Visual Aggregation’. In: *ICME*. 2018 (cit. on p. 12).
- [32] 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. 9).
- [33] R. Carlini Sperandio, S. Malinowski, L. Amsaleg and R. Tavenard. ‘Time Series Retrieval using DTW-Preserving Shapelets’. In: *SISAP*. 2018 (cit. on p. 9).
- [34] 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. 8).
- [35] A. Delvinioti, H. Jégou, L. Amsaleg and M. E. Houle. ‘Image Retrieval with Reciprocal and shared Nearest Neighbors’. In: *VISAPP*. 2014 (cit. on p. 11).
- [36] 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. 8, 10).
- [37] H. Farid. *Photo Forensics*. The MIT Press, 2016 (cit. on p. 9).
- [38] M. Gambhir and V. Gupta. ‘Recent automatic text summarization techniques: a survey’. In: *Artif. Intell. Rev.* 47.1 (2017) (cit. on p. 12).
- [39] I. Goodfellow, Y. Bengio and A. Courville. *Deep Learning*. MIT Press, 2016 (cit. on p. 7).
- [40] 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. 10).
- [41] A. Iscen, L. Amsaleg and T. Furon. ‘Scaling Group Testing Similarity Search’. In: *ICMR*. 2016 (cit. on p. 11).
- [42] A. Iscen, G. Tolia, Y. Avrithis and O. Chum. ‘Mining on Manifolds: Metric Learning without Labels’. In: *CVPR*. 2018 (cit. on pp. 8, 11).
- [43] B. Þ. Jónsson, G. Tómasson, H. Sigurþórsson, Á. Eríksdóttir, L. Amsaleg and M. K. Larusdóttir. ‘A Multi-Dimensional Data Model for Personal Photo Browsing’. In: *MMM*. 2015 (cit. on p. 10).
- [44] 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. 10).
- [45] 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. 9).
- [46] 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. 8).
- [47] S. Leroux, P. Molchanov, P. Simoons, 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. 8).

- [48] A. Lods, S. Malinowski, R. Tavenard and L. Amsaleg. ‘Learning DTW-Preserving Shapelets’. In: *IDA*. 2017 (cit. on p. 9).
- [49] 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. 9).
- [50] D. Shahaf and C. Guestrin. ‘Connecting the dots between news articles’. In: *KDD*. 2010 (cit. on p. 12).
- [51] M. Shi, H. Caesar and V. Ferrari. ‘Weakly Supervised Object Localization Using Things and Stuff Transfer’. In: *ICCV*. 2017 (cit. on p. 8).
- [52] R. Sicre, Y. Avrithis, E. Kijak and F. Jurie. ‘Unsupervised part learning for visual recognition’. In: *CVPR*. 2017 (cit. on p. 8).
- [53] R. Sicre and H. Jégou. ‘Memory Vectors for Particular Object Retrieval with Multiple Queries’. In: *ICMR*. 2015 (cit. on p. 8).
- [54] 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. 9).
- [55] O. Siméoni, A. Iscen, G. Tolias, Y. Avrithis and O. Chum. ‘Unsupervised Object Discovery for Instance Recognition’. In: *WACV*. 2018 (cit. on p. 8).
- [56] H. O. Song, Y. Xiang, S. Jegelka and S. Savarese. ‘Deep Metric Learning via Lifted Structured Feature Embedding’. In: *CVPR*. 2016 (cit. on p. 8).
- [57] 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. 12).
- [58] 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. 12).
- [59] V. Vukotić. ‘Deep Neural Architectures for Automatic Representation Learning from Multimedia Multimodal Data’. Theses. INSA de Rennes, 2017 (cit. on pp. 8, 10).
- [60] 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. 8).
- [61] V. Vukotić, C. Raymond and G. Gravier. ‘Generative Adversarial Networks for Multimodal Representation Learning in Video Hyperlinking’. In: *ICMR*. 2017 (cit. on p. 8).
- [62] J. Weston, S. Chopra and A. Bordes. ‘Memory Networks’. In: *CoRR* abs/1410.3916 (2014). arXiv: [1410.3916](https://arxiv.org/abs/1410.3916) (cit. on p. 8).
- [63] 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. 12).
- [64] J. Zahálka and M. Worring. ‘Towards interactive, intelligent, and integrated multimedia analytics’. In: *VAST*. 2014 (cit. on p. 11).
- [65] L. Zhang, M. Shi and Q. Chen. ‘Crowd Counting via Scale-Adaptive Convolutional Neural Network’. In: *WACV*. 2018 (cit. on p. 8).
- [66] 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](https://arxiv.org/abs/1707.01083) (cit. on p. 8).